ICA 95

TRONDHEIM, NORWAY

Proceedings of the 15th
International Congress on Acoustics

Trondheim, Norway, 26-30 June 1995

VOLUME III
Executive committee:
Co-Chairmen: T. Gjestland & J. Tro
Exhibition manager: T. Berge
Proceedings editor: M.J. Newman
Secretariat: Sabine Nørstrand, SEUV Congress Department

Commission on Acoustics:
Chairman: T. Kihlman, Sweden
Vice-Chairman: G. Rasmussen, Denmark
Secretary: A. Alippi, Italy
Members:
G. Daigle, Canada
R. Zhang, China
G.J. Quentin, France
G.M. Sessler, Germany
S. Namba, Japan
A. Krokstad, Norway
L.M. Lyamshev, Russia
P. Bury, Slovakia
C.G. Rice, UK
R. Lyon, USA

Associate Members:
H. Myncke, Belgium
F. Kolmer, Czech Rep.
J. Blauert, Germany
J. Lighthill, UK

Scientific Committee:
A. Krokstad, NTH, University of Trondheim, chairman
B. Angelsen, University of Trondheim
J. Borenius, Acoustics Laboratory, Helsinki
P.V. Brøtø, Denmark
G. Flottorp, University of Oslo
J.M. Hovem, NTH, University of Trondheim
T. Kihlman, CTH, University of Göteborg
U. Kristiansen, NTH, University of Trondheim
J. Sundberg, KTH, University of Stockholm
J.N. Tjøtta, University of Bergen
T.E. Vigran, NTH, University of Trondheim
U.K. Laine, University of Technology, Helsinki
A.K. Foldvik, AVH, University of Trondheim
FOREWORD

These proceedings are for the 15th triennial International Congress on Acoustics which is being held in Trondheim, hosted by the Acoustical Society of Norway. At the time of writing, the proceedings contain over 400 papers on all subjects within acoustics. The proceedings are divided into broad subject categories and papers are in alphabetical order of the first author’s family name. The page numbers for each paper are printed in the congress program.

The Executive Committee takes this opportunity to thank those involved with the organization of the congress, there are many who have but only space to name a few. Firstly we must thank Asbjørn Krokstad for his unstinting efforts in making the congress possible as President of the Acoustical Society of Norway, member of the International Commission on Acoustics, and Chairman of the Scientific Committee. The financial assistance from IUPAP has enabled our small society to organize such a large congress without undue financial risk. The Congress would not have been possible without the professional help of SEVU, the congress secretariat, with special thanks to Sabine Nørstrud. We wish to thank all of the organizers of the structured sessions for providing a solid foundation for the scientific content of the congress. Last but not least we thank the NTH-SINTEF Acoustics Group for their contribution.

Mike Newman
Editor

Trondheim, April 1995
ACOUSTIC CONSULTANTS

IGP AS covers a broad range of acoustic consulting activities.

The following subjects are our specialities.

- Building Acoustics
- Room Acoustics
- Noise and Vibration, measurement and control
- PA and Sound Systems

Our acoustics group consists of the following specialists:

**Kjell Nordmark**
Chartered Eng. building tech.
Group leader
16 years experience

**Endre Grimsmo**
Chartered Eng. telecoms.
Quality assurance manager
20 years experience

**Svein Strøm**
M.Sc. physics, acoustics
25 years experience

**Frode Bye**
Engineer, audio-visual
13 years experience

**Michael Newman**
Ph.D. vibration and noise
15 years experience

**Ole Martin Ulvan**
Measuring engineer
5 years experience

Cooperative agreements with:

**Professor Asbjørn Kroksstad**
Faculty of telecommunication and computer science, acoustics
The Norwegian Institute of Technology

and

Acoustics Research Center
SINTEF DELAB

Address: IGP AS
P.O. Box 2564
N-7001 TRONDHEIM
Norway

Telephone: (+ 47) 73 89 58 00
Telefax: (+ 47) 73 89 58 89
E-mail: **@igp.no (** = initials)

IGP AS is one of the leading electrical consultants in Norway, with approx. 170 employees in 14 offices. The main office is in Trondheim close to the Olav’s Hall where the ICA’95 opening session is being held.
# TABLE OF CONTENTS

## VOLUME III

### SPEECH

**Sharon M. Abel**  
SPEECH PERCEPTION DEFICITS IN NOISE: CONTRIBUTING FACTORS .................................. 3

**Shuichi Arai, Ken'ichi Kido**  
A SPEAKER VERIFICATION METHOD BASED ON THE TALKER SIMILARITY SCALE DEFINED IN EACH PHONEME .......................................................................................................................... 7

**Yasuhiko Arai, Hisanori Kanasashi, Toshimitsu Minowa**  
MEASUREMENT OF USER-SATISFACTION RATES OF THE VOICE INTERFACE BASED ON WORD RECOGNITION AND ESTIMATION OF ACCEPTABLE RECOGNITION ACCURACY .................................................. 11

**Dawn M. Behne, Bente H. Moxness**  
EFFECTS OF POSTVOCAL VOICING AND DISTINCTIVE VOWEL LENGTH ON SYLLABLE-INTERNAL TIMING: RESULTS FROM REAL- AND PSEUDO-WORDS IN NORWEGIAN ................................................................. 15

**Ocke-Schwen Bohn, Winifred Strange**  
ACOUSTIC DETERMINANTS OF VOWEL IDENTITY FOR GERMAN LISTENERS: TARGET SPECTRAL INFORMATION NOT NEEDED ............................................................. 19

**P.O.A.L. Davies, A. Barney, C.H. Shadle**  
AEROACOUSTIC SOURCES IN A VOCAL TRACT MODEL .................................................. 23

**César Díaz, Concepción Velázquez**  
A LIVE SPANISH LANGUAGE EVALUATION OF THE RASTI-METHOD ................................ 27

**Guan Dinghua, Chu Min, Lu Shinan**  
A CHINESE TEXT-TO-SPEECH SYSTEM WITH HIGH INTELLIGIBILITY AND HIGH NATURALNESS .................................................................................................................. 31

**Titoj Duono, Naomitsu Ikeda, Akira Watanabe**  
EFFECT OF SPEECH PARAMETERS ON SPEAKER INDEPENDENT WORD RECOGNITION BASED ON PHONEME-RECOGNITION AND DICTIONARY ............................ 35

**Johan Sundberg**  
ACOUSTICS OF THE SINGING VOICE .................................................................................. 39

**Christophe Gerard, Marc Baunry**  
ONDELETTES ET PARAMÉTRISATION DU SIGNAL DE PAROLE EN MILIEU BRUITÉ .......................................................... 45

**Bruno Gramatica, Robert Ruiz, Claude Legros**  
COMPARISON BETWEEN TIME-SCALE AND TIME-FREQUENCY APPROACHES FOR THE DETECTION OF STRESS IN CONSONANTS .......................................................... 49
Berard Harmegnies, Marielle Bruyninckx
STABILITÉ TEMPORALE D’INDICATEURS
SPECTRAUX DE LA QUALITÉ VOCALE ........................................ 53

Osamu Hashimoto, Sho Kimura
EVALUATION OF SPEECH AUDIBILITY IN A ROOM BY
MEANS OF TRI-SYLLABLE ARTICULATION TEST TAKING
CONSIDERATION FREQUENCY OF SPEECH RATE IN CONVERSATION ...... 57

Hynek Hermansky
EXPLORING TEMPORAL DOMAIN FOR
ROBUSTNESS IN SPEECH RECOGNITION.................................... 61

Melyvn J. Hunt
A ROBUST METHOD OF DETECTING THE PRESENCE OF VOICED SPEECH .......... 65

Takashi Ikeda, Akira Watanabe, Yuichi Ueda, Kouji Tasaki
CHARACTERISTICS OF A HEARING AID FOR TELEPHONIC
SPEECH BY SINGLE RESONANT ANALYSIS.................................... 69

Ken’iti Kido, Kenji Tanahashi, Yasuhiro Ohuchi, Masako Matsuzaki, Takayuki Kuroda
A METHOD OF SPEECH QUALITY EVALUATION FOR PATIENTS
OF ORTHODONTIC TREATMENT .................................................. 73

Dae-won Kim
AN EMG STUDY OF THE FEATURE “TENSITY” WITH
REFERENCE TO ENGLISH AND KOREAN...................................... 77

Birger Kollmeier, Torsten Dau, Inga Holube, Matthias Wesselkamp
MODELING THE EAR’S “EFFECTIVE” SIGNAL PROCESSING
TO PREDICT PSYCHOACOUSTICS AND SPEECH PERCEPTION
IN NORMAL AND IMPAIRED LISTENERS .................................... 81

Winfried Krebber, Jürg Becker
PREDICTION OF SPEECH TRANSMISSION
QUALITY OF TELEPHONE HANDSETS USING
SHORT-TIME SUBBAND ANALYSIS AND PSYCHOACOUSTIC MODELS .......... 85

François Malbos, Marc Baudry
DETECTION D’EVENTEMENTS IMPULSIONNELS EN PAROLE .................. 89

Hiroshi Matsumoto, Yoshinori Nimura
AN IMPROVED SPECTRAL SUBTRACTION WITH
SMOOTHING FOR NOISY SPEECH RECOGNITION .......................... 93

Masafumi Matsumura, Takuya Niikawa, Koji Shimizu,
Yasuji Hashimoto, Tatsuya Morita
MEASUREMENT AND ANALYSIS OF 3D SHAPES OF VOCAL
TRACT, DENTAL CROWN AND NASAL CAVITY USING MR .................. 97

Richard S. McGowan
KNOWLEDGE FROM SPEECH PRODUCTION USED IN
SPEECH TECHNOLOGY: ARTICULATORY SYNTHESIS ....................... 101

Yasuhiro Minami, Sadaoki Furui
UNIVERSAL ADAPTATION METHOD BASED ON HMM COMPOSITION ........ 105
Pedro J. Moreno, Bhiksha Raj, Richard M. Stern
APPROACHES TO ENVIRONMENT
COMPENSATION IN AUTOMATIC SPEECH RECOGNITION......................... 109

Takayoshi Nakai, Naoki Imagawa, Hisayoshi Suzuki
A NEW REAL TIME HELIUM SPEECH UNSCRAMBLER USING SINGLE DSP ...... 113

Paroo Nihalani
DISTINGUISHING APICAL AND LAMINAL STOPS: A STUDY OF RELEVANT
ARTICULATORY AND ACOUSTIC PROPERTIES........................................ 117

Yasuhisa Niimi, Xiao-Yang, Gian-Geng Ke, Yutaka Kobayashi
TONE RECOGNITION OF MANDARIN SPEECH......................................... 121

Ryouchi Nishimura, Futoshi Asano, Yōiti Suzuki, Toshio Sone
SPEECH ENHANCEMENT BASED ON THE WAVELET TRANSFORM USING
DECAYING SINUSOIDAL WAVELET...................................................... 125

Georg E. Ottesen, Berit Horvei, Sverre Stensby
SPEECH SYNTHESIS OF TELETEXT ..................................................... 129

Jose Manuel Pardo, Francisco M. Giménez de los Galanes, José Angel Vallejo, Miguel
Angel Berrojo, José Manuel Montero, Emilia Enríquez, Angeles Romero Cortés
SPANISH TEXT-TO-SPEECH, FROM PROSODY TO ACOUSTICS .................. 133

Bernd Pomponio-Marschall
MIDSAGITTAL DISTANCES AND AREA FUNCTIONS
– BASIC CONSIDERATIONS FOR AN ARTICULATORY MODEL ..................... 137

Robert Ruiz, Claude Legros
VOWEL SPECTRAL CHARACTERISTICS TO DETECT VOCAL STRESS ............ 141

Christine H. Shadle
MODELLING THE NOISE SOURCE IN VOICED FRICATIVES ....................... 145

Sumi Shigeno
VOWEL–CONTINGENT EFFECTS IN THE PERCEPTION OF
STOP CONSONANTS UNDER AUDIO-VISUAL PRESENTATION ..................... 149

Sverre Stensby, Berit Horvei, Georg E. Ottesen, Arne Kjell Foldvik
PRONUNCIATION LEXICONS FOR
SPEECH SYNTHESIS AND SPEECH RECOGNITION ................................ 153

Andrzej Skumiel, Mikolaj Labowski
ELECTRIC FIELD EFFECT ON THE ABSORPTION
OF ULTRASONIC WAVES IN CONDENSER OIL ...................................... 157

Jouji Suzuki, Miyuki Setoh, Tetsuya Shimamura
EXTRACTION OF PRECISE FUNDAMENTAL FREQUENCY
BASED ON HARMONIC STRUCTURE OF SPEECH .................................. 161

Yōiti Suzuki, Futoshi Asano, Shoei Sato, Toshio Sone
APPROPRIATE GROUPING OF PHONEME BASED ON DISTINCTIVE FEATURES
FOR EVALUATING THE PERFORMANCE OF HEARING AIDS ..................... 165

Yuichi Ueda and Akira Watanabe
SPEECH FEATURE EXTRACTION USING A DSP
SYSTEM FOR TACTILE DISPLAY ...................................................... 169
Hidenori Usuki, Jouji Suzuki, Tetsuya Shimamura
CHARACTERISTICS OF MISPRONUNCIATION AND HESITATION
IN JAPANESE TONGUE TWISTER .................................................. 173

Dirk Van Compernolle
NONLINEAR FILTERING OF SPECTRAL FEATURES
FOR ROBUST SPEECH RECOGNITION ............................................ 177

Robert Vich
PITCH SYNCHRONOUS LINEAR
PREDICTIVE CZECH AND SLOVAK TEXT-TO-SPEECH SYNTHESIS ...... 181

Akira Watanabe, Shigeki Tokunaga, Shingo Tomishige
SPEECH VISUALIZATION BY EXTRACTING
FEATURES WITH NEURAL NETWORKS ........................................... 185

Yunxin Zhao, Ted H. Applebaum, Brian A. Hanson
ACOUSTIC NORMALIZATION AND ADAPTION FOR
MICROPHONE-CHANNEL CHARACTERISTICS .................................. 189

PHYSIOLOGICAL AND PSYCHOLOGICAL ACOUSTICS

S. M. Abel, E. O. Pelausa, J. Simard, L. Dempsey
THE CANADIAN FORCES HEARING CONSERVATION
PROGRAM: A THREE-YEAR PROSPECTIVE EVALUATION .................. 195

Birgitta Berglund, Anna Preis
LOUDNESS AND ANNOYANCE OF INTERMITTENT
AND CONTINUOUS ENVIRONMENTAL SOUNDS ............................... 199

Frans A. Bilsen, Han H. Dols
NEW RESULTS ON THE PITCH OF INTERAURALLY
DELAYED WHITE NOISE .......................................................... 203

Jens Blauert, Hilmar Lehnerl
THE AUDITORY REPRESENTATION IN VIRTUAL REALITY ................. 207

Marie-Claire Botte, Carolyn Drake, Nathalie Panissal-Vieu, Caroline Bey
EFFECT OF TONE FREQUENCY RANGE ON TEMPO SENSITIVITY ............... 211

P. Camiz, M. Beato
A MATHEMATICAL DESCRIPTION OF PSYCHOACOUSTIC SPACE ............. 215

M. A. Cheatham
MECHANISMS OF COCHLEAR FUNCTION;
OBSERVATIONS FROM MAMMALIAN HAIR CELLS ............................. 221

Gordon Flottorp
TEN YEAR TEST OF THE NORWEGIAN
REGULATIONS RELATING TO NOISE AT THE WORKPLACE .................. 225

R. W. Gatehouse, Wilma Veenhof
DO INSERT EARPHONES REALLY MAKE
A DIFFERENCE IN ACHIEVED TRESHOLDS? ..................................... 229
Klaus Genuit
EXAMINATION OF OBJECTIVE EVALUATION OF BINAURAL ANALYSIS .......... 233

Christian Giguère, Hans Kunov, Guido F. Smoorenburg
COMPUTATIONAL MODELLING OF PSYCHOACOUSTIC COMBINATION
TONES AND DISTORTION-PRODUCT OTOACOUSTIC EMISSIONS .................. 237

Beno Groothoff
A STUDY INTO ASYMMETRIC HEARING LOSS AMONG A TRUCK DRIVER
POPULATION ON THE SUNSHINE COAST IN QUEENSLAND ................... 241

Anthony W. Gummer, Werner Hemmert, Serena Preyer, Hans-Peter Zenner
A RESONANT TECTORIAL MEMBRANE WITH
AN ELECTROMECHANICAL OUTER HAIR CELL
AS THE MECHANISM FOR COCHLEAR FREQUENCY SELECTIVITY ............. 245

Simo Hemilä, Sirpa Nummela, Tom Reuter
ISOMETRIC MAMMALIAN MIDDLE EARS:
THE EFFECT OF SIZE ON AUDIOGRAMS ........................................ 249

E. Hojan, R. Drazazga
NEW METHOD OF HEARING AIDS FITTING ...................................... 253

Magne Kringlebotn
VOLUME FLOWS IN THE INNER EAR WINDOWS .................................. 257

Kazuoki Kuramoto, Kensei Otsmatsu, Shin’ya Kuwahara, Shizuma Yamaguchi
AN EQUAL-LOUDNESS CONTOUR FOR UNDERWATER
ACOUSTIC SIGNAL AND ITS DEPTH DEPENDENCY .............................. 261

Hilmar Lehnert, Marcus Richter
AUDITORY VIRTUAL ENVIRONMENT:
SIMPLIFIED TREATMENT OF REFLECTIONS .................................. 265

Tomasz Letowski, Diana W. Emanuel
MOST COMFORTABLE LISTENING LEVELS
FOR LISTENING THROUGH EARPHONES AND IN A SOUND FIELD ........... 269

Wojciech Majewski
AUTOMATIC AND AURAL-PERCEPTUAL SPEAKER
VERIFICATION IN THE PRESENCE OF MASKING NOISE ....................... 273

Sirpa Nummela, Simo Hemilä, Tom Reuter
THE MIDDLE EARS OF BIRDS AND MAMMALS:
A COMPARISON OF STRUCTURE AND FUNCTION .............................. 277

P. Rubak, K. Hermansen, F. K. Fink, U. Hartmann
PSYCHOACOUSTIC EVALUATION OF SPEECH
ENHANCEMENT USING THE PARTRAN CONCEPT .............................. 281

Jesper Sandvad, Dorte Hammershøi
SPATIAL RESOLUTION AND INTERACTIVE CHANGES OF HEAD-RELATED
TRANSFER FUNCTIONS FOR VIRTUAL REALITY ................................ 285

Michael Friis Sørensen, Morten Iydolf, Peder Chr. Frandsen, Henrik Møller
COMPARISON OF EIGHT PSYCHOMETRIC
METHODS FOR MEASURING THRESHOLD OF HEARING ......................... 289
BIOACOUSTICS AND MEDICAL ACOUSTICS

Lawrence A. Crum, Wing Law
THE RELATIVE ROLES OF THERMAL AND NONTHERMAL
EFFECTS IN THE USE OF HIGH INTENSITY FOCUSED ULTRASOUND
FOR THE TREATMENT OF BENIGN PROSTATIC HYPERPLASIA ........................................... 315

Michael Delius
BIOLOGICAL EFFECTS AND MECHANISM OF
ACTION OF EXTRACORPOREAL. SHOCK WAVES .................................................. 319

Emad S. Ebbini, Ralf Seip
NONINVASIVE FEEDBACK FOR THERAPEUTIC
APPLICATIONS USING HIGHLY-FOCUSED ULTRASONIC FIELDS ......................... 323

Helge E. Engan
ULTRASOUND TRANSDUCERS FOR MEDICAL USE .............................................. 327

Andreas Helmdal, Torgrim Lie, Alf O. Brubakk
MEASUREMENTS OF GAS BUBBLE SIZES USING HIGH
FREQUENCY ULTRASOUND IMAGING .......................................................... 331

Ange Hernandez, Olivier Basset, Pierre Chirossel, Christian Cachard, Gerard Gimenez
UTILIZATION OF A PANTOGRAPH LOCALIZATION SYSTEM FOR SPECKLE
REDUCTION BY SPATIAL COMPOUNDING, IN B-SCAN ECHOGRAPHY ............. 335

Sverre Holm
MEDICAL ULTRASOUND TRANSDUCERS AND BEAMFORMING ......................... 339

Hans T. Kaarmann, Reinhard Lerch
DEVELOPMENT OF ULTRASOUND TRANSDUCERS
FOR MEDICAL IMAGING .................................................................................. 343
Xiaoming Lai, Hans Torp, Kjell Kristoffersen
EXTENDED AUTOCORRELATION METHOD FOR COLOR FLOW IMAGING .......... 347

Torgrim Lie, Aage Gronningsaeter, Klaus-Dieter Bolz, Andreas Heimdal
ULTRASONOGRAPHIC STENT-IMAGING ARTIFACTS .................................. 351

O. Lotsberg, J. M. Hovem, B. Aksum
EXPERIMENTAL OBSERVATION OF SUBHARMONIC
OSCILLATIONS IN INFOSON BUBBLES ............................................. 355

Michael Nickel, Aage Gronningsaeter, Bjørn Angelsen
REDUCTION OF STATIONARY REVERBERATIONS
IN CARDIAC ULTRASOUND IMAGING ............................................ 359

William D. O’Brien, Jr., Eric J. Chien, W. Kenneth Jenkins, Jan Novakofski
ACCURACY, LIMITS AND TRADEOFFS IN
ULTRASOUND DISPLACEMENT ESTIMATES
AND THEIR APPLICATION TO TISSUE ELASTICITY ASSESSMENT .............. 363

Lars Ødegaard
EVALUATION OF PHASE ABERRATION CORRECTION DUE TO
DIFFRACTION OF THE ABERRATED WAVEFRONT; A SIMULATION STUDY ...... 367

Hans Torp
ULTRASOUND DOPPLER BLOOD FLOW MEASUREMENTS:
SIGNAL MODELS, AND SIGNAL PROCESSING METHODS .......................... 371

Zhang Yu Xia, Zhou Jian Hua
APPLICATION OF COMMON FAN ULTRASONIC DIAGNOSTIC
APPARATUS IN OCULAR DIADEASES ............................................... 375

Juha Ylitalo, Torgrim Lie, Hans Torp
COMPUTED SYNTHETIC APERTURE ULTRASOUND IMAGING:
APPLICATION TO MEDICAL DIAGNOSIS ...................................... 379

MUSICAL ACOUSTICS

Seji Adachi, Masa-aki Sato
BRASS SOUND SIMULATION USING A TWO-DIMENSIONAL
LIP VIBRATION MODEL .................................................................. 385

Ana Barjaui Condomines, Vincent Gibiat
STABLE AND UNSTABLE BEHAVIOURS IN WOODWINDS ....................... 389

Matthias Bertsch
ASPECTS OF TRUMPET PLAYING .................................................. 393

Antoine Chaigne
TRENDS AND CHALLENGES IN PHYSICAL
MODELING OF MUSICAL INSTRUMENTS ....................................... 397

Tong Chen
CONSTRUCTION AND FREQUENCY CHARACTERISTICS OF
CHINESE BOWED STRING INSTRUMENT ........................................ 401
Francesc Daumal i Domènech and Diana Möller Parera
TONAL ARCHITECTURE, SOUND CHARACTER SPACES ........................................ 405

Philippe Dérogis, René Caussé
COMPUTATION AND MODELISATION OF THE SOUND
RADIATION OF AN UPRIGHT PIANO USING MODAL
FORMALISM AND INTEGRAL EQUATORS ...................................................... 409

Pierre Dutilleux, Thomas Ruoff
AN EVENT RECORDING SYSTEM FOR THE AUTOMATIC
TRANSCRIPTION OF MUSIC PLAYED ON IDIOPHONES
APPLICATION TO THE GAMELAN GENDER WAYANG ............................. 413

R.-F. Fehlmann
FINITE ELEMENT SIMULATION
AND MEASUREMENTS OF CURVED SWISS HORN ........................................ 417

R.-F. Fehlman
FINITE ELEMENT SIMULATION OF CONICAL WAVEGUIDE ......................... 423

Helmut Frank, Wolfgang Vogl, Christoph Muss, Horst Wagner
COMPUTER OPERATED STRING
INVESTIGATION UNDER MUSICAL ASPECTS ............................................. 427

Anders Friberg, Johan Sundberg, Lars Frydén
THE KTH RULES FOR MUSICAL PERFORMANCE:
OVERVIEW AND RECENT ADDITIONS ..................................................... 431

Alexander Galembo
SOUNDS OVERLAPPING IN MELODY PLAYED ON A PIANO ......................... 435

Alexandra Hettergott
SPECTRAL VISIBILITY OF TIMBRAL DIFFERENCES IN TWO
REALIZATIONS OF A. SCHOENBERG'S ORCHESTRA PIECE OP. 16 NO.3 ......... 439

Kazuhiro Iida, Masayuki Morimoto
ABSOLUTE ANGLE AND JND OF AUDITORY
SOURCE WIDTH FOR A MUSIC MOTIF .......................................................... 443

Jyri Huopaniemi, Matti Karjalainen, Vesa Välimäki
PHYSICAL MODELS OF MUSICAL INSTRUMENTS
IN REAL-TIME BINAURAL ROOM SIMULATION ......................................... 447

Matti Karjalainen, Jyri Huopaniemi, Vesa Välimäki
DIRECTION-DEPENDENT PHYSICAL
MODELING OF MUSICAL INSTRUMENTS .................................................. 451

Armin Kohlrauch, Andreas Sander
HOW DOES THE BANDWIDTH OF A HARMONIC
COMPLEX-TONE MASKER INFLUENCE ITS MASKING BEHAVIOR? ................ 455

Neil P. McAngus Todd, Eric F. Clarke
THE PERCEPTION OF RHYTHMIC STRUCTURE
IN EXPRESSIVE MUSICAL PERFORMANCE ............................................. 459
J. Meyer
THE ACOUSTICS OF THE ORCHESTRA .................................................. 463

Kenji Ozawa, Yōiti Suzuki, Toshio Sone
A "MASKED FREQUENCY SPECTRUM" OF A
BROAD-BAND NOISE AND ITS TIMBRE ........................................... 465

Maria Pavlidou, Bernard E. Richardson
MODELLING THE STRING-FINGER
INTERACTION IN THE CLASSICAL GUITAR ..................................... 469

Eric Prame
STUDY OF VIBRATO EXTENT AND F0 VARIATIONS AMONG SINGERS .......... 473

Bernard E. Richardson
NUMERICAL MODELLING OF VIBRATIONS AND
SOUND RADIATION FIELDS OF STRINGED MUSICAL INSTRUMENTS .... 477

David Sharp, Arnold Myers, Raymond Parks and Murray Campbell
BORE RECONSTRUCTION BY PULSE REFLECTOMETRY AND
ITS POTENTIAL FOR THE TAXONOMY OF BRASS INSTRUMENTS .... 481

Laurent Simon, Jean-Christophe Veliere
SUBJECTIVE MEASUREMENTS OF DENOISED
MUSIC SIGNALS: REALISATION AND ANALYSIS ............................ 485

Julius O. Smith III
DIGITAL WAVEGUIDE MODELS FOR SOUND
SYNTHESIS BASED ON MUSICAL ACOUSTICS ................................. 489

Anatoll Stuvlov
FLEXIBLE STRING VIBRATIONS EXCITED BY
A NONLINEAR HEREDITARY HAMMER ............................................ 493

Sadaoki Furui
RECENT ADVANCES IN SPEECH PROCESSING TECHNOLOGY .............. 497

Tomoyasu Taguti
DESIGN OF PIANO PERFORMANCES IN STRUCTURED EXPRESSION ......... 505

Hisashi Uematsu, Kenji Ozawa, Yōiti Suzuki, Toshio Sone
A CONSIDERATION ON THE TIMBRE OF COMPLEX
TONES ONLY CONSISTING OF HIGHER HARMONICS ......................... 509

Vesa Välimäki, Matti Karjalainen, Jyri Huopaniemi
MEASUREMENT, ESTIMATION, AND MODELING
OF WIND INSTRUMENTS USING DSP TECHNIQUES ........................ 513

Gregor Widholm
BRASS WIND INSTRUMENT QUALITY MEASURED
AND EVALUATED BY A NEW COMPUTER SYSTEM ........................... 517
Howard A.K. Wright, Bernard E. Richardson
PSYCHOACOUSTICAL EVALUATION OF SYNTHESISED GUITAR TONES ............521

Sølvi Ystad, Richard Kronland-Martinet
A SYNTHESIS MODEL FOR THE FLUTE TRANSIENT BY
ANALYSIS OF REAL SOUND THROUGH TIME-FREQUENCY METHODS ..........525

AEROACOUSTICS, ATMOSPHERIC SOUND
AND INFRASOUND

Hans Bodén
LINEARITY MEASURES FOR IN-DUCT ACOUSTIC ONE-PORT SOURCES .......... 531

M. K. Bull, A. M. Blazewicz, Yi Li, J. M. Pickles
THE EFFECT OF FLOW VARIATIONS ON ACOUSTIC
RADIATION FROM A PLATE IN A VORTEX WAKE .................................. 535

J. E. Flowcs Williams
CAN HIGH SPEED FLOW BE SILENCED? .............................................. 539

Jyri Lavrentjev, Mats Åbom
MEASUREMENT OF SOURCE .............................................................. 543

Richards E. Musafir
DIRECTIVITY OF JET NOISE .............................................................. 547

O. Planade, J. Ducourneau, J. P. Guilhot
MODÉLISATION DE SILENCIEUX
PASSIFS EN PRÉSENCE D’ÉCOULEMENT ........................................... 551

A. Snakowska, H. Idczak
ON THE POSSIBILITY OF APPLYING A HARD-WALLED CIRCULAR
DUCT AS A REFERENCE SOUND SOURCE – THEORY AND MEASUREMENT .... 555

Igor P. Tchountchouzov
ACOUSTIC PROBING OF THE ATMOSPHERE ..................................... 559

Mats Åbom, Börje Nilsson
SOUND TRANSMISSION AT A SUDDEN AREA
EXPANSION IN A FLOW DUCT ............................................................ 563
SPEECH
SPEECH PERCEPTION DEFICITS IN NOISE: CONTRIBUTING FACTORS

Sharon M. Abel
Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada

SUMMARY

This paper reviews factors which contribute to speech perception deficits in noise. These include non-fluency, hearing loss, aging, speech-to-noise ratio, and the wearing of hearing protectors. Data are presented for a study in progress, which compares protectors with active noise reduction and conventional passive muff and plugs. Sixteen young normal-hearing subjects were tested. The results showed that all the devices examined resulted in improved performance, particularly for words embedded in sentences with poor contextual cues. The active muff provided no special advantage.

INTRODUCTION

Many well-controlled laboratory studies have documented the effects of hearing loss on the intelligibility of speech. In a study conducted in our own laboratory (Abel, Krever and Alberti, 1990), it was demonstrated that in quiet surroundings, subjects with moderate high-tone sensorineural hearing loss achieved consonant discrimination scores in a closed set test that were 10% lower than normal. This difference increased when the speech materials were mixed with speech spectrum noise. With an S/N of -4 dB, the scores for normal listeners had decreased by 15% and those for the hearing-impaired listeners, by approximately 30%. A similar pattern was noted in a word recognition test, in which subjects repeated the last word in sentences with either good or poor contextual cues. The results were correlated with the detection threshold for a 2-kHz pure tone. Aging (20-65 yrs) did not affect outcome. Aging has been shown to affect speech perception, when time processing abilities are involved (Bergman, 1980).

Hearing protectors are worn in the occupational setting to mitigate the injurious effects of high-level noise on hearing. An important consideration is whether these devices interfere with communication. In a study to answer this question, we explored the effects of hearing status, non-fluency with the language spoken, and the wearing of conventional passive hearing protectors on speech understanding (Abel, Alberti, Haythornthwaite and Riko, 1982). Lists of common English monosyllabic words were presented at levels of 80 and 90 dBA in quiet or in backgrounds of continuous 85 dBA-white noise or crowd noise. The subjects tested were either normal-hearing or hearing-impaired. Half of each group were fluent in English and half had learned English as a second language and were poorly conversant. Speech intelligibility decreased with a decrease in the speech-to-noise ratio, and was worse in crowd noise than in white noise. The hearing protectors caused a substantial decrement for the hearing-impaired but had no effect for normal listeners. For both groups non-fluency resulted in an additional decrement of 10-20%.
In a more recent study, we investigated the possible benefit for hearing-impaired subjects of active muffs with limited amplification of 5-10 dB for levels below 85 dB SPL (Abel, Armstrong and Giguère, 1993). Consonant discrimination and word recognition were studied in quiet and in a continuous background of impulsive noise presented at a level of 75 dB SPL. For normal-hearing listeners, amplification had a deleterious effect on both consonant discrimination and word recognition in noise. In the hearing-impaired group, the active muff resulted in less of a decrement than the passive devices, compared with unoccluded listening, in quiet but not in noise.

In a study currently in progress, we are comparing the effects on word recognition of a conventional passive muff and plug with high noise reduction ratings (Bilsom Viking muff and E-A-R foam plug), a plug with a low noise reduction rating (E-A-R HI-FI plug) and an active muff incorporating low-frequency noise cancellation (Pelto 7004 muff), tested in both the off and on modes. The relationship between speech understanding, hearing threshold and acuity for a change in stimulus frequency and duration is being investigated.

EXPERIMENTAL DESIGN

Three groups of 16 subjects are participating: two groups with normal hearing, aged 20-35 yrs. and 40-55 yrs., and one group with high-frequency hearing loss in the older age category. Each subject is tested with the ears unoccluded, and subsequently fitted binaurally with each of the three conventional passive protectors and the active muff with and without active noise reduction (ANR) operational. In each of these six ear conditions, measurements are made of hearing thresholds for one-third octave noise bands with centre frequencies ranging from 250 Hz to 8000 Hz, word recognition, the duration difference limen for a 200 ms one-third octave noise band centred at 4000 Hz, and the frequency difference limen for a 4000 Hz pure tone. Hearing thresholds and word recognition are assessed in quiet and in a continuous background of 75 dB SPL cable swager noise. The speech perception data for the younger normal-hearing group will be presented in this report.

METHODS AND MATERIALS

Subjects:
The subjects were 16 volunteers ranging in age from 21-29 years. In all, hearing threshold levels were equal to or less than 20 dB SPL and 25 dB SPL at 500 Hz and 4000 Hz, respectively. All subjects were fluent in English.

Apparatus:
The apparatus has been previously described (Abel, Armstrong and Giguère, 1993). Subjects were tested individually in a semi-reverberant chamber that met the requirements of the ANSI Standard for hearing protector testing (ANSI S12.6-1984). The speech materials used to assess speech perception were presented at a level of 80 dB SPL over a set of three Celestion DL10 three-way loudspeakers. In the noise background condition, speech and noise were mixed.
Procedure:
Speech perception was assessed by means of the SPIN test (Bilger, Nuetzel et al., 1984). This test comprises a set of eight alternative lists of fifty sentences pre-recorded on audio cassette by a male speaker. In half the sentences in each list, the final word is highly predicted from the context, and in half, poorly predicted. The subject's task is to write the last word in each sentence. In the present study, one list was presented for each of the twelve background by ear conditions. The unoccluded condition was presented first, followed by the five protected conditions, their order counterbalanced across subjects. Within each, quiet preceded noise. Across subjects, the eight lists were presented equally often within each listening condition.

RESULTS AND DISCUSSION
The results are presented in Table 1. The responses to each list were scored in three ways: total percentage correct (T), percentage of highly predicted words correct (H), and percentage of poorly predicted words correct (L). The data show that subjects' performance was relatively worse in noise than in quiet, regardless of the way in which the lists were scored. Within each background by ear condition, scores for the subset of poorly predicted sentences were relatively lower than those for the set of highly predicted sentences. The differential effect of the various ear conditions was particularly apparent for the poorly predicted items. In quiet, the results for the E-A-R HL-FI plug (EHF) were similar to normal. The conventional muff (BV) and the muff with ANR off (PF) or on (PN) resulted in a small decrement of about 5%, while the E-A-R foam plug (EF) resulted in a 17% decrement. By contrast, in the noise background condition, subjects performed more poorly in the unoccluded compared with any of the five protected conditions by 18% on average. The results for the latter appeared to be fairly similar.

The results of the preliminary analysis of the data presented above are in line with the conclusions of our previous studies. Normal-hearing subjects showed improvements in speech perception when they wore hearing protectors in noise. As the present study shows, this effect was particularly pronounced when contextual cues were absent. ANR did not appear to provide an advantage over the other devices, as suggested in other reports (McKinley and Nixon, 1993). Our finding is not surprising, given that the Peltor 7004 device chosen reduces sounds by approximately 5-10 dB below 500 Hz, i.e., below the speech frequency region. In quiet, attenuation had a deleterious effect. This was especially the case for the conventional E-A-R plug which, unlike the muffs, attenuated both high and low frequencies.

CONCLUSIONS
(1) In normal-hearing subjects, the wearing of hearing protectors resulted in an improvement in speech understanding under degraded listening conditions, i.e., when background noise was present, and contextual cues were poor.

(2) Hearing protectors with low-frequency ANR did not appear to be more beneficial than conventional muffs and plugs.
ACKNOWLEDGEMENTS

This research was supported by a contract from the Directorate of Health Protection and Promotion, Department of National Defence, Canada. The author is indebted to Ms. Valerie Hay and Ms. Deborah Spencer for their assistance with testing of subjects and analysis of data.

REFERENCES


Table 1. Word recognition scores as a function of ear condition and background. The effect of high (H) versus low (L) contextual cues.

<table>
<thead>
<tr>
<th>Ear Condition</th>
<th>Background/Score</th>
<th>Quiet</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>UN</td>
<td>93(5)</td>
<td>98(3)</td>
<td>88(10)</td>
</tr>
<tr>
<td>EHF</td>
<td>95(4)</td>
<td>100(1)</td>
<td>91(8)</td>
</tr>
<tr>
<td>EF</td>
<td>84(13)</td>
<td>96(8)</td>
<td>71(18)</td>
</tr>
<tr>
<td>BV</td>
<td>92(7)</td>
<td>99(2)</td>
<td>85(12)</td>
</tr>
<tr>
<td>PF</td>
<td>90(6)</td>
<td>98(3)</td>
<td>83(10)</td>
</tr>
<tr>
<td>PN</td>
<td>90(4)</td>
<td>99(2)</td>
<td>81(8)</td>
</tr>
</tbody>
</table>

*Mean percent correct (SD), N = 16
A SPEAKER VERIFICATION METHOD BASED ON THE
TALKER SIMILARITY SCALE DEFINED IN EACH PHONEME

Shuichi Arai, Ken'ichi Kido
Dept. of Computer Science, CHIBA Inst. of Tech., 2-17-1, Tsudanuma,
Narashino-shi, 275 Japan

SUMMARY

This paper describes a new talker verification system based on representing talker feature as
instantaneous and transitional spectral information of the particular phonemes. Recently, we
found out that the time transitional features of particular phoneme contexts are effective in talker
verification. Using this feature, talker characterization is carried out by some rules with the
similarity membership functions, until the decision rule is sufficient using the 7 words maximum.
The experimental result shows that this method achieves quite low error rate: false acceptance
rate is 0.3% and false rejection rate is 0.6%, and requires only 1.75 words utterance average (ap-
proximately 1 sec long) for testing 1 talker.

INTRODUCTION

Automatic talker recognition has long been studied, but the existing speech analysis meth-
ods cannot extract the individual speech characteristics exactly. The acoustical features rep-
resent the phonetic characteristics rather than the individual one. Therefore, it is main prob-
lem for talker recognition to eliminate the phonetic information from these acoustic features.
There have been proposed many talker recognition methods, and they are categorized into fol-
lowing two methods in this point of view. One method eliminates the phonetic information by
statistical averaging over every acoustic feature [2],[3],[4]. Another method eliminates by lo-
cating comparable phonetic events in a test utterance and the reference dictionary using the
speech recognition techniques [5],[6],[7]. In this method, acoustical feature space is divided into
some sub-spaces. Each sub-space has the specific phonetic features, so the phonetic feature
can be ignored in the same sub-space. The former method uses the statistical feature, there-
fore longer training material is required to char-
acterize a talker ( to eliminate phonetic vari-
ability ). On the other hand, the latter method
needs short time testing or training material be-
cause the phonetic information is uniform in the
sub-space that is assigned by the phoneme recog-
nition technique. However, the talker recogni-
tion performance rely on the performance of pho-
netic or phonetic-like classifier that eliminate
phonetic features. In general, performance of
the phoneme recognition is not sufficient now,
so we employ the text-dependent approach and
give the phoneme sequence of test utterances in
advance. The test utterance is divided into the
phoneme segments by DP matching using the
phoneme sequence and the delta cepstrum.
This method can extract the particular phoneme
segment explicitly, then this enables the higher
performance talker recognition that uses the dif-
fences of the individual information between
phonemes. Recently, we found out that some
Japanese phonemes contain more information
for individuality than other phonemes, so only
these phonemes are used in this talker recogni-
tion.

In practical use, the utterance time for en-
rolling and testing should be short. So, we de-
fine the talker similarity scales based on the dif-
fference of cepstral distance distribution between
intra-speakers and inter-speakers. Talker char-
acterization is carried out by some rules with
the talker similarity scales, so just seven words
utterances are required for training.

ACOUSTICAL PARAMETER

There have been already proposed various
acoustical parameters, however most of all rep-
resent the phonetic characteristics rather than
the individual one. Recently, we evaluated the
representation performance of talker individual-
ity for various acoustical parameters. This eval-
uation was performed under the following exper-
imental conditions. The acoustical parameters
are the pitch frequency, the formant frequency (1st.-3rd.), the gradient of spectrum and the various type spectral parameters derived from linear prediction. The test material is Japanese 50 words and are recorded 24 times during 1 year long by 10 males.

The result shows that every parameter can represent the talker individuality, but the performance is varied with the phoneme. Especially, Mel LPC Cepstral parameter (analysis order:32) has high performance for the individuality representation in the particular phonemes. The result shows in Fig.1. From this result, we decided to use just the vowel, nasal and fricative for characterizing the individual in this study.

**TALKER SIMILARITY**

In general, a long training and/or testing material is required to characterize a talker. But, long utterances are not comfortable for users in practical use. This is owing to characterize the individuality absolutely in the parameter space. So, we define the talker individuality relatively in the parameter space between intra and inter speaker. Generally, the distributions of inter-speaker and intra-speaker are spread over and overlapped because of phonetic information, so the distribution are calculated for every phoneme in this study. Fig.2 shows the distributions of the vowel /I/. The horizontal axis is the Cepstral Distance defined in Eq.1

\[
d = 10 \ln 10 \left( \sum_{i=1}^{p} (c_{p} - c_{q})^2 \right) [dB] \quad (1)
\]

Fig.2: Cepstrum distribution of intra and inter speaker

Using these distributions, the talker similarity scale \( M_p(d) \) is defined in each phoneme \( p \) as follows.

\[
M_p(d) = \begin{cases} 
1 & (d < d_1) \\
\frac{\int_{d_1}^{d_2} \min(h_p(d), t_p(d)) \, dd}{\int_{d_1}^{d_2} \min(h_p(d), t_p(d)) \, dd} & (d_1 \leq d \leq d_2) \\
0 & (d_2 < d) 
\end{cases} \quad (2)
\]

where \( h_p(d) \) and \( t_p(d) \) are the intra and inter speaker distribution respectively.

Fig. 3 and Fig. 4 are the talker similarity scales of /I/ and /n/ in each.

To obtain these similarity functions, many utterances spanning long term are required, but quite short utterances are required in enrollment and testing.

**SIMILARITY EVALUATION**

The block diagram to evaluate the similarity is shown in Fig.5. The talker similarity scales mentioned above are independent of the talker, so they can use for any talker. This talker similarity scales are computed for each phoneme in advance and stored in the verification system.
The test utterance is evaluated the similarity using this scale. This similarity scale is defined for each phoneme, so that the phoneme segmentation should be required. In this study, the test material is text-dependent so the standard templates that have the phoneme label $P_T(w, i)$ and the cepstral parameter $C_E(w, i)$ can be given. Using this standard templates, the phoneme segmentation for a input utterance is performed.

![Diagram](image)

**Fig.6: Segmentation and Time alignment**

The talker similarity $V_w$ is computed through the word using only the effective phonemes to represent the talker individuality.

$$V_w = \sum_{i \in P} M_p(d(C_E(w, i), C_T(w, j)))$$ (4)

where $P_e$: vowel, nasal and fricative mentioned above.

**VERIFICATION METHOD**

In this verification method, the Japanese seven words that contains the effective phonemes mentioned above. The system operation flow is shown in Fig.7.

![Diagram](image)

**Fig.7: System Operation Flow**

When the verification claim is occurred, the system requires the specify word utterance to
the user. The system computes the talker similarity \( V_w \), then judges accept, reject or ambiguous.

\[
\begin{align*}
V_w \leq th1 & : \text{reject} \\
th1 < V_w < th2 & : \text{ambiguous} \\
th2 \leq V_w & : \text{accept}
\end{align*}
\] (5)

If the talker similarity is ambiguous, the system requires another word utterance to the user. When the system cannot judge up to the last word, the average of \( V_w \) is used to judge comparing with the threshold \( th3 \).

**EXPERIMENTS and RESULTS**

Before testing the proposed method, the thresholds \( th1 \), \( th2 \), and \( th3 \) should be fixed. Various combination of threshold values were tested using 150 male testing sets and obtained Fig.8. From the result, \((th1, th2, th3) = (0, 95, 40)\)

![Fig.8: Error Rate on Various Thresholds](image)

Table 1: Verification Rate

<table>
<thead>
<tr>
<th>False Reject</th>
<th>False Acceptance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Fig.9 shows the completion rate for the required words. This result shows that only 1.75 words utterances are required on the average in testing a talker.

**CONCLUSIONS**

In this paper, we proposed the talker similarity scales for the particular phonemes that represent talker feature effectively. Using this talker similarity scale, the talker verification system was constructed and tested. The result shows that this method achieves quite low error rate: false acceptance rate is 0.25% and false rejection rate is 0.55%, and requires only 1.75 words utterances average for testing 1 talker even if just one time enrollment is needed.

**REFERENCES**


MEASUREMENT OF USER-SATISFACTION RATES OF THE
VOICE INTERFACE BASED ON WORD RECOGNITION AND
ESTIMATION OF ACCEPTABLE RECOGNITION ACCURACY

Yasuhiko Arai, Hisanori Kanasashi, Toshimitsu Minowa
Matsushita Communication Industrial Co., Ltd., Yokohama, Japan

SUMMARY

In order to argue about usability of the voice interface based on word recognition in the practical environments, the user-satisfaction rate is measured psychologically. It is made clear that the user-satisfaction rate is the linear function of the receiving entropy derived from the state transition diagrams of the voice interface. From that relation, acceptable recognition accuracy is estimated as 95%. And it is concluded that today's voice interface based on word recognition is usable in the limited environments where the 95%-accuracy is ensured.

INTRODUCTION

Although the word recognition accuracy is getting close to 100% in a laboratory, it is easily degraded in practical environments. It is our great concern whether today's voice interface based on word recognition is usable in practical environments or not. In order to make clear this question, we measured psychological user-satisfaction rate of the voice interface, and estimated the acceptable recognition accuracy.

The user-satisfaction rate was measured for two types of the interface. One was the type which accepted re-utterances in case of incorrect recognition (Type 1), and the other was that which required "yes" or "no" utterances according to correct or incorrect recognition (Type 2). In the latter type, the next candidate was obtained in an occasion of recognition failure.

EXPERIMENTS

Experimental conditions for measuring the user-satisfaction rate are shown in table 1. Every subject uttered 30 Japanese family names for the type-1 interface. When the interface failed to receive a correct family name, the subject re-entered the same family name up to 3 utterances. After all operations the subject marked his/her psychological satisfaction score on the 7-point rating scale ranging from -3 to +3. Experimental results are shown in table 2. All subjects are classified into 5 categories according to the numbers of correct words of the first utterances, such as 30-28(10 subjects), 27-25(9), 24-21(10), 20-17(9), and 16 and less(2). The recognition accuracy is averaged in each category. Numbers of all utterances including re-utterances and rejections are counted for the calculation of the recognition accuracy.

In the type-2 interface, the software simulator pretended to recognize a hundred Japanese city names in accuracy ranging from 60% to 100%. When the interface succeeded to receive a correct city name (as scheduled by the soft-
ware programming), the subject uttered "yes", and when it failed, the subject uttered "no" in order to obtain the next candidate. This procedure might be repeated until the third candidate was obtained. Every subject marked his psychological satisfaction score on the 7-point rating scale, every after a hundred operations for each accuracy condition. Experimental results are shown in table 3. When the actual recognition accuracy was degraded by re-utterances and rejections from the value scheduled, psychological satisfaction score for that operations was put together with that for the equivalent category (other rank of recognition accuracy condition with the same receiving entropy). Consequently, the number of accepted evaluation data in each category (the rank of recognition accuracy condition) became the value as shown in parentheses in table 3.

As shown in tables 2 and 3, it is clear that the user-satisfaction rate falls down as the recognition rate of the first utterance (or the first candidate) goes down, even if the ratio of correct words is virtually maintained by repeating utterances or requesting next candidates. Further, relationship between the accuracy and the user-satisfaction rate is different between two types of the interface.

RELATION BETWEEN USER-SATISFACTION RATE AND RECEIVING ENTROPY

Generally, recognition systems often fail to recognize entry words. Failure in recognition reduces the efficiency of the systems on receiving information. This reduction of the efficiency is measured by entropies.

It is found that the user-satisfaction rate is the linear function of the receiving entropy calculated from the state transition diagrams of the voice interfaces. And, it is considered that the user-satisfaction rate doesn't depend on the interface types as shown in figure 1. Therefore, following equation is obtained.

\[ S = 3.0 - 2.87E_r \]  

(1)

Where, \( S \) is the user-satisfaction rate, and \( E_r \) is the receiving entropy.

Assuming the second highest level (42) of the user-satisfaction rate is the acceptable boundary, the receiving entropy of 0.36[bit] is the lower limit of the acceptable interface performance.

ESTIMATION OF ACCEPTABLE RECOGNITION ACCURACY

When the receiving entropy (\( E_r \)) is given, corresponding recognition accuracy is calculated from the state transition diagrams by taking the recognition accuracy for the transition probability [1]. The recognition accuracy of the \( m \)-th utterance (\( A_{mth} \)) is expressed as \( A_{mth} = A_m - A_{m-1} \). Where \( A_m \) is the cumulative recognition accuracy up to the \( m \)-th utterance. And, \( A_m \) is experimentally obtained as follows.

\[ A_m = A_{m-1} [1+1.6067(1-A_{m-1})^{1.74}] \]  

(2)

It is same as for the recognition accuracy of the \( n \)-th candidate (\( B_{nth} \)), that is, \( B_{nth} = B_n - B_{n-1} \). And, the cumulative recognition accuracy up to the \( n \)-th candidate (\( B_n \)) is experimentally obtained as follows.

\[ B_n = 1-0.52(1-B_{n-1})(N-n)/(N-n+1) \]  

(3)

Where \( N \) is the number of recognition words. Using formulae (2) and (3), when the receiving entropy (\( E_r \)) is 0.36[bit], corresponding recognition accuracy
is obtained as $A_1=0.945 (94.5\%)$ for the type-1 interface and $B_1=0.950 (95.0\%)$ for the type-2 interface respectively. Therefore, approximately 95% is considered as the acceptable recognition accuracy of the voice interfaces.

In the practical environments, the recognition accuracy is easily reduced by noisy sound, room echoes, weaker voice, wrong speaking timing, and so on. In order to ensure enough user satisfaction, it is necessary to keep the recognition accuracy up to the value assumed above against the practical environments [2]. Therefore, strategic system design against noisy environments and human factors is required to achieve this subject.

CONCLUSION

Psychological user-satisfaction rates are measured for two types of the voice interface based on word recognition. Consequently, it is considered that the psychological user-satisfaction rate is the linear function of the receiving entropy. By using this relation, the lower limit of acceptable interface performance is assumed as 0.36[bit]. Further, the acceptable recognition accuracy is calculated from this figure, as it is approximately 95%.

In order to ensure enough user satisfaction, it is suggested that strategic system design for keeping up the 95%-accuracy against practical environments is necessary. And it may be given as a conclusion that today's voice interface based on word recognition is usable in the limited environments where the 95%-accuracy is ensured.

REFERENCES


Figure 1. Psychological user-satisfaction rate.
Table 1. Experimental conditions.

<table>
<thead>
<tr>
<th>Interface type</th>
<th>Type 1</th>
<th>Type 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognizer</td>
<td>Speaker independent isolated word recognizer. (Hardware system.)</td>
<td>Same as left. (Software simulator on PC.)</td>
</tr>
<tr>
<td>Recognition words</td>
<td>10 Japanese family names.</td>
<td>100 Japanese city names.</td>
</tr>
<tr>
<td>Recognition accuracy</td>
<td>36.7% to 96.7% (measured).</td>
<td>100%, 99%, 98%, 97%, 95%, 90%, 85%, 80%, 70%, and 60% (controlled by software).</td>
</tr>
<tr>
<td>Counterplan in case of failure</td>
<td>Re-utterance up to 2 times. (Maximum 3 utterances.)</td>
<td>Request next candidate up to 2 times. (Maximum 3 candidates.)</td>
</tr>
<tr>
<td>Subjects</td>
<td>40 subjects including male and female persons, age of 20 to 50.</td>
<td>11 male adults.</td>
</tr>
<tr>
<td>Number of words uttered for evaluation</td>
<td>30 words.</td>
<td>100 words for every accuracy condition. (total 1,000 words.)</td>
</tr>
</tbody>
</table>

Table 2. Experimental results in case of the type 1.

<table>
<thead>
<tr>
<th>Average recog. accuracy measured.</th>
<th>Receiving entropy [bit].</th>
<th>Psychological user- satisfaction rate.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st utterance. Up to 2nd. Up to 3rd.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9467</td>
<td>0.9967</td>
</tr>
<tr>
<td></td>
<td>0.8630</td>
<td>0.9667</td>
</tr>
<tr>
<td></td>
<td>0.7433</td>
<td>0.9600</td>
</tr>
<tr>
<td></td>
<td>0.6222</td>
<td>0.9259</td>
</tr>
<tr>
<td></td>
<td>0.4000</td>
<td>0.9167</td>
</tr>
</tbody>
</table>

Value in parentheses means the number of subjects ranked in that category.

Table 3. Experimental results in case of the type 2.

<table>
<thead>
<tr>
<th>Scheduled cumulative recog. accuracy.</th>
<th>Receiving entropy [bit].</th>
<th>Psychological user- satisfaction rate.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st candidate. Up to 2nd. Up to 3rd.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>0.95</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>0.90</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>0.85</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>0.80</td>
<td>0.89</td>
<td>0.94</td>
</tr>
<tr>
<td>0.70</td>
<td>0.84</td>
<td>0.92</td>
</tr>
<tr>
<td>0.60</td>
<td>0.79</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Value in parentheses means the number of data ranked in that category.
EFFECTS OF POSTVOCAL VOICING AND DISTINCTIVE VOWEL LENGTH ON SYLLABLE-INTERNAL TIMING: RESULTS FROM REAL- AND PSEUDO-WORDS IN NORWEGIAN

Dawn M. Behne, Bente H. Moxness
University of Trondheim, N-7055 Dragvoll, Norway

SUMMARY
The relative timing of consonants and vowels of a syllable can be affected by a variety of factors, including postvocalic voicing and distinctive vowel length. With both postvocalic voicing and distinctive vowel length, a vowel and postvocalic consonant duration are inversely related. When these factors converge on a syllable, the relative vowel and consonant durations realized in the acoustic signal simultaneously reflect the concurrent influence of both postvocalic voicing and distinctive vowel length. The goal of the present study is to characterize the concurrent effects of postvocalic and distinctive vowel length on syllable-internal timing of C1 VC2 components and examine whether syllable-internal timing distinguishes postvocalic voicing and distinctive vowel length. To this end, segment durations of relevant real and pseudo Norwegian words in context were measured and analyzed. Results demonstrate the expected inverse timing relationships between the duration of V and C2 for both distinctive vowel length and postvocalic voicing. In conjunction with the increased V duration and decreased C2 duration due to distinctive vowel length, the closure duration of C1 increased. However, with the same rhyme-internal pattern due to a voiced C2, the duration of C1 tended to decrease, or not differ, compared to a syllable containing a voiceless C2.

INTRODUCTION
The duration of syllable-internal components are known to be affected by diverse aspects of their linguistic context. Consequently the relative timing of vowels and consonants within a syllable may be affected. Among the factors known to affect segment durations in different languages are postvocalic voicing and distinctive vowel length which occur in Norwegian and constitute the focus of this project.

For a variety of languages, voicing of a postvocalic consonant can affect the duration of a preceding vowel. Vowels preceding a voiced consonant are typically longer than those preceding a voiceless consonant (e.g., [1][2][3][4][5]). In addition, a postvocalic voiceless consonant is generally longer than a postvocalic voiced consonant [6][7][8][9][10]. In his early work with Norwegian words such as takk [tak:] “thanks” versus tagg [tagg:] “thorn”, Fintoft [5] observed the effect of postvocalic voicing on the duration of a vowel in Norwegian. More recent observations of syllable-internal timing in Norwegian have substantiated these findings and have demonstrated the effects of voicing on the duration of a postvocalic consonant that have been widely observed for other languages [11].

Phonological distinctions can also be realized by means of vowel duration. Norwegian has traditionally been described as having a phonological distinction between short and long vowels. For example, the word takk [tak:] “thanks” has a phonologically short vowel compared to tok [tok] “hold” which has a phonologically long vowel. Accompanying this vowel length distinction is a difference in postvocalic consonant length. The phonotactics of Norwegian are such that, in a closed syllable, a distinctively long vowel is typically followed by a short consonant (VC), and a distinctively short vowel tends to be followed by a long consonant (V:C). For instance in the example above, the short vowel in [tok:] is followed by a long consonant whereas the long vowel
in [tɔːk] contains a short postvocalic consonant. This quantity distinction of Norwegian vowels [5][11] and corresponding postvocalic consonants [11] is also realized acoustically; phonologically long vowels are longer in duration than phonologically short vowels and in a closed syllable will be followed by a consonant which is shorter.

Previous research suggests that the pattern of effects associated with both postvocalic voicing and distinctive vowel length holds an inverse relationship between the duration of a vowel and postvocalic consonant. In fluent speech the relative timing of syllable-internal components reflect the concurrent influence of postvocalic voicing and distinctive vowel length. Does syllable-internal timing distinguish postvocalic voicing and distinctive vowel length? Previously [11] we have observed that the duration of a prevocalic consonant may differentiate the otherwise similar syllable-internal timing patterns associated with postvocalic voicing and distinctive vowel length. In the present investigation we examine this point further comparing results from real and pseudo words. The goal of this study is twofold: (1) to examine the effects of postvocalic voicing and distinctive vowel length on vowel and postvocalic consonant duration in Norwegian, and (2) to investigate the effects of these factors on prevocalic consonants.

**METHOD**

**Stimuli.** The 24 target words in Table 3 were used in this investigation. All target words were CVCs containing /s, t, d, n, / or /ts, dz, n/. Half of the target items were real Norwegian words in which the initial consonant was either a stop or a fricative, /t, d, s, h/ and the postvocalic consonant was /k/ or /g/. The other 12 targets were legitimate Norwegian pseudo words in which the initial stop was consistently /k/ and the postvocalic consonant was /t/ or /d/.

**Table 3. Target real and pseudo CVCs**

<table>
<thead>
<tr>
<th>Vowel Length</th>
<th>REAL WORDS</th>
<th>PSEUDO WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Postvocalic Voicing</td>
<td>Postvocalic Voicing</td>
</tr>
<tr>
<td>/k/</td>
<td>/tɔːk/</td>
<td>/tɔːk/</td>
</tr>
<tr>
<td></td>
<td>/tɪk/</td>
<td>/tɪk/</td>
</tr>
<tr>
<td></td>
<td>/tɪk/</td>
<td>/tɪk/</td>
</tr>
<tr>
<td></td>
<td>/tɪk/</td>
<td>/tɪk/</td>
</tr>
<tr>
<td></td>
<td>/tɪk/</td>
<td>/tɪk/</td>
</tr>
<tr>
<td></td>
<td>/tɪk/</td>
<td>/tɪk/</td>
</tr>
<tr>
<td></td>
<td>/tɪk/</td>
<td>/tɪk/</td>
</tr>
<tr>
<td>/g/</td>
<td>/pɪɡ/</td>
<td>/kɪt/</td>
</tr>
<tr>
<td></td>
<td>/pɪɡ/</td>
<td>/kɪt/</td>
</tr>
<tr>
<td></td>
<td>/pɪɡ/</td>
<td>/kɪt/</td>
</tr>
<tr>
<td></td>
<td>/pɪɡ/</td>
<td>/kɪt/</td>
</tr>
<tr>
<td></td>
<td>/pɪɡ/</td>
<td>/kɪt/</td>
</tr>
<tr>
<td></td>
<td>/pɪɡ/</td>
<td>/kɪt/</td>
</tr>
<tr>
<td></td>
<td>/pɪɡ/</td>
<td>/kɪt/</td>
</tr>
<tr>
<td>/d/</td>
<td>/kɪd/</td>
<td>/kɪd/</td>
</tr>
<tr>
<td></td>
<td>/kɪd/</td>
<td>/kɪd/</td>
</tr>
<tr>
<td></td>
<td>/kɪd/</td>
<td>/kɪd/</td>
</tr>
<tr>
<td></td>
<td>/kɪd/</td>
<td>/kɪd/</td>
</tr>
<tr>
<td></td>
<td>/kɪd/</td>
<td>/kɪd/</td>
</tr>
<tr>
<td></td>
<td>/kɪd/</td>
<td>/kɪd/</td>
</tr>
<tr>
<td></td>
<td>/kɪd/</td>
<td>/kɪd/</td>
</tr>
</tbody>
</table>

**Conversations.** A simple conversation consisting of a question and a response was developed for each target word. For instance, for the target word Skjåk, the question was *Hvor ligger Skjåk?* "Where is Skjåk situated?" and the response was *Skjåk ligger i Oppland* "Skjåk is situated in Oppland." For each target word the set of conversations was balanced to include the target word as focused and nonfocused in both initial and final sentence position.

**Subjects.** The subjects were 12 native speakers of Norwegian (6 males and 6 females) between 20 and 30 years old with no history of speech or hearing impairment. All participants used Trønder as their principle dialect. Trønder is the Norwegian dialect spoken in the region around Trondheim.

**Procedure.** Recordings were made of each subject producing the full set conversations for real and pseudo words with an experimenter in a sound attenuated room. For each conversation the
The experimenter asked the question and the subject read the response. Subjects were encouraged to speak at their usual speaking rate as if participating in a natural conversation.

**Measurements.** Three measurements were made within each target C1VC2 from subjects' responses in each conversation: (1) fricative/closure duration of C1, (2) vowel duration, and (3) closure duration of C2. Frication was measured from the beginning to the end of the aperiodic energy. Closure durations were measured from the start of the closure to the beginning of the release. Vowel duration was measured from the onset to the end of periodic energy.

**RESULTS**
Main effects of distinctive vowel length and postvocalic voicing on segment durations for target words are summarized in Figures 1-2 respectively. Accepted statistical differences (p<.05) are marked "*".

![Figure 1: Mean C1, V, and C2 durations for distinctively short and long vowels (V) in real (a) and pseudo (b) words.](image)

![Figure 2: Mean C1, V, and C2 durations for conditions with voiced and voiceless postvocalic consonants (C2) in real (a) and pseudo (b) words.](image)

**Distinctive Vowel Length.** Effects of distinctive vowel length on segment durations are presented in Figure 1. Findings show that the duration of C1 is longer before distinctively long vowels than before distinctively short vowels in both real and pseudo words [real: F(1,575)=44.52, p<.0001; pseudo: F(1,575)=4.01, p<.04]. As expected [3][11], the duration of distinctively long vowels is longer than the duration of distinctively short vowels [real: F(1,575)=552.87, p<.0001; pseudo: F(1,575)=436.43, p<.0001] and C1 is shorter following distinctively long vowels than following distinctively short vowels [real: F(1,575)=92.43, p<.0001; pseudo: F(1,575)=97.71, p<.0001].
These findings suggest that distinctive vowel length may be realized in the acoustic signal by the duration of the vowel, by the inverse relationship between the V and C₂ duration, and/or by the duration of C₁ and it's coordination with the duration of V.

Postvocalic Voicing. Figure 2 illustrates the main effects of postvocalic voicing on segment durations. In real words, which contained both stops and fricatives as prevocalic consonants, the duration of C₂ is shorter when the postvocalic consonant is voiced than when it is voiceless [F(1,575)=38.88, p<.0001], whereas in pseudo words which contained only stop prevocalic consonants, no reliable difference was observed [F(1,575)=2.69, n.s.]. Vowel duration was found to be longer before a voiced consonant than before a voiceless consonant in both real and pseudo words, which is consistent with previous research [5][11] [real: F(1,575)=62.80, p<.0001; pseudo: F(1,575)=15.94, p<.0001]. In addition, as has been observed previously [11], in both real and pseudo words C₁ is shorter when it is voiced than when it is voiceless [real: F(1,575)=171.60, p<.0001; F(1,575)=117.52, p<.0001]. These findings for postvocalic voicing support the expected inverse relationship between the duration of the vowel and postvocalic consonant. Furthermore, the results for C₁ suggest that it, too, may assist in cueing postvocalic voicing in the acoustic signal. Future investigations will examine this possibility further with focus on differences due to manner of articulation.

CONCLUSIONS

The results verify that distinctive vowel length and postvocalic voicing affect vowel duration and the timing of neighboring components within the syllable. Findings are consistent with previous research and suggest a global effect on the segment durations within a syllable. Although both postvocalic voicing and distinctive vowel maintain an inverse timing relationship between the duration of a vowel and postvocalic consonant within the rhyme, they differ in their effects on the duration of a syllable onset which may also distinguish their overall timing patterns.

ACKNOWLEDGEMENTS

This research has been supported by a grant from the Norwegian Marshall Fund to the first author.

REFERENCES

ACOUSTIC DETERMINANTS OF VOWEL IDENTIFY FOR GERMAN LISTENERS: TARGET SPECTRAL INFORMATION NOT NEEDED

Ocke-Schwen Bohn, Winifred Strange*
English Department, Kiel University, Kiel, Germany
* Dept. of Comm. Sciences and Disorders, Univ. of South Florida, Tampa, Florida, USA

SUMMARY
Two series of perceptual experiments tested the relative importance of two types of acoustic information (dynamic spectral and target spectral) in coarticulated German vowels. Separate groups of German listeners identified (experiments 1.1-1.5, reported in detail in Strange & Bohn 1995) and discriminated (experiments 2.1-2.5) naturally produced /d/-vowel-/n/-syllables which were electronically modified to manipulate the availability of formant target information (traditionally considered the primary information for vowel identity) and of dynamic spectral information. The results of the two series of experiments were very similar, even though different stimulus sets and testing paradigms were used. The most important finding was that German listeners identified and discriminated German silent-center syllables (in which only the dynamic portions of the syllable onsets and offsets were presented) very well, and that performance levels for vowel-center stimuli, in which vocalic nuclei were presented with onsets and offsets removed, were not significantly higher than for silent-center syllables. The results for German listeners identifying and discriminating German vowels are very similar to previously reported results for American listeners identifying American English (AE) vowel-center and silent-center stimuli. This provides important support for the Dynamic Specification Theory (Strange 1989), which states that vowels are specified by dynamic information defined over syllable onsets and offsets.

INTRODUCTION
In the traditional view of vowel perception, the target frequencies of the first two formants constitute the primary acoustic information for the perceptual identity of vowels. This Simple Target Model of vowel perception is inadequate because it fails to account for how listeners perceive speakers' intended messages in the face of various sources of variation in the acoustic signal. One kind of variation in vowel targets comes from coarticulation of vowels with consonants in consonant-vowel-consonant (CVC) syllables. Research by Strange and her collaborators (reviewed in Strange 1987, 1989) has shown that vowels produced in CVC syllables are identified with far greater accuracy than vowels produced in isolation, even though targets are often not reached in coarticulated vowels (Lindblom 1963). This led Strange to hypothesize that important information for vowel identity must be contained in the dynamic contour of the formants within the syllable.

Strange and her collaborators developed the Silent Center paradigm to test their hypotheses on the role of dynamic sources of information in vowel perception. Their methodology involves the systematic modification of CVC syllables to explore the perceptual relevance of various sources of potential information contained in these syllables. The methodologically most important modification in this paradigm is the generation of silent-center (SC) syllables, which are created by attenuating to silence the entire syllable nucleus, leaving only the initial and final transitions in their appropriate temporal relationship. The converse of SC syllables are vowel-centers (VCs), which are created by deleting the initial and final transitions, so that the syllable nuclei with target information are retained. Perceptual experiments employing SCs and VCs as stimuli allow one to test the perceptual importance of acoustic information associated either with the opening and closing gestures of the vocal tract in the production of CVC syllables (i.e., SCs), or with the (approximate) target configuration of the vocal tract in vowel production (i.e., VCs). - Experiments employing the SC paradigm typically also test the perception of initials (INIs) and
finals (FINs), which are, respectively, the initial and final transitions alone. This is done to test whether SCs in their entirety, or their initial or final part alone, contribute to perceived vowel identity.

Several studies have examined the perception of AE vowels by AE listeners in the SC paradigm and found high levels of identifiability for SCs. Strange (1989) concluded a review of these studies by stating that no single spectral cross-section adequately captures the perceptually-relevant information; rather, the acoustic information for vowel identity resides in the changing spectral structure. Because previous SC studies examined the perception of AE vowels, some ambiguity as to the nature of the dynamic information remained. According to Nearey's Compound Target Theory (CTT, Nearey 1989), the more or less diphthongized English vowels can be differentiated by contrasting patterns of vowel-inherent spectral change. Andruski & Nearey (1992) suggested that this information alone was sufficient to account for the perception of English vowels in SCs. Strange's Dynamic Specification Theory (DST, Strange 1989), on the other hand, states that vowels are specified by dynamic information defined over syllable onsets and offsets. The dynamic information reflects each vowel's characteristic opening and closing phases in their appropriate temporal relationship and style of movement of the vocal tract.

In order to examine the generality of previous findings on the perception of AE vowels in the SC paradigm, and to assess alternative hypotheses about the nature of dynamic information, the experiments reported here examined the perception of coarticulated German vowels. Like English, German has a large vowel inventory whose monophthongs differ in tenseness and/or length, but which have little or no diphthongization. Strange's DST predicts that dynamic spectral information plays an important role in the perception of vowels which have little or no diphthongization. Nearey's CTT, on the other hand, predicts that the importance of dynamic spectral information is restricted to the perception of diphthongized vowels.

METHODS

STIMULI: For experiments 1.1-1.5, the 14 monophthongs of North German (/ɪ/, /ɪ/, /ɛ/, /ɛ/, /æ/, /æ/, /ɪ/ /ɪ/, /ɛ/, /ɛ/, /ʊ/, /ʊ/, /ʊ/ /ʊ/) were recorded four times each by a male native speaker of North German. The vowels were produced in /dVt/ syllables in the sentence "ich habe /dVt/ gesagt" at the speaker's self-selected normal speaking rate. For experiments 2.1-2.5, the same speaker produced six tokens each of the vowels /ɪ/, /ɪ/, /ɛ/, /ɛ/, /ʊ/, /ʊ/ in isolated /dVt/ syllables. These vowels were selected for the discrimination experiments because the pairs /ɪ/-/ɛ/, /ɛ/-/ʊ/, /ʊ/-/ɛ/ and /ʊ/-/ʊ/ were most frequently confused in the identification experiments. From the digitized waveforms, measurements of target syllable duration, voice onset time (VOT) and fundamental frequency (F0) were used to make the final selection (criteria: stable F0, short VOT) of two instances of each of 14 vowels for experiments 1.1-1.5 and of four instances of each of six vowels for experiments 2.1-2.5.

SCs were generated by attenuating to silence the center portion of each of the target syllables, leaving onset and offset portions in their original temporal position. The onset and offset portions included the major part of the transitions associated with opening and closing gestures for all vowels - VC's were the converse of SCs. They were generated by attenuating to silence the onset and offset portions. FINs were generated by silencing both center and offset portions in each syllable. FINs were generated by silencing both onset and center portions of each syllable.

SUBJECTS: 121 subjects who met the selection criteria (no history of hearing loss according to self-report, native speaker of North German, limited exposure to languages other than German) were recruited from introductory linguistics courses at Kiel University and participated as unpaid volunteers. The mean age of the 87 female and 34 male subjects was 24 years (SD = 3.2).

PROCEDURE: For experiments 1.1-1.5, separate stimulus tapes were made to test the perception of unmodified syllables (experiment 1.1), VC's (experiment 1.2), FINs (experiment 1.3), INJs (experiment 1.4), and FINs (experiment 1.5). 2 instances each of 14 vowels were arranged in 4 random sequences and recorded onto audio tape. The response interval between utterances was 4 sec, with an 8 sec interval between blocks of 14 utterances. Groups of 10 (FINs) to 16 (SCs) subjects were assigned to one of the five conditions and group-tested in the language laboratory of Kiel University. The stimuli were presented via headphones, and subjects' responses were collected from response sheets which listed the response alternatives in German orthography.

For experiments 2.1-2.5, the stimulus syllables were redigitized on stored on the hard disk of a 486-PC. Groups of 10 subjects each were assigned to one of the five listening conditions (analogously to experiments 1.1.1.5) and tested individually in a sound treated chamber, where
stimuli were presented from a loudspeaker. Each subject was tested for discrimination of the contrasts /i/-/eI, /el/-/oI, /I/-/el, /oI/-/U/ in a pseudo-randomized order. Because the results of experiments 2.1-2.5 were to provide baseline data for a study of infant vowel perception (Bohn & Polka 1995), the adult subjects were tested in an age-appropriate version of the change/no change procedure. The subjects listened to presentations of the background stimuli (e.g., the four tokens of /dit/ for the /i/-/el/ contrast), and were instructed to raise their hand when they detected a change to the foreground (e.g., /det/ for the /i/-/el/ contrast). The ISI was set at 1.5 sec; a change trial consisted of the presentation of three foreground stimuli. Presentations of stimulus categories from the hard disk as foreground or background was counterbalanced within each group. The change from background to foreground was initiated by an assistant, who observed the subject through a one-way mirror. The software kept track of correct responses, false alarms, correct rejections, and misses throughout the 25 trials (15 changes and 10 controls) for each contrast.

**RESULTS**

Figure 1 gives the overall results for the five stimulus conditions in the identification experiments 1.1 - 1.5 with the full set of German monophthongs, expressed as percentage of correct responses (out of 112 opportunities) averaged across subjects within each group. The figure shows that overall performance in the unmodified, the SC and the VC conditions was much better than in INI and FIN conditions. SCs (mean % correct: 90) were not identified as well as unmodified syllables (mean % correct: 98.5), but vowel identity was nevertheless well maintained in most cases, despite the rather drastic modifications of the original syllables. Both VC (mean % correct: 84.7) and SC groups gave significantly more correct responses than the INI (mean % correct: 45.3) and the FIN (mean % correct: 49.5) groups. The SC and VC groups on the one hand and the INI and FIN groups on the other did not differ significantly from each other. The extremely high error rates in both INI and FIN conditions show that neither onsets alone nor offsets alone were sufficient to maintain vowel identity.

Figure 2 gives the overall results for the five stimulus conditions in the discrimination experiments 2.1 - 2.5 with the vowel contrasts /i/-/eI, /el/-/oI, /I/-/el, /oI/-/U/. expressed as percentage of correct responses averaged across subjects within each group. Discrimination levels for unmodified syllables (mean % correct: 97.5), SCs (mean % correct: 96.1), and VCs (mean % correct: 99.1) did not differ significantly. As in the identification experiments, vowel identity was well maintained in the SC condition, even though the vocalic nucleus with information on formant targets was not presented in that condition. All four vowel contrasts were discriminated highly accurately in the SC condition (96.5%, 95.6%, 95.6%, 96.8%) - Overall discrimination of FINs (78.3% correct) was significantly better than of INIs (61.2% correct), but more detailed analyses revealed that this was true only for contrasts differing in tenseness (/el/-/l/ and /oI/-/U/). FINs and INIs were discriminated equally well in the tense /i/-/el/ and the lax /U/-/eI/ contrast - Both FINs and INIs on the one hand were discriminated significantly less accurately than unmodified syllables, SCs, and VCs on the other.
CONCLUSIONS

The most important finding from the two series of experiments reported here was that German listeners identified and discriminated German vowels highly accurately when only dynamic spectral information of the syllable onsets and offsets in their appropriate temporal relationships was presented. The pattern of results for German vowels is very similar to that reported previously for AE vowels. For an adequate perceptual representation of vowels in these two languages with large vowel inventories, target spectral information is not necessary; rather, trajectory information specified over syllable onsets and offsets is a very good source of information for vowel identity in both German in AE.

The finding that German vowels presented as SCs were identified and discriminated highly accurately is incompatible with Nareyek's (1989) Compound Target Theory (CTT). The CTT predicts that distinct patterns of vowel-inherent spectral change contribute importantly to perceived vowel identity, at least the when the more or less diphthongized vowels of AE are presented as SCs. However, acoustic analyses of the German vowels presented in the experiments reported here revealed that all but one token of /e/ lacked diphthongization (s. Strange & Bohn, forthcoming). Formant movement, particularly for F2, was observed in many of the vowels, but this movement was associated with coarticulatory influences from the preceding and following alveolar consonants. This means that the dynamic sources of information which German listeners used so successfully in the SC conditions were not vowel-inherent but associated with the opening and closing gestures at the margins of the CVC syllables, as predicted by Strange's (1989) Dynamic Specification Theory.

Further studies are underway to establish the generality of these perceptual results with German vowels and adult German listeners. One series of studies in progress examines whether German vowels produced in multiple consonant contexts by multiple speakers are identified accurately when listeners are presented with SCs. Another series examines how the three types of acoustic information present in coarticulated German vowels (dynamic spectral, target spectral, and temporal) contribute to perceived vowel identity in prelingual German infants (Bohn & Polka, 1995). Preliminary results from these experiments suggest that almost all infants who can discriminate vowel contrasts with unmodified syllables can also discriminate the same contrast when vowels are presented as SCs.

ACKNOWLEDGEMENTS

This research was supported by a grant from the U.S. National Institutes of Health - NIDCD - to the second author and by a grant from the German Research Society - DFG - to the first author. The authors wish to thank Desiderio Saludes, Xiange Ling, and Sonja Trent for their assistance in generating stimulus materials, and Patrick Jung, Kirsten Schriever, and Anja Steinlen for assistance in testing subjects.

REFERENCES


AEREOACOUSTIC SOURCES IN A VOCAL TRACT MODEL

P.O.A.L. Davies, A. Barney*, C.H. Shadle**
Institute of Sound and Vibration Research, * Department of Electronics and Computer Science, ** University of Southampton, Southampton SO17 1BJ, UK

SUMMARY

Predictive models of phonation appropriate for inclusion in articulatory synthesizers include a relatively slowly time varying part, or acoustic filter, that is acoustically excited by a relatively rapidly time varying source. An investigation of the characteristic aeroacoustic sources associated with the voicing of open vowels was carried out on a simplified dynamic mechanical model with leading dimensions corresponding to those of an adult male. This consisted of a cylindrical tube open at one end and supplied at the other with a metered flow modulated by opening and closing of a pair of electromechanically driven shutters.

The flow distribution measured with calibrated hot wires provided clear evidence of periodic vortical flow synchronized with the shutter motion that travelled from its origin near the shutters to the open end, convecting at the mean flow speed. A similar flow structure has been observed in subjects voicing open vowels. Simultaneous calibrated pairs of wall pressure time histories defined the associated pressure field throughout the tract. Established techniques were then adopted to process the measurements and estimate the strength of the aeroacoustic sources associated with the potential (acoustic) and the rotational components of the flow in the tract and thus predict the sound radiated from the open end. The results were generally in close agreement with the corresponding free field measurements of the radiated sound spectrum at three representative mean flow speeds.

INTRODUCTION

Identification and quantification of the aerodynamic source mechanisms present during phonation are essential not only for high quality speech synthesis, but also for the correct interpretation of experimental results: these include inverse filtering and techniques involving measurement of pressures and air flows within the vocal tract. Likewise, experimental results, correctly interpreted, are necessary to validate any source mechanisms that are predicted theoretically. In this paper we use measurements made within a vocal tract-vocal fold model to identify and quantify the aerodynamic sources that constitute phonation.

The aeroacoustic factors concerned during the voicing of open vowels comprise the acoustic response of the vocal tract, acting as an acoustic filter, combining with the fluid dynamic features of the fluctuating glottal and vocal tract flow that provide the acoustic excitation. Generally it has been assumed that sound is normally produced at the glottis by the fluctuating volume velocity that
results from vocal fold vibration. However, this is not necessarily the only source [1], since other aeroacoustic sources may be present in the tract. For example, the relatively abrupt area expansion from the glottal opening to the vocal tract is known to produce a separating flow. This will result in the production of a vortical motion resulting in the presence of a rotational velocity field in the vocal tract. Rotational motion is by definition incompressible, so cannot be concerned in the propagation of sound along the tract. Nevertheless, it is also well known [2] that where they interact appropriately with boundary surfaces, rotational fields will produce sources of sound.

SOURCE MECHANISMS

McGowan [1] has suggested that the rotational velocity field associated with the separating flow out of the glottis can produce a dipole source there. He made an estimate of its relative contribution to the acoustic excitation, based on the observed pressure losses that accompany the flow through the glottis. It is also a well established feature of rotational flows that they persist for relatively long distances downstream. So it is not surprising that there is some experimental evidence [3] for their presence throughout the vocal tract. To investigate whether such additional sources exist, one needs more specific details of the fluctuating velocity and pressure distribution throughout the tract. A detailed experimental study to establish the fluctuating flow and pressure distribution throughout the vocal tract in vivo during speech seems hardly practicable. Therefore a mechanical model with leading dimensions corresponding to those of an adult male was chosen instead. This consisted of a cylindrical tube, open at one end and supplied at the other with a carefully controlled flow, modulated by the opening and closing of an electromechanically driven shutter. The resulting fluid motion in the tube was periodic, while the shutter motion provided a time reference.

The flow distribution throughout the model was measured with calibrated hot wires, and provided clear evidence of the presence of a periodic vortical flow throughout the tract that remained in concert with the shutter motion. The observations showed that the rotational disturbances grew rapidly, until they spanned the tube and were then found to convect along it at the velocity of the mean flow. Calibrated pressure measurements made at the walls of the tube defined the associated pressure field at a number of fixed points along the model tract. Corresponding sound radiation measurements were obtained with the model in a semi-anechoic chamber, with background levels well below those produced by the model.

SOURCE IDENTIFICATION

A four channel digital data acquisition system was used to acquire and store the data, with one channel allotted to record the shutter motion and so synchronise all the other records. The signals were periodic in concert with the steady shutter motion and sufficiently repetitive for harmonic analysis to establish stable harmonic component amplitudes up to the twelfth. These signal records could then be processed to describe the fluctuating flow and pressure distribution in the model tract for each harmonic, with the corresponding amplitude of the radiated sound. The signal records from pairs of suitably placed wall microphones can be processed [4] to define the acoustic pressure and velocity distributions throughout the model for each harmonic component in turn. These are first calculated as the forward and backward travelling acoustic component wave amplitudes given respectively by $p^+(x, f)$, $p^-(x, f)$. The corresponding acoustic velocity associated with the acoustic wave motion is then defined by
\[ \rho_0 c_0 u_0(x, f) = p^+(x, f) - p^-(x, f) \]  

(1)

One can then estimate the periodic acoustic velocity time history \( u_0(x, t) \) at any transverse plane \( x \) by summing the harmonic amplitudes resulting from equation (1).

The resulting acoustic velocity is compared in figure (1) with the corresponding velocity fluctuations measured with the hot wire. The obvious difference arises from the fact that the hot wire measurements include the contributions from the rotational flow also present in the model tract. To see how this comes about one can represent the fluctuating velocity field \( u(x, r, t) \) at any position by a sum of three components, defined as

\[ u(x, r, t) = u_0 + u_a(x, t) + u_s(x, r, t) \]  

(2)

where \( u_0, u_a \) and \( u_s \) are respectively the mean, acoustic (irrotational) and rotational (solenoidal) component velocities and \( x \) and \( r \) are respectively the axial and radial coordinate displacements. One should note that the hot wire measures the velocity magnitude \( |u_0| \), which it turns out did not appear to vary significantly across any plane in the downstream half of the duct. This is doubtless due to the influence of the image vortices necessary to satisfy the condition of zero normal (radial) velocity at the duct wall, combined with the fact that the frequencies measured were too low for the propagation of higher order modes, so that the acoustic wave motion remained plane.

The irrotational acoustic field in the duct is the result of excitation by acoustic sources associated with the fluctuating volume velocity of the shutter flow together with the source arising from the associated shed vorticity observed there. Predictions based on equation (1) of the sound radiated by this measured acoustic field with velocity \( u_0 \) [5] are compared with observations of the radiated field in figure 2a. One should note that the resulting estimates of the corresponding fluctuating acoustic particle velocities did not always correspond with the fluctuating velocities measured with the hot wires, since these include the contributions from the rotational flow. However, they did almost correspond for harmonics 4 to 6 which lie close to the first quarter wave acoustic resonance of the model tract. The discrepancies imply the existence of a further source associated with the fluctuating rotational velocity field \( u_s \) included in the hot wire velocity measurements. From the signal records one can estimate the vorticity associated with this field that accompanies the flow through the open termination [5]. If one includes the contribution from the associated source in the prediction of the radiated sound and compares this with the observations one obtains the results in figure 2b. The improved agreement at all but the 3rd harmonic is obvious.

CONCLUSION

The results suggest that further sources associated with rotational flow might contribute to the radiated sound during phonation. These may be situated at significant area discontinuities along the tract, for example the teeth.

ACKNOWLEDGEMENT

The financial support of the Leverhulme Trust is gratefully acknowledged.

REFERENCES


Figure 1  Comparison of the velocity fluctuations measured by a hot wire with the estimated acoustic velocity (dotted line).

Figure 2  Comparison of the measured radiated acoustic field with predictions (dotted line). (a) Predicted with acoustic velocity \( v_a \) alone; (b) Predicted with contribution from rotational field \( u_r \) included.
A LIVE SPANISH LANGUAGE EVALUATION
OF THE RASTI-METHOD

César Díaz & Concepción Velázquez
Departamento de Física Aplicada, E.T.S. de Arquitectura,
Universidad Politécnica de Madrid, 28040 Madrid, España

SUMMARY

The RASTI-Method is a simplified version of Speech Transmission Index, that has been evaluated in several languages. In this paper, a comparative study of the scores of Speech Intelligibility obtained by the traditional speaker-listener tests in Spanish Language and the RASTI-Method has been carried out in classrooms. The list of monosyllabic words used are phonetically balanced and representative of the Spanish Language. The percentage of understanding monosyllables for poor and bad RASTI-Index decreased more slowly when the words are pronounced without a prerecorded voice.

INTRODUCTION

The speech intelligibility is a subjective factor that measures the facility to understand a spoken signal emitted by a sound source and its valuation in a room decreases due to background noise and reverberation. The speaker-listener tests are a direct way of obtaining the speech intelligibility in a room. One method can be the following: a list of meaningless monosyllables (logatoms) phonetically balanced and representative of a specific language is read out in the room under test. The scores obtained for this procedure are influenced by the characteristics of the speakers and listeners, and a large number of tests are required to achieve reliable average results. Consequently, the method is very slow.

There are several procedures for the physical measurement of speech intelligibility \( 1^\text{-}10 \) (AI, SIL, AL\text{\_}consp, STI,\ldots). The RASTI-Method, a simplified version of Speech Transmission Index Method, quantifies speech intelligibility by means of Modulation Transfer Function (MFT). This method has been evaluated with speaker-listener tests \( 11 \) and an IEC \( 12 \) Standard RASTI provides an objective physical method for assessing speech intelligibility under conditions of steady noise and/or reverberation.

This paper deals with the results of a comparative study of speech intelligibility by speaker-listener tests in the Spanish Language and using the RASTI-Method. The tests have been carried out face-to-face, without prerecorded speaker voices and electroacoustical systems.
EXPERIMENTAL METHOD

Measurements of RASTI Index
The measurements have been carried out in four rectangular classrooms, occupied by 37, 22, 19 and 18 students respectively. To compute the RASTI Index we have used a Speech Transmitter, B&K, type 4225 and a Speech Receiver, B&K, type 4419. This picks up the signal and calculates the RASTI value in different places in the classrooms. The Speech Transmitter was placed 1.5 m above the platform, in the same place that occupies the speaker. The reception was performed in the places which occupy the listeners. In each place we have made 3 measurements with a period of 32 s each, in order to reduce random error. To obtain RASTI Index lower than 0.4 we have used Sound Source, B&K, type 4224, which emitted a wide band noise.

Sound pressure levels in the classrooms.
The background noise inside the classrooms ranges between 35-38 dB(A). During the performance of the tests, speaker-listener distances ranged between 2 and 12 m. Table 1 shows the different types of sound sources and the variation of sound pressure level inside the classrooms.

<table>
<thead>
<tr>
<th>Type of sound source</th>
<th>Range of sound pressure level inside the classrooms dB(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connecting sound source</td>
<td>61 - 69</td>
</tr>
<tr>
<td>B&amp;K 4224</td>
<td>61 - 64</td>
</tr>
<tr>
<td>Speakers</td>
<td>62 - 69</td>
</tr>
<tr>
<td>Speakers plus B&amp;K 4224 on</td>
<td>70 - 74</td>
</tr>
</tbody>
</table>

Speaker-listener tests
The tests were carried out by four speakers, two women and two men, they talked face-to-face in the classrooms, they did not used prerecorded voices and public address system. The listeners were 92 university students, 32 women and 60 men, with ages between 19 and 25 and an average of 21 years old. The listeners had normal hearing and they were not trained.

The speaker-listener tests were carried out by the following way: every test is composed of 100 meaningless monosyllables, distributed in groups of 10. The beginning of test lecture was preceded by a sentence. The rhythm of lecture was a monosyllabic word every four seconds. For every new test we have interchanged the listeners' position and the order of reading of monosyllabic groups. Ten different listeners have been in each point and the PB-word score is the average percentage of understanding monosyllables. The list used of meaningless monosyllables are phonetically balanced and representative of the Spanish Language.

EXPERIMENTAL RESULTS

Table 2 shows the range of PB-word score of speakers-listener tests, RASTI-Index and average equivalent signal-to-noise ratio (S/N) in the different points of measurement. The speaker-listener tests have been carried out ten time in each point. Figure 1 shows the relation between PB-word score (%) and RASTI-Index.

REFERENCES

A CHINESE TEXT-TO-SPEECH SYSTEM WITH HIGH INTelligibility AND HIGH NATURALNESS

Guan Dinghua, Chu Min, Lu Shinnan
Institute of Acoustics, Academia Sinica, 100080, Beijing, China

ABSTRACT

This paper presents a Chinese TTS system which is based on the concatenation of Chinese syllable wavelet segments and the utilization of the TD-PSOLA method. The system accepts Chinese text with a few prosodic markers as input, and generates fluent news broadcast style speech. The two main components of the system: the Chinese syllable wavelet segment dictionary and the prosodic model, are discussed successively. The result of a formal evaluation of the synthetic speech quality are also presented. The intelligibility and naturalness of the output speech of the proposed system are much higher than others. Rule-based waveform concatenation synthetic method can be used in Chinese speech synthesizer with rather high naturalness. Building up a suitable prosodic rule system is the key point for increasing the naturalness of synthetic Chinese speech.

1. INTRODUCTION

Research on the naturalness of synthetic Chinese has shown that fundamental frequency, duration, co-articulation and intensity are important factors. Among them the fundamental frequency and duration affect the naturalness most. The TD-PSOLA algorithm can modify the pitch, duration and intensity of wavelet segment in time domain with little distortion. We have built up a TTS system which is based on the concatenation of Chinese syllable wavelet segments and the utilization of the TD-PSOLA method. The system can produce rather natural speech, which can be accepted by most listeners.

There are many kinds of Chinese dialects in use today. Even for the Standard Chinese, people from different age groups or from different educational backgrounds speak differently. So we choose the CCTV news broadcast speech as the model of the TTS system and all the prosodic rules are based on the prosodic analysis of news broadcast speech.

A wide range of studies on Chinese prosody have been done. Wu studied the F0 patterns of polysyllabic combinations of Standard Chinese in details. Shen advances that the focal prominence are expressed by increasing the up-line of the pitch box. Based on these previous researches and prosodic analysis of CCTV news broadcast speech carried out in our laboratory, we setup a prosodic model for the TTS system. The prosodic model processes the input text on word level and sentence level.

2. OUTLINE OF THE TTS SYSTEM

The block diagram of the proposed system is shown in Fig 1. It works in real time on an ordinary PC-386 with a Sound Blaster board. The system accepts Chinese text with a few prosodic symbols as input data, and turns them into fluent news broadcast style Chinese. Prosodic markers uesed in this system are prominence marker and pause marker for very long sentences.
The two main components of the system: syllable dictionary and the prosodic model are discussed successively.

### 2.1 Creating the Syllable Dictionary

![Diagram](image)

- **Chinese test with prosodic markers**
  - Test analysis
    - Syllable dictionary
      - Target Fo contour
      - TD-PSOLA algorithm
        - Original wavelet segment
          - Pause insertion
            - Concatenation
              - Synthetic Chinese

Fig. 1 The block diagram of the TTS system

Reasons for selecting syllables as synthetic units are as follows:

1. Most Chinese morphemes are monosyllables and the speakers are very aware of syllable boundaries;
2. Chinese has a very small syllable set and most Chinese syllables have CV structures.

We choose monosyllables with tones (including light tone) as the synthetic units. There are about 1500 syllables in the syllable dictionary.

The syllable dictionary is based on an acoustics record made by a female broadcast announcer. All of monosyllables with the four lexical tones are articulated separately and the light tone syllables are articulated with a preceding syllable. The recording is digitized at 12 kHz sampling rate and 14 bit. The intensity and duration of each syllable are controlled carefully in order to make them sound identically. It is necessary to assign pitch marks to wavelet segments of voiced portions to meet the need of the TD-PSOLA method. All pitch marks are made manually to assure the accuracy.

Though the TD-PSOLA method can modify fundamental frequency, duration and intensity of wavelet segments, the scale of modification is limited. To assure the quality of the synthetic speech of the system, a synthesis-listening test has been carried out. First we made a 90 seconds news broadcast speech second of a female speaker whose average pitch is close to that of the syllable dictionary, and analyzed the pitch contour, duration and intensity of each syllable. Then we modified these prosodic features of syllables taken out from the syllable dictionary to those values and concatenated the modified syllables together. We got a rather natural speech. In the recorded speech the highest pitch is 342Hz and lowest is 103Hz. The longest syllable is 400ms and the shortest is 124ms. These ranges of F0 and duration cover the normal range of the speaker.

### 2.2 Prosodic Model

The F0 contours of the four Chinese lexical tones are put into a pitch grid to show their relations. The upper border of the grid is called up-line and the lower border of the grid is called base-line. We use the two-line model to describe intonation of Chinese speech.

There are two level processes in the prosodic model: word and combination level process and sentence level process.

On the word and combination level, F0 patterns of monosyllables, bi-syllabic combinations and tri-syllabic combinations, which are generated after the statistical analysis of a 10 minutes real-speech corpus and are normalized into the same pitch grid, are stored in the prosodic library. F0 patterns of poly-syllabic combinations which have more than three syllables are combinanted by F0 patterns of monosyllables according to special rules. There are three kinds of syllabic duration:

-32-
long, middle and short. Duration pattern of monosyllable is middle. For bi-syllabic combinations, if the accent is on the first syllable, the duration pattern is long-middle, otherwise the duration pattern is middle-long. For tri-syllabic combinations, the duration pattern is middle-short-long. The duration pattern of poly-syllabic combinations which have more than three syllables is short and middle interchanged. The last syllable is always long.

On the sentence level, the up-line of the pitch grid of the prominence combination rises. The base-line of each combination is a little lower than its precedence until it reach a more than 300 ms pause. After the pause the base-line goes up to its starting value and another falling period begins. Both the base-line and up-line of the last combinations of a sentence or a subsentence drop noticeably.

Here we give an example to show how the prosodic model works. We want to synthesize a sentence: “中国女子足球队卫冕成功” (zhong1 guo2 nu3 zi3 zi2 qiu2 dui4 wei4 mian3 cheng2 gong1). Each syllable is fetched from the syllable dictionary, the pitch contours of them are shown in Fig.2(a). First is the word level process. Use the bisyllabic $F_0$ pattern to “中国”, “女子”, “卫冕” and “成功”, and use tri-syllabic $F_0$ pattern to “足球队”. The duration pattern of “中国” and “女子” is long-middle, and that of “卫冕” and “成功” is middle-long. The duration pattern of “足球队” is middle-short-long. After word level process, the $F_0$ contour of the sentence is shown in Fig.2(b). Then begin the sentence level processing. The prominence of the sentence is “女子”, so the up-line of the combination rising. The base-line of each combination is lower than its precedence. Both the up-line and the base-line of the last combination “成功” drop noticeably. At this time, the final $F_0$ contour of the sentence is obtained as in Fig.2(c). Fig.2(d) shows the $F_0$ contour of sentence of natural speech. Fig.2(c) and Fig.2(d) are similar.

The system inserts different pauses automatically according to different kind of punctuation. For a long sentence without any punctuation, it is better to place a pause marker for breathing in the proper position to tell the system to insert a pause.

3. RESULT OF EVALUATION AND DISCUSSION

The system took part in a formal evaluation of speech quality of synthetic Chinese which is
held by the office of the State High Technology Development Project of China in May, 1994. Here we give out some results in Fig. 3. KX-PSOLA is the proposed system in this paper. KX-FSS is another TTS system of our laboratory, which is a formant synthesizer and uses the same prosodic model as KX-PSOLA. TH-SPEECH is a simple wavelet concatenation system, and CELP and VQ-LPC are two systems using LPC method. The intelligibility evaluation consists of syllabic clarity, word clarity and sentence intelligibility. The naturalness is only evaluated for sentences. There are altogether 16 listeners. The average intelligibility and sentence naturalness of the synthetic speech of KX-PSOLA are 94.1\% and 78\% respectively.

![Graph](image)

**Fig.3 Evaluation result of speech quality of Synthetic Chinese** (a) average intelligibility and sentence naturalness; (b) clarity of syllable and word, intelligibility of sentence.

In Fig. 3 (a), the average intelligibility of the two wavelet concatenation systems KX-PSOLA and TH-SPEECH is much higher than others, and the sentence naturalness of KX-PSOLA and KX-FSS, which have a good prosodic model, are higher. Both the average intelligibility and naturalness of KX-PSOLA are the highest. From Fig. 3 (b), we can see though the syllable and word clarity of KX-FSS is lower than that of TH-SPEECH, the sentence intelligibility of KX-FSS is higher than that of TH-SPEECH. This is due to the contribution of the prosodic model.

### 4. CONCLUSION

Rule-based waveform concatenation synthetic method can be used in Chinese speech synthesizer with rather high naturalness. Building up a suitable prosodic rule system is the key point for increasing the naturalness of synthetic Chinese speech.

Though the evaluation proves the high quality of the speech generated by KX-PSOLA, there are still obvious differences in naturalness between synthetic and natural speech. Improvements are being made in our laboratory.

### REFERENCES

EFFECT OF SPEECH PARAMETERS ON SPEAKER INDEPENDENT
WORD RECOGNITION BASED ON PHONEME-RECOGNITION AND
DICTIONARY

Titon Dutono, Naomitsu Ikeda*, Akira Watanabe*
Faculty of Engineering, Kumamoto University
* Yatsushiro National College of Technology

SUMMARY

This paper presents effects of speech parameters on speaker independent word recognition system based on phoneme recognition. The recognizer consists of a frame-based measuring distortion followed by dictionary-based "flag-DTW" as a decision algorithm.

In this study we tried to implement 43 parameters that consist of three parameters (RGB), i.e., the circular ratios of three formant frequencies that estimated by inverse-filter-control method, eight parameters (NN), i.e., the outputs of artificial neural network that trained by back propagation method in order to estimate eight phonemic classes, and 32 channels critical-band filter-bank (FB) that use Mel-scale for deciding center frequencies.

We used 32 standard patterns (i.e., 32 phoneme-like segments), each standard pattern consists of only a single template.

Recognition test was carried out by using 90 Japanese words separated into 3 groups of words, each spoken by independent 30 male Japanese native speakers. In applying only FB, for the top candidate, the recognition rate in average is 89.7%, moreover, by applying FB-NN and FB-NN-RGB, the average recognition rates are 93.4%, and 95.5% respectively.

INTRODUCTION

Speaker-dependent recognition has moved from the small vocabulary systems, 10-100 words, that were commercially available in the beginning of the 1980's, to the large vocabulary systems, 5,000-50,000 words, currently available. However, in the case of speaker-independent recognition, the number of words is still limited [1]. Therefore, Efforts for improving performance of the Speaker Independent Word Recognition still being an interesting issue.

Speaker Independent Word Recognition based on phoneme recognition and dictionary had been proposed by Nakatsu [2]. Such recognizer allows us to change the dictionary word easily without updating standard patterns that contain only phonemic-like segment patterns.

In general, the implementation of the large number of parameters is seen to be effective and leads to substantial improvement in the recognition performance. In this study we tried to implement 43 parameters that consist of three parameters (RGB), i.e., the circular ratios of three formants, eight parameters (NN), i.e., the outputs of artificial neural network that trained by back propagation method in order to estimate eight phonemic classes, and 32 channels critical-band filter-bank (FB).

RECOGNITION SYSTEM

The structure of the recognizer consists of a frame-based measuring distortion followed by a dictionary-based flag-DTW which has a role in measuring distortion score for each word written in the dictionary. The minimum distortion score implies the top candi-
date of the word decision.

**Feature estimation**

43 parameters estimation is made in each frame that has a duration of 20 ms at 12 kHz sampling rate and shifted 5 ms apart. As shown in Fig. 1, feature can be separated into three groups of parameter.

1. **Filter banks (FB)**
   - 32 channels critical-band filter-bank that use Mel-scale for center filter decision.

2. **Formant circular ratio (RGB)**
   - The circular ratios of first, second, and third formants. This conversion can be expected to normalize influence of vocal tract lengths. Therefore, it will be possible by this conversion to estimate normalized formants independent to age and sex of speakers [3]. In this study formant were estimated by inverse filter control method.

3. **Neural network (NN)**
   - The output of three layers neural network which is trained using utterances by independent speakers by backpropagation method to estimate five phonemic classes based on manner of articulation [NN1] (i.e., vowel, nasal, buzz-bar, burst and fricative) and three phonemic classes based on nature of the sound source [NN2] (i.e., voiced, unvoiced and silent).

**Phoneme-standard pattern**

In this study we used 32 standard patterns (i.e., 32 phoneme-like segments), for the simplicity of implementation, each standard pattern consists of only a single template. In order to create a single frame vowel standard template 60 tokens spoken by 20 male speakers, were single-frame fetched respectively, then averaging was carried out. In this study we applied 5 vowel standard patterns. The other 27 standard patterns (each of them was fetched from a word, spoken by a male speaker), considered has phoneme correspond to a standard pattern being created. Standard patterns represent voiced-stop and semivowel which have 5 ~ 10 frames long.

**Phoneme-distance matrix calculation**

For the frame distance between the input word and standard pattern phoneme-like segment in each group of the parameters we used Euclidian distance. The local distance calculation concerning a multi frames standard pattern phoneme-like segment has been carried out by dynamic programming. Two kinds of outputs in neural network i.e., manner of articulation based outputs (NN1) and nature of sound source based outputs (NN2), have been calculated independently. The distances are then linearly summed of four parameter groups. The frame distance can be written as

\[
D_{ij}(k) = \left( \sum_{m} (X_i^{(k)}(m) - S_{j,i}^{(k)}(m))^2 \right)^{1/2}
\]

where \(X_i^{(k)}(m)\) denotes input speech parameter at \(i^{th}\) frame, \(m^{th}\) parameter of \(k^{th}\). \(S_{j,i}^{(k)}(m)\) denotes \(j^{th}\) phoneme-like standard pattern parameter at \(f^{th}\) frame. \(D_{ij}(k)\) is calculated based on the following condition:

- \(k = 1, 2, 3, 4; M(1) = 3\) for RGB; \(M(2) = 5\) for NN1; \(M(3) = 3\) for NN2; \(M(4) = 32\) for FB.

The total local distance \(D_{ij}\) between \(i^{th}\) frame of input speech and \(j^{th}\) phoneme-like standard parameter will be
(2) Transition from /e/ to /a/ produces semivowel /y/.

SPEECH DATABASE
Recognition test was carried out by using 90 Japanese words spoken two times by 30 male Japanese native speakers. To evaluate the performance of recognizer, we divided 90 words into three groups of words.

RECOGNITION EXPERIMENT AND RESULT

To compute the total local distance \( D_{ij} \) in Eq. (2), appropriate weighting coefficients should be given for each group of parameter. For this purpose, we first combine FB and NN and used word group 2 as the vocabulary word. Based on preliminary experiment in a rather heuristic way, we fixed \( W_{FB} \) equal 1 and varied \( W_{NN} \) from 0 to 0.5. Fig. 2 shows a plot of recognition rate as a function of \( W_{NN} \). The maximum recognition rate is given when \( W_{NN} \) reach 0.1 and remain steady at \( W_{NN} \) equal 0.2. We fixed 0.2 as the weighting coefficient \( W_{NN} \). The same way is carried out concerning the combination of FB and RGB. Fig. 3 shows a plot of recognition rate as a function of \( W_{RGB} \). It seems that the maximum recognition rate at top candidate is given when \( W_{RGB} \) equal 0.1, however, by considering recognition rate at second candidate, the recognition rate become has maximum value when \( W_{RGB} \) equal 0.2. In order to decide which of the both value is an appropriate one, evaluation concerning the both value is carried out by applying all groups of parameter and \( W_{RGB} \) decided before as a fixed value. The result of this evaluation is shown in a table at Fig. 3. It can be seen that 0.2 is an appropriate value.

In order to evaluate the effect of combination in applying the group of parameter on recognition rate, we used three groups of vocabulary words and implemented weighting coefficient that have already decided. The condition of experiment is shown in the following table of experiment. Fig. 4 shows a plot of the growth in the recognition rate as a function of parameter group combination.
for three word groups. Consider FB parameter as a standard evaluation. In applying only FB, the average recognition rate is 89.7%. Moreover, by applying FB-NN and FB-NN-RGB, the average recognition rates are 93.4%, and 95.5% respectively. It can be seen that combination of parameter groups lead to performance improvement. Although each standard pattern consists of only a single template, framed from one speaker who has arbitrarily been selected, a good performance still can be achieved by implementing three groups of parameter.

**Table of Experiment**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NN,RGB</td>
</tr>
<tr>
<td>2</td>
<td>FB</td>
</tr>
<tr>
<td>3</td>
<td>FB,RGB</td>
</tr>
<tr>
<td>4</td>
<td>FB,NN</td>
</tr>
<tr>
<td>5</td>
<td>FB,NN,RGB</td>
</tr>
</tbody>
</table>

**CONCLUSION**

Effects of speech parameters combination on speaker independent word recognition system based on phoneme recognition and dictionary has been evaluated for three word groups. Results indicate that combination of FB, NN and RGB lead to performance improvement. Although the recognizer components (i.e., dictionary, standard patterns) were built by simple manner, a good performance still can be achieved. These three groups of parameters were taken from *Speech Training System* that have been developing in our laboratory, expected, this recognizer can be applied as an auxiliary device.

**References**


ACOUSTICS OF THE SINGING VOICE

Johan Sundberg
Department of Speech Communication and Music Acoustics,
KTH (RIT), Box 70014, S-10044 Stockholm

SUMMARY
This article reviews some recent contributions to the acoustical and physiological characteristics of singing. Singers accurately vary subglottal pressure in synchrony with both vocal loudness and pitch. Under conditions of loud singing at high pitches, singers tend to maintain a large amplitude of the transglottal airflow pulses by avoiding unnecessary increases of glottal adduction, thus retaining a strong voice source fundamental. Magnetic resonance imaging data of a singer's vocal tract confirmed that the singer's formant is generated by establishing acoustic mismatch between the larynx tube and the pharynx cavity. Magnetometer measurements of singers' jaw opening show that this articulator is used only for the vowels [aː] and [uː] to avoid the situation that the fundamental frequency exceeds the first formant frequency.

INTRODUCTION
The acoustics of the singing voice is an important area of music acoustics. The fact that the Voice Foundation recently arranged its 24th annual symposium Care of the Professional Voice, and that in September the first Pan European Voice Conference, PEVOC, will happen in London demonstrates high research activity. In this article some examples will be presented of recent investigations which have contributed to the acoustical understanding of singing in three different respects: breathing, phonation, and articulation.

BREATHEING
The phonatory function of the breathing apparatus is to supply an overpressure of air suitable for sound production. This subglottal pressure, is the main tool for loudness variation: the higher the subglottal pressure, the louder the sound.

Fig. 1 illustrates how subglottal pressure is typically varied in neutral speech and singing. In neutral speech the pressure pattern is simple; subglottal pressure is used merely to vary loudness of phonation, while emphasis is generally conveyed by F0 gestures and syllable duration. Furthermore, when vocal loudness is increased, mean F0 also tends to increase. The mean pitch increase, averaged across subjects, is about 0.4 semit/dB Lpr, a value to be expected if the increase of subglottal pressure were the sole agent underlying the pitch increase.
Fig. 1. Subglottal pressure in neutral speech (left) and in singing (right) according to Lieberman, 1967. The upper curves represent pressure captured by tracheal puncture and as the oral pressure during [p]-occlusion, the lower curves show F0.

In singing, by contrast, subglottal pressure is carefully varied with pitch (Cleveland & Sundberg, 1985; Leanderson et al., 1987). Pitch is raised by stretching the folds which thus become stiffer and need higher driving pressures. Note that in the graph the highest pressure was used not for the highest pitch but for the tone following after it. Thus, the singer made a crescendo culminating on this tone, which is the first tone in a new harmony. As pitch also depends on subglottal pressure, a failure to reach the target pressure will be manifest as a failure to reach target pitch, i.e., to singing out of tune. This exercise is a striking demonstration of the singer’s need for a virtuosoic control of the respiratory system.

VOICE SOURCE

The voice source is the pulsating airflow through the slit between the vocal folds, the glottis. Its acoustic characteristics can be conveniently captured in terms of the waveform of this air flow, the flow glottogram. The unique advantage of flow glottograms is that they represent the sound of the voice source in a physiologically realistic manner. Thus, the closed phase of the vibration cycle is represented by a horizontal part of the curve and the triangular pulses correspond to the open phase.

Flow glottograms vary considerably depending on the mode of phonation, see Fig. 2. The peak amplitude is a particularly interesting parameter. It is small in pressed phonation, large in breathy and in flow phonation, and intermediate in neutral. Physiologically is it varied by means of the degree of glottal adduction. Acoustically it affects the relative level of the voice source fundamental as shown in Fig. 3. Along with pitch and loudness, mode of phonation is a phonatory dimension; continuous variation is possible along all these three dimensions.

Fig. 4 illustrates what seems to be a typical voice source difference between nonsingers and singers which can be observed under conditions of loud phonation at high pitches. The graph shows log peak amplitude of the flow glottogram versus SPL for a singer and a nonsinger phonating at different degrees of loudness at low, medium, and high pitch. The solid reference lines represent a 1:1 relationship between this amplitude and sound pressure. The singer increased his peak amplitude at the same rate as the SPL. Acoustically, this means that he kept the balance between the fundamental and the lower overtones fairly constant. The nonsinger, on the other hand, did not increase the peak amplitude as much as he increased the SPL; on average, a 10 dB rise in SPL was associated with only a 4 dB gain of the fundamental, approximately.
Fig. 2 (left). Flow glottograms for a male subject who deliberately varied mode of phonation. $P_g$ is the subglottal pressure.

Fig. 3 (above). Relation between the peak-to-peak amplitude of the flow glottogram and the amplitude of the voice source fundamental. The data were collected from phonations of 4 singers singing the vowel [ae:] at different pitches.

Fig. 4. Relation between overall SPL and the level of the voice source fundamental in a singer and a nonsinger (right and left graphs) phonating at indicated F0 values in different degrees of vocal loudness. Solid and dotted lines represent a 1:1 and a 10:4 relationship.

Fant (1959) found exactly this relationship, represented by the dotted line in the graph, in an investigation of nonsinger voices. This suggests that the data shown in this figure are typical for untrained voices.

**SINGER’S FORMANT**

A typical acoustic characteristic of Western operatic singing is the *singer’s formant*, a high peak in the spectrum envelope of voiced sounds produced by male singers and altos, see Fig 5. It
improves the singer’s chances of being heard over a loud orchestral accompaniment (Sundberg, 1987). Acoustically it can be explained within the framework of classical theory of vowel production, if F3, F4, and F5 are assumed to be densely clustered along the frequency scale.

By experiments with real tube resonators it was demonstrated that such a formant cluster can be produced, if there is an abrupt change in cross sectional area between the outlet of the larynx tube and the pharynx. If so, an acoustic mismatch occurs between the pharynx and the exit of the larynx tube (Sundberg 1974). The larynx tube is about 2 cm long, its bottom is the vocal folds and it inserts into the pharynx at an angle, some 2 cm above the bottom. In case of acoustical mismatch, F4 is strongly affiliated with the larynx tube and is thus rather insensitive to changes of vocal tract shape outside the larynx tube.

Recently I had a chance to check this explanation by magnetic resonance (MR) imaging technique at the Iowa University. Fifty exposures were taken of a sung and a spoken version of the vowels i and a and combined to yield an area function. Fig. 6 shows the frequency sweep response of the area function for the sung [a] derived from an acoustic model where the cross sectional areas were arranged axisymmetrically. It can be seen that F3, F4 and F5 occur somewhat like a cluster.

To find out the sensitivity to errors in cross sectional area, the change of the formants due to a halving of the cross sectional area at different locations along the tract was measured. The results showed that the fine tuning of F3, F4, and F5 was particularly sensitive to the details of the area function in the larynx tube.

---

**Fig. 5 (left).** Spectrum envelope of the vowel [a:] spoken and sung by a professional singer.

**Fig. 6 (right).** Frequency response of an axisymmetrical acoustic tube model representing the area function of the vowel [a:] as sung by a male baritone singer. The area function was derived from magnetic resonance imaging of the vocal tract by Dr Brad Story, Iowa University.

Do these results support the interpretation of the singer’s formant as a resonance effect resulting from the clustering of F3, F4, and F5? To produce a clear singer’s formant, F3, F4, and F5 have to be clustered somewhat narrower than indicated in the figure, but small area function errors can have great effects on these formant frequencies. Furthermore, potentially important differences exist between the real vocal tract and the axisymmetrical models used for calculating the formant frequencies of the MR data. For example the sinus piriformes make the glottal end of the vocal tract far from axisymmetrical in reality. For a more accurate assessment of the interpretation of the singer’s formant as the result of a clustering of F3, F4, and F5 we have to await sweep frequency measurements of hi-fi, three dimensional models of the vocal tract.
FORMANTS AT SUPER PITCHES

As long as pitch is low, F0 is lower than the typical values of F1. At the high pitches occurring particularly in female singing, this is no longer true. For example, depending on the vowel F1 varies between about 300 and 800 Hz, while tenors’ and sopranos’ high C:s (C5 and C6) correspond to the frequencies of 523 Hz and 1050 Hz.

When the target pitch is higher than F1, singers tend to tune F1 to a frequency near that of the fundamental (Sundberg, 1975). This adds substantially to the SPL. This technique has been documented in sopranos but can be expected in the upper parts of the ranges of other singers’ too, at least in some vowels.

A main articulatory tool for increasing F1 is the jaw opening: the wider the jaw opening, the higher the first formant (Lindblom & Sundberg, 1971). In an X-ray investigation we found that two sopranos singing three vowels over a two octave range widened it systematically with pitch (Johansson & al., 1985). Recently the dependence of jaw opening on pitch and vowel was further analyzed in ten singers of different classifications (Sundberg & Skoog, forthcoming). A set of magnetometers were used to track jaw opening, as the singers sang ascending two octave scales on the vowels [a; aː; oː; uː; iː; eː]. Typical results are shown in Fig. 7.

![Formant and Jaw Opening Diagram]

Fig. 7. Jaw opening as function of pitch in a professional mezzo soprano observed during an ascending, two octave scale sung on the indicated vowels. Bars represent formant frequencies.

Most subjects kept the jaw opening constant up to the pitch were F0 was close to F1. For higher pitches the singers increased the jaw opening in the vowels [a; aː] while for the other vowels the jaw opening was kept constant up to much higher pitches.

These findings do not disprove the assumption that singers avoid raising F0 beyond F1. Rather they reflect the fact that for most vowels there are other articulatory means to raise F1 than by widening the jaw opening.
DISCUSSION AND CONCLUSIONS

Above some recent findings have been presented which add to our understanding of acoustical and physiological aspects of singing. Singers have been found to tune their subglottal pressure not only to regulate vocal loudness but also depending on the pitch. They avoid unnecessary increases of glottal adduction under conditions of loud singing at high pitches. They maintain the singer's formant by adjusting the larynx and the deep pharynx. At super pitches, they use the jaw opening or other articulatory means to avoid that F0 exceeds F1.

All these examples of special behavior in singers seem to reflect three basic needs: (1) a wide pitch range (2) a wide variety of voice timbre (3) a access to a wide dynamic range. The subglottal pressure patterns seem prompted by the use of a wide pitch range; high pressures are needed to produce high pitches, as demonstrated by Titze (1992). Singers' voice source behavior implies an orthogonalization of three main voice source dimensions, pitch, vocal loudness and mode of phonation. The gain is access to a wider range of voice quality variation which would allow a richer musical expressivity. The singer's formant is a resonatory phenomenon which not only increases the audibility of the singer's voice when accompanied by a loud orchestral accompaniment and thus is an example of replacing vocal effort by resonance. It also seems to serve as a timbral uniform cap for vowels, such that they all sound in some sense similar. This improves the singer's possibilities to mark the musical structure in the performance; timbral dissimilarity is a frequently used mean to represent structural boundaries. The principle of tuning F1 at high pitches increases the overall sound level of the sound produced and thus is another example of replacing vocal effort by resonance. Thus, all these examples of singers' special vocal behavior can be understood from a musical point of view.

How could singers invent all these smart solutions to complicated acoustical problems? The method has probably been trial and error applied over a long period. Those singers who happened to develop smart solutions became successful and hence they became ideals for their less successful colleagues.

By analysing singing we may learn how the voice can be used to produce loud sounds without excessive vocal effort. We also get an insight into the mysteries of musical expressivity and its role in music communication. Finally, we can see a number of inspiring examples of a smart and subtle use of acoustic principles.

REFERENCES


Sundberg, J (1975) "Formant technique in a professional female singer", Acustica 32, 89-96.


ONDELETTES ET PARAMÉTRISATION DU SIGNAL DE PAROLE EN MILIEU BRUITÉ

Christophe Gerard, Marc Baunry
Laboratoire d’Informatique de l’Université du Maine
Avenue O. Messiaen, B. P. 535, Le Mans 72017, Cedex France

ABSTRACT

In the field of automatic speech recognition in noisy environments, much research is currently focusing on signal parameterization. Among the time–frequency representations usually employed, the wavelet transform presents attractive time and frequency localization properties.

The parameterization studied is based on a set of analytic wavelets distributed over the Mel scale and then followed by a reverse cosine transform. One objective is to exploit the good time localization of the wavelet to model transient phenomena as well.

In this first comparative research, localized maxima of wavelet coefficients are used for every frequency analysed. Such maxima enable to keep optimal accuracy on the modelized pertinent event.

This wavelet parameterization was evaluated on a recognition system based on dynamic time warping methods. Recognition results were compared with MFCC parameterization (Mel Frequency Cepstrum Coefficients). Our first conclusions on the global corpus show that wavelets are as noise–resistant as the MFCC. A better modelization of transient feature phonetic classes might be revealed when a more suitable recognition protocol is used.

Keywords: speech, wavelet, parameterisation
INTRODUCTION
En parole, l'étape de paramétrisation conditionne les performances des systèmes de reconnaissance ; une difficulté supplémentaire est d'obtenir une représentation du signal robuste au bruit [BAUDOIN 93]. L'architecture des systèmes de reconnaissance basés sur des techniques d'alignement temporel (DTW) ou de chaînes de Markov (HMM) impose des représentations paramétriques à fenêtre de taille fixe (STFT, LPC, MFCC).Cette contrainte se heurte à la structure interne du signal de parole composé de segments acoustiques de durées variables. Si la modélisation des sons voisés aux caractéristiques spectrales stationnaires est acceptable, celle des phonèmes à caractère impulsionnel comme les plosives pourrait être améliorée par une autre approche [MALBOS 95].
Une nouvelle représentation du signal de parole mieux adaptée aux phonèmes rapides est proposée. Ses performances seront évaluées en milieu bruité par rapport à la paramétrisation de type MFCC.

ONDELETTES ET PAROLE
L'information pertinente du signal de parole est répartie en temps et en fréquence de façon variable. Temporellement, les durées des segments homogènes varient de quelques millisecondes pour les phonèmes transitoires des plosives jusqu'à plusieurs centaines de millisecondes pour les sons voisés. Du point de vue fréquentiel, la prise en compte de phonèmes psychacoustiques (échelle perceptive des hauteurs et décomposition en bandes critiques) conduit à de meilleures représentations. Ces contraintes proscrittent donc l'usage d'ondelettes orthogonales. En effet, les ondelettes comme celle de Daubechies serait inadaptée en raison de leur décomposition fréquentielle en octaves [DAUCHECHIES 92].

L'ondelette oblique de Morlet \( \psi(t) = e^{-\frac{t^2}{2}} e^{imt} \) a été choisie car elle minimise le critère d'incertitude d'Ilieisenberg-Gabor. Le plan fréquentiel est décomposé en 24 bandes critiques à partir d'ondelettes de la forme \( \psi_{ab}(t) = \frac{1}{\sqrt{a}} \psi \left( \frac{t-b}{a} \right) \quad a > 0. \)

Cette décomposition présente quelques particularités : les fréquences centrales sont réparties selon l’échelle Mel et les largeurs de bande selon l’échelle Bark [ZWICKER 89]. Afin de réduire les temps de calcul, un algorithme de calcul rapide de la décomposition en ondelette de Morlet a été utilisé [BARRAT 91].

PARAMÉTRISATION
Afin de valider notre paramétrisation dans un systèmes de reconnaissance, nous nous sommes replacés dans le cadre d'une analyse à fenêtre fixe de 10ms. Notre paramétrisation appelée Mel Wavelet Maxima Coefficients (MWMC) vise à exploiter les propriétés de bonne localisation temporelle de l'ondelette. Nous avons donc choisi de modéliser chaque fenêtre de signal par le coefficient d'ondelette le plus énergétique dans chaque bande de fréquence. Ainsi, les phénomènes stationnaires aussi bien que les transitions rapides sont représentés précisément. Ceci est du au fait que la fenêtre d'analyse n'engendre pas de dégradation du type lissage temporel. On remarquera aussi que la représentation est invariante à une petite translation temporelle de la fenêtre ce qui confère à notre paramétrisation une plus grande stabilité.
Afin de dresser une étude comparative avec la paramétrisation MFCC, une transformation en cosinus inverse est également réalisée sur notre paramétrisation. La dimension finale du jeu de paramètres est ainsi réduite à 10 coefficients MWMC. Les figures 1 et 2 représentent la localisation du coefficient le plus énergétique dans chaque bande. On y visualise une fenêtre de signal (a) et l’analyse correspondante (b).

![Figure 1. son voisé /a/](image1)

![Figure 2. son occlusif /k/](image2)

**Évaluation**

Le protocole d’évaluation commun a été établi au cours d’un précédent groupe de recherche coordonnée sur la comparaison des méthodes de paramétrisation [GERARD 94].

**Corpus**

Le corpus est constitué des chiffres prononcés en français par quatre locuteurs (deux masculins, deux féminins) soit au total 1200 mots extraits de la base EUROM0. Trois sources de bruit ont été sélectionnées dans NOISE-ROM0 : un bruit blanc, un bruit de voiture et un bruit de foule. On génère les signaux définitifs par superposition d’un signal et d’un bruit selon le rapport signal sur bruit désiré.

**Protocole et système de reconnaissance**

La reconnaissance monlocuteur porte sur des mots isolés non segmentés. Le corpus de référence est constitué d’une seule occurrence de chaque chiffre. Le corpus test est constitué de tous les autres éléments de la base. Les signaux tests et références ont été bruités avec différents types de bruits pour plusieurs rapport signal à bruit. Un score moyen est obtenu à partir des résultats associés à chaque locuteur.

Le système de reconnaissance de mots enchaînés utilisé est basé sur les techniques d’alignement temporel et d’analyse grammaticale. La grammaire utilisée est du type silence-mot-silence.

**Paramétrisation MFCC**

Cette paramétrisation est constituée d’un jeu de 10 coefficients MFCC obtenus à partir d’un banc de 24 filtres répartis selon l’échelle Mel. L’analyse est effectuée sur une durée de 20 ms avec un pas temporel de 12 ms. Une fenêtre de Hanning et une préaccentuation avec un coefficient de 0.90 sont utilisées [MOKBEL 92].

- 47 -
RESULTATS

Le tableau 1 résume les performances des paramétrisations expérimentées. Les taux moyens de reconnaissance pour les trois bruits utilisés sont représentés pour les deux paramétrisation évaluées.

<table>
<thead>
<tr>
<th></th>
<th>bruit blanc</th>
<th>bruit de voiture</th>
<th>bruit de foule</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSB (dB)</td>
<td>30 20 10 0</td>
<td>30 10 0 -10</td>
<td>30 20 10 0</td>
</tr>
<tr>
<td>Paramétrisation MFCC</td>
<td>99 98 96 87</td>
<td>99 99 99 97</td>
<td>99 97 93 59</td>
</tr>
<tr>
<td>Paramétrisation MWMC</td>
<td>99 98 95 86</td>
<td>99 99 99 97</td>
<td>97 97 92 60</td>
</tr>
</tbody>
</table>

CONCLUSIONS

L’analyse centiseconde du signal de parole établit un compromis entre une sur-segmentation des sons voisés et une sous-segmentation des phonèmes à caractère impulsionnel. Si la sur-segmentation peut être gérée par le système de reconnaissance, la sous-segmentation équivaut à une perte d’information pouvant être dommageable. Notre paramétrisation basée sur l’utilisation des maximums locaux dans chaque bande tente de remédier à ce problème. L’idée générale est de localiser au sein d’une fenêtre et pour chaque bande de fréquence, l’événement le plus énergétique, supposé le plus pertinent.

Les résultats actuels confirment la robustesse les MWMC mais ne permettent pas de mettre en évidence les points forts de notre paramétrisation mieux adaptée aux phénomènes à caractère impulsionnel. En effet le rôle des plosives dans l’identification des éléments du corpus des chiffres reste minime. Notons aussi que la reconnaissance globale et l’alignement temporel utilisés ne permettent pas de mettre en exergue une meilleure représentation des plosives.

Un autre protocole de reconnaissance plus adapté permettra de déterminer l’apport en reconnaissance d’une meilleure modélisation des classes phonétiques présentant des variations rapides. Quoi qu’il en soit certains résultats montrent que les représentations en ondelettes représentent une alternative intéressante en analyse centiseconde du signal.

REFERENCES

[BARRAT 91]

[GERARD 94]

[MALBOS 95]

[MOKBEL 92]
C. MOKBEL, "Reconnaissance de la parole dans le bruit : bruitage / débruitage", thèse de doctorat, ENST Paris, 1992

[ZWICKER 81]
COMPARISON BETWEEN TIME-SCALE AND TIME-FREQUENCY APPROACHES FOR THE DETECTION OF STRESS IN CONSONANTS

Bruno Gramatica, Robert Ruiz, Claude Legros
Laboratoire d'Acoustique, Université Toulouse-Le Mirail, 5 allées Antonio Machado, 31058 Toulouse Cedex, France

SUMMARY

The modification of the voice is a great problem for vocal recognition systems. Moreover, voice analysis is a non-intrusive operation that allows to know the emotional state of a speaker. We tried to determine if the analysis of consonants could permit to reach this goal.

After a short presentation of the paper and of the previous researches, we give a theoretical formulation of the Wavelet transform and of the smoothed pseudo Wigner-Ville transform which shall be used herein. We then show an example of the graphs we obtain during the rest period and under stressful conditions.

We compare the two methods for the determination of pertinent cues helpful to the detection of stress. The first results are presented here, showing many differences between the two methods and between the rest period and the stressful one. The great difficulty of the interpretation of these time-scale and time-frequency diagrams is discussed.

INTRODUCTION

The aim of this research is to put in evidence spectro-temporal changes in the vocal signal of a pilot under stressful conditions. We suppose that voice is modified by the apparition of a flight incident which leads the pilots to be under stress. We developed some methods to analyze the vowels parameters, particularly with the time contour of the fundamental frequency [1], and the Cumulative Spectral Probability Diagram [2]. Several authors were interested in vowels modifications, with spectrographic analysis and pitch extraction [3,4]. They show that vowels are sensitive to stress.

Our problem is to know if the consonants are also modified under stressful conditions. We use signal processing tools such as the wavelet transform and the Wigner-Ville transform. After the presentation of their theoretical formulation, we present and comment an example of each of them during the rest period and the stressful one.

THE WAVELET TRANSFORM

Theoretically formulated in the 80's by J. MORLET & A. GROSSMANN the wavelet transform is found to be a very useful tool for signal processing applications. It allows to observe signals with different precisions.

Let $\psi(t)$ be the wavelet. The wavelet transform of a signal $x(t)$ is then defined as :

$$W_{a,b} = \langle x(t)|\psi_{a,b}(t) \rangle = \int x(t)\psi^*_{a,b}(t)dt \quad \text{with} \quad \psi_{a,b}(t) = a^{1/2}\psi\left(\frac{t-b}{a}\right)$$

- 49 -
\( \psi(t) \) is the mother wavelet. It is translated in time using the translation parameter \( b \) (to select the part of the signal to be analysed) and dilated or contracted using the scale parameter \( a \). The \( W_{a,b} \) coefficient gives an idea of the content of the signal at time \( b \) and at scale \( a \).

Although it is not admissible (\( \hat{\psi}(0) \equiv 10^{-3} \)), we use the MORLET's wavelet which is given by:

\[
\psi(t) = \pi^{-1/4} e^{i\omega t} e^{-t^2/2}
\]

The two parameters \( a \) and \( b \) can take discrete values [5]:

\[
a = 2^{-j} \quad \text{and} \quad b = 2^{-j}k \quad (j, k \in \mathbb{Z})
\]

then \( \{ \psi_{j,k} \}_{j,k \in \mathbb{Z}} = \{ 2^{j/2} \psi(2^j t - k) \} \) and

\[
x = \sum_{j,k \in \mathbb{Z}} < x \mid \psi_{j,k} > \psi_{j,k}
\]

We then obtain a time scale plane with \( j \) and \( k \) as coordinates respectively of scale and translation. Taking this formulation, we obtain HILBERT's bases of wavelets [6]. We choose to compute the wavelet transform on 4 octaves with 10 voices by octave. It allows to have a frequency scale where the scale 40 represents 5000 Hz and the scale 10 is 625 Hz. In each octave band, the frequency precision is of \( 1/10 \) octave. The scalogram \( |W_{a,b}|^2 \) of the word "OK" is shown below.

Figure 1 represents the rest period and figure 2 the stressed one computed thanks to NI!Power software.

Figures 1 and 2: Wavelet transform of the word "OK" during the rest period (left) and the stressed one (right).
THE SMOOTHED PSEUDO WIGNER-VILLE TRANSFORM

Introduced by E.P. WIGNER and J. VILLE the Wigner-Ville transform was theoretically studied in CLAASSEN & MECKLENBRÄUKER [7], and has been found to be in the heart of time-frequency theories. For the spectral and temporal resolutions to be independant, we have to use a two degree of freedom analysis in taking a smoothing function which can be separated in its two variables [8]:

\[ F_p(t, v) = g(t) \cdot Q(v) \]

Let \( z_1(t) \) be the analytic signal associated to the real signal \( x(t) \). The smoothed pseudo-WIGNER-VILLE Distribution (SPWVD) is defined as:

\[ PW_{\alpha}(t, v) = \int_{-\infty}^{\infty} e^{-2\pi i \tau} q(\tau) \int_{-\infty}^{\infty} g(u-t) \cdot z_1 \left( u + \frac{\tau}{2} \right) \cdot z_1^* \left( u - \frac{\tau}{2} \right) \, du \, dv \quad \text{with} \quad q(\tau) = \hat{Q}^{-1}(\tau) \]

For the computation, we take the discrete time version of the SPWVD:

\[ PW_{\alpha}(t, v) = 2 \sum_{t=-N+1}^{N-1} |h_N(\tau)|^2 \times \left[ \sum_{k=-M+1}^{M-1} g_M(k) \cdot z_1 \left( t + k + \tau \right) \cdot z_1^* \left( t + k - \tau \right) e^{-i2\pi k \tau} \right] \]

\( h_N \) and \( g_M \) are respectively the weighting window and the smoothing window.

Our speech signal is sampled at 10 kHz. We choose the weighting window to contain \( N=151 \) samples and the smoothing window \( M=21 \), which leads us to have 2.1 ms of temporal precision and \( \Delta f = \frac{1}{NT} \approx 66 \text{ Hz} \).

The figures 3 and 4 represent the same words than the figure 1 and 2.

![Figure 3 and 4: Smoothed Pseudo Wigner-Ville Distribution of the word "OK" during the rest period (left) and the stressed one (right).](image-url)
DISCUSSION

Let us compare the wavelet and the SPWVD diagrams. With the SPWVD, we can have independent precisions in time and in frequency. It depends on the choice of $h_N$ and $g_M$. The wavelets are limited by the HEISENBERG-GABOR uncertainty principle : $\Delta \lambda \Delta f \geq \frac{1}{4\pi}$. The frequency scale is a linear one for the SPWVD and a logarithmic one for the wavelet transform. We can recognize on the SPWVD graph the formants of the environment vowels, even when we can not on the figures 1 and 2. The temporal events are well localized with the wavelet transform in high frequencies, but also with the Wigner transform. The Wigner transform gives a more habitual representation of the evolutive spectrum.

Now, we are going to show the differences appearing in the consonant signal between the rest period and the stressed one. First, we observe that the explosion bar of the word “ok” is clearly visible both with wavelets and SPWVD. Both diagrams allow to see that for the rest period (figures 1 and 3) the occlusive silent is not really silent. It is for the stressed one (figures 2 and 4). Time during the explosion bar and the onset of the following vowel is more pertubated for the stressed period. The temporal frontiers, then, are more marked. The lenght of the pronunciation is closely the same in the two situations, the V.O.T (Voice Onset Time) is not modified either.

The frequential informations can only be given by the SPWVD. In fact, this representation is close to the one obtained with a spectrographic analysis. The formants frequencies off the vowels are not modified and the frequential content of the explosion bar seems to be identical between the two states.

Finally, the more useful representation for us is the Smoothed Pseudo Wigner-Ville Distribution. There is no numerical results for this study yet, but we could put in evidence a modification of the evolutive spectrum of the consonants on a large corpus, particularly on the stops consonants : the pilot seems to articulate more “cleanly” during the stress period.

CONCLUSION

After discussing the theoretical properties of the wavelet transform against the Wigner-Ville transform, we can say that, in our application, the SPWVD seems to be the best way to observe spectro-temporal modifications of the consonants. This first approach showed that the interpretation is difficult. Two identical sounds are never pronounced by the same person in the same way. In our case, we can only conclude that some modifications appears between the two states of emotion. They seem to be present for several stop consonants, but not for all. The vocalic context determines the way to uttered a consonant. In the same context, for the same consonant, we then observe modifications going in the direction of more clarity on the diagram and more precision in the articulation.

REFERENCES

STABILITÉ TEMPORELLE D'INDICATEURS SPECTRAUX DE LA QUALITÉ VOCALE

Berard Harmegnies, Marielle Bruyninckx
Département de communication parlée, Université de Mons-Hainaut, 18 place du parc, B-7000 Mons, Belgium

SUMMARY

The paper investigates the consistency of Long Term Average Spectra (LTAS) across time. Five males and five females have been recorded at a 5-year interval. Long term average spectra drawn from their productions are compared in two ways: comparisons of contemporary spectra and comparisons of non contemporary spectra. The dissimilarity coefficients drawn from comparisons of non contemporary spectra have higher values than the others. The effect of time as a source of variation of Long Term Average Spectra is therefore confirmed. Moreover, an additional treatment of the data suggests that the low-frequency part of the spectra could be less time-dependent than the others. This seems nevertheless to be the case mainly in male speakers. This observation could be related to presence/absence of harmonic-like structures in the low-frequency part of LTAS. These structures dissipate as frequency increases, quicker in men than in women. Further experiments should therefore check this hypothesis to determine whether the observed sex-related variability depends upon the appearance of the structures of the spectra or is really due to actual differences in spectral variabilities due to sex-related behavioral differences in speakers.

INTRODUCTION

On trouve couramment exprimée dans la littérature l'idée que le spectre moyen à long terme (SMLT) constitue un corrélat acoustique fiable de la qualité vocale [8]. Si la variance du spectre à long terme peut être établie par diverses approches corrélatives ([2], [3], [11]), sa fidélité semble néanmoins plus douteuse. Ainsi Purui, se livrant à l'analyse de cepstres à long terme y décèle une variabilité temporelle non négligeable [4]. Celle-ci ne peut certes s'expliquer que par l'existence de variations structurelles du spectre à long terme. Bruyninckx et Harmegnies [1] ont étudié la variation du SMLT entre deux périodes distantes, dans le temps, de 5 années. Ils ont attesté le rôle du facteur temporel en tant que source de variation de la forme du SMLT. Leur étude suggérait en outre une meilleure résistance des formes spectrales en basse fréquence. L'expérience était uniquement centrée sur des sujets de sexe masculin. On sait néanmoins [10] que le sexe est un facteur important de la variabilité des formes du SMLT.

Dans cet article, nous confronterons dès lors les résultats dérivés d'un groupe de sujets masculins à ceux obtenus à partir d'un ensemble de sujets féminins. Nous chercherons à déterminer dans quelle mesure la variable sexe interagit avec la variable temps. Nous nous attacherons également à investiguer les éventuelles différences sexuelles dans la résistance du SMLT en basse fréquence.

DISPOSITIF EXPERIMENTAL

Dix sujets (5 masculins et 5 féminins) francophones âgés d'une vingtaines d'années ont été enregistrés lors de 2 sessions, distantes dans le temps de 5 ans. Chaque session d'enregistrement s'est déroulée en chambre anéchoïque, à l'aide d'un matériel invariant de qualité HIFI digitale.
Un même corpus, constitué d'un texte phonétiquement équilibré fut proposé aux sujets lors de chacune des 2 sessions. Chacun fut invité à en produire 5 lectures. L'expérimentation présentée ici porte, par conséquent, sur 100 productions (5 sujets x 2 sexes x 5 productions x 2 sessions).

TRAITEMENT DES DONNEES

Analyses acoustiques

Les productions recueillies furent traitées grâce à un analyseur spectral Bruel Kjaer 2033 dont la fréquence d'échantillonnage était fixée à 12.8 kHz. Les spectres produits étaient ainsi définis sur 400 canaux avec une résolution constante de 12.5 Hz dans la bande 0-5 kHz. Un algorithme de calcul de moyenne à pondération linéaire a fourni un SMLT pour chaque production.

Comparaisons spectrales et quantification des dissimilarités

Différentes comparaisons inter spectrales ont été réalisées. Toutes consistaient en des comparaisons intra sujet, soit de type intra session (ses.1/ses.1 et ses.2/ses.2), soit de type inter sessions (ses.1/ses.2). Pour chacune des deux sessions, nous avons ainsi procédé à la comparaison de l'ensemble des productions d'un locuteur et ce, pour chacune des dix paires non redondantes de SMLT, soit au total 200 comparaisons (2 paires de sessions x 5 sujets x 2 sexes x 10 comparaisons). De la même manière, nous avons comparé l'ensemble des productions issues de la session 1 avec celles issues de la session 2, soit au total, 250 comparaisons (1 paire de sessions x 5 sujets x 2 sexes x 25 comparaisons).

Des traitements identiques ont également été effectués sur des portions des spectres caractérisées par des fréquences supérieures d'analyse graduellement plus importantes. La première portion se limitait aux quarante premiers points du spectre (0-500 Hz), la deuxième aux cinquante premiers (0-625 Hz) et la troisième aux soixante premiers (0-750 Hz).

Chaque comparaison inter spectrale a fait l'objet d'une quantification au moyen de l'indice de dissimilarité SDDD [7].

RESULTATS

Les indices SDDD moyens obtenus par chaque sujet à la faveur de chaque type de comparaison inter spectrale sont reproduits au tableau 1.

<table>
<thead>
<tr>
<th></th>
<th>SM1</th>
<th>SM2</th>
<th>SM3</th>
<th>SM4</th>
<th>SM5</th>
<th>SF1</th>
<th>SF2</th>
<th>SF3</th>
<th>SF4</th>
<th>SF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ses. 1/1</td>
<td>2.24</td>
<td>2.67</td>
<td>2.41</td>
<td>2.24</td>
<td>2.43</td>
<td>3.36</td>
<td>2.72</td>
<td>2.73</td>
<td>2.73</td>
<td>2.67</td>
</tr>
<tr>
<td>Ses. 2/2</td>
<td>2.46</td>
<td>2.38</td>
<td>2.42</td>
<td>2.32</td>
<td>1.94</td>
<td>2.74</td>
<td>3.51</td>
<td>2.63</td>
<td>2.52</td>
<td>3.32</td>
</tr>
<tr>
<td>Ses. 1/2</td>
<td>6.06</td>
<td>5.19</td>
<td>6.22</td>
<td>4.52</td>
<td>3.76</td>
<td>5.56</td>
<td>6.51</td>
<td>6.65</td>
<td>4.64</td>
<td>5.17</td>
</tr>
</tbody>
</table>

Tableau 1: indices SDDD moyens des 5 sujets masculins (SM) et des 5 sujets féminins (SF) selon le type de comparaison inter spectrale

Ces données montrent une tendance nette des SMLT féminins à présenter des taux de variabilité supérieurs à ceux des SMLT masculins. Cette observation confirme des résultats antérieurs [10]. En outre, on note, chez les femmes, à l'instar de ce qui avait été constaté chez les hommes, des dissimilarités inter périodes supérieures aux dissimilarités intra périodes. Il se confirme donc bien que le temps peut agir comme un facteur de variation du spectre à long terme.

A première vue, on ne constate pas d'interaction entre la variable sexe et la variable temps: rien ne permet de suggérer que l'un des sexes serait touché plus que l'autre par l'effet du délai entre les périodes de recueil des données. En revanche, une étude attentive des valeurs de dissimilarité tirées des diverses parties basse fréquence du spectre (données non présentées ici) suggère l'existence d'un dimorphisme sexuel. En effet, chez les hommes, la supériorité
de la dissimilarité inter sessions par rapport aux dissimilarités intra sessions est surtout sensible lorsque les spectres sont calculés de 0 à 625 Hz ou de 0 à 750 Hz. Elle est minimale lorsque les spectres sont calculés de 0 à 500 Hz. Par contre, chez les femmes, on n'observe guère de tendance aussi tranchée.

Afin de rendre plus apparente cette différence sexuelle, nous avons établi un rapport F inspiré de la statistique de Snédecor [9], qui permet de rapporter la dissimilarité inter sessions aux dissimilarités intra sessions. Les valeurs de F sont présentées graphiquement à la figure 1. Celles-ci reflètent bien la différence de tendances entre les deux groupes sexuels: le rapport calculé s'accroît lorsque l'empânement du spectre s'élargit vers les hautes fréquences. Cet accroissement est néanmoins beaucoup plus important chez les hommes que chez les femmes. Il semble donc bien se confirmer que la résistance du spectre basse fréquence aux écarts du temps est nettement plus marquée chez les hommes que chez les femmes.

![Figure 1: valeurs du rapport F chez les sujets masculins et féminins pour les différents types de spectres (1: 0-500 Hz; 2: 0-625 Hz; 3: 0-750 Hz).](attachment:figure1.png)

**CONCLUSION**

À la faveur de cette expérience, nous avons confirmé, à partir d'un échantillon mixte de 10 sujets, l'action de la variable temps sur le spectre à long terme. Des enregistrements des mêmes personnes recueillies à 5 ans d'intervalle se sont révélés produire des spectres à long terme plus différents entre eux que ne le sont les SMLT contemporains. Cette constatation, préalablement opérée sur base d'un échantillon de sujets masculins, est ici confirmée dans le cas de sujets de sexe féminin. L'amplitude de l'effet ne semble pas pouvoir être liée au sexe. Néanmoins, la résistance de la partie basse fréquence du spectre aux écarts du temps, qui avait été constatée sur un échantillon d'hommes, ne semble pas s'appliquer aux sujets féminins.

Les variations observées ici résultent peut-être des différences structurelles caractérisant, en général, les spectres à long terme d'hommes et de femmes. On sait en effet que tout spectre à long terme calculé avec une résolution suffisante fait apparaître, en sa partie basse-fréquence, des structures quasi harmoniques qui disparaissent en général très rapidement avec l'accroissement de la fréquence. Ces caractéristiques s'expliquent par le fait que pour une variation donnée de fréquence fondamentale, un harmonique subit une
variation fréquentielle dépendant certes de la variation du fondamental, mais aussi de son rang. À partir d'un certain rang, la variation fréquentielle absolue de l'harmonique dépasse l'écart inter-harmonique correspondant à la fréquence fondamentale d'origine. Il y a, à ce moment, chevauchement des zones d'influence des harmoniques et le spectre se lisse dès lors.

On peut montrer que pour une voix d'homme ordinaire, le lissage intervient aux environs de 500 Hz [6]. Par contre, pour une voix ordinaire de femme, il se produit aux environs de 850 Hz. Les différences que nous avons constatées ici pourraient être dues à des spécificités respectives des spectres masculins et féminins. Il se pourrait en effet que la meilleure résistance du spectre en basse fréquence des hommes soit due au fait que notre première fréquence de coupure (500 Hz) correspondait justement avec la fréquence où apparaît le lissage du SMLT. La constance du SMLT dans le temps pourrait dès lors être liée à l'existence de structures harmoniques dans la portion du spectre prise en considération. Si tel est le cas, c'est aux environs de 850 Hz que devrait se situer la première frontière spectrale chez les sujets féminins: en effet, jusqu'à cette fréquence, le SMLT est dominé par des structures quasi harmoniques.

Si un effet du même type que celui constaté sur les spectres masculins influence les spectres féminins, il faudrait dès lors, pour le mettre en évidence, comparer des spectres dont la fréquence supérieure d'analyse correspond à la zone fréquentielle où le SMLT commence à se liser (environ 850 Hz) à des spectres dont la fréquence supérieure d'analyse soit située dans une zone où le SMLT présente déjà une apparence lisse.

Tel sera l'objet de nos prochains travaux en la matière. Seule une investigation de ce type permettra de définir si la meilleure résistance du spectre en basse fréquence, chez les hommes, est due aux caractéristiques structurelles du spectre à long terme, ou plutôt à des caractéristiques intrinsèques de la qualité vocale directement liées au sexe du locuteur.

REFERENCES


EVALUATION OF SPEECH AUDIBILITY IN A ROOM BY MEANS OF TRI-SYLLABLE ARTICULATION TEST TAKING CONSIDERATION FREQUENCY OF SPEECH RATE IN CONVERSATION

Osamu Hashimoto, Sho Kimura
College of Science & Technology, Nihon Univ., Chiyoda-ku, Tokyo, Japan

SUMMARY

Tri-syllable articulation\(^1\), which has been proposed as an articulation method for the Japanese language, is a method that treats the extent of interference in speech intelligibility due to the influence of reflected sounds in a room as the influence of the mutual masking of syllables. Taking note of the characteristic that syllables in Japanese are always CV form utterance, evaluations are carried out of the results of hearing, with speech composed of three nonsense syllables successively uttered as the sound source.

It is generally thought that articulation decreases as the influence of reflected sounds increases; however, since the tri-syllable speech time is short when long-pass echoes are generated in large spaces, etc., as well as under conditions in which reverberation time is quite long, there have been cases confirmed in which tri-syllable words are repeatedly heard by the subjects and, unlike the impression of hearing conversation speech, evaluations of test results have yielded high points. It has been pointed out that there is a problem with the fact that the influence of the time distortion of reverberation, echo, etc., is not reflected in test results, since tri-syllable is at a virtually uniform speed at the speed of the limits of audibility.\(^2\)

Accordingly, in this paper, we considered first varying the speech rate corresponding to the speech time of a single syllable using a speech analytical method based on monosyllable speech. We devised tri-syllable articulation test sound source that were variable between four different patterns of speech rates, carried out evaluation tests in sound fields having differing acoustic conditions and demonstrated the effectiveness of the tri-syllable articulation evaluation method, taking the frequency of speech rates in conversation into consideration, through the correspondence of test results to the results of psychological measurements relating to audibility of speech in conversation using paired comparisons.

TRI-SYLLABLE ARTICULATION TEST
SOUND SOURCE WITH VARIED SPEECH RATES

Upon analyzing the waveform of the monosyllable sound sources, it was found that speech is realized through the succession of waveforms having a certain frequency(pitch cycle), as indicated in Fig. 1. We became aware of this characteristic in processing speech signals and consider one pitch cycle to be the unit element. The flow of speech expansion and contraction is indicated in Fig. 2.

First, as the process of analysis, the subject monosyllable waveform is converted to digital data through A/D conversion. A time window is set (300ms) within a range at which a one-cycle segment of the waveform does not fluctuate greatly in this monosyllable waveform, the original waveform is extracted and the pitch location within the time window is calculated using the linear prediction method. This time window is shifted little by little on the time axis (100ms) in order to maintain the succession of the waveform, while pitch analysis of the overall monosyllable waveform is carried out. This pitch location becomes the unit for the pitch cycle and time window in the integration process. Next, by re-setting the time window for the integration process and shifting the waveform on the time axis to coincide with the desired expansion or contraction time while using the pitch cycle as the unit, sound sources for individual syllable
segments having a set speech rate are formed. Tri-syllable test source are thereupon established by coupling three syllables having the same speech rate on the axis.

The range of the speech rates, as indicated in Table 1, is set from a speech rate of 200ms, which is through to be the speed of the virtual limits of speech and hearing, to 800ms, through to be the length at which hearing would not be unnatural, and using a frequency of envelope spectrum, 3.15Hz (312ms speech rate) and 2.00Hz (500ms speech rate) at each 1/3 oct. in that time range, devised sound sources having total of four patterns of speech rates within a range thought to have a comparatively high frequency of appearance in conversational speech.

Examples of the waveforms of the test sound source from the four speech rate patterns are indicated in Fig.3.

We prepared total of twelve sound tables, with three sounds tables each for each of the four speech rate patterns and fifty groups per sound table. In addition, we used a female voice for the test speech.

We had the subjects listen to all of the tri-syllable groups and write down the results of hearing. We considered that even one mistake among the three syllables would result in disqualification of that group in the results and carried out an evaluation according to the rate of correct responses. Consequently, the result of each syllable and each group was calculated as the results; however, the authors treat the group results as the results for overall evaluation.

TRI-SYLLABLE ARTICULATION TEST USING TEST SOUND SOURCES WITH DIFFERING SPEECH RATES

In order to examine the effectiveness of sound sources used in the tri-syllable articulation test having varied speech rates, we carried out hearing tests in sound fields in a gymnasium, together with virtually uniformly uttered tri-syllable articulation test sources with the speech time of the tri-syllables within approximately 500ms.

Here, in order to use the acoustic conditions of the gymnasium itself as the basis while creating different acoustic conditions within the same sound field, sound was dispersed throughout with a digital reverberator using multiple loudspeakers and artificially set up sound fields that would have greater echo or reverberation. The example of impulse response waveforms of each sound field pattern at the approximate center of the gymnasium are indicated in Fig 4.

Hearing tests were carried out with speech recorded on a DAT tape recorder heard by means of headphones through the medium of a dummy head and articulation was evaluated. Furthermore, the subjects were men and women with normal hearing ability and testing was carried out with six subjects at each of measurement points and the mean values were taken as the test results.

With testing at representative points in standard sound

| Table 1. Speech rate corresponding frequency of envelope spectrum |
|---------------------|---------------------|---------------------|---------------------|
| frequency           | 1.25Hz              | 2.00Hz              | 3.15Hz              | 5.00Hz              |
| speech rate          | 800ms               | 500ms               | 312ms               | 200ms               |

Fig.1 Pitch cycle of a speech waveform

Fig.2 Flow chart of speech contraction and expansion

Fig.3 Examples of speech waveform (Tri-syllable articulation test sound source)
fields, sound fields with supplemented echo and sound fields with supplemented reverberation, the results of tri-syllable articulation of groups with sound sources having varied speech rates (hereafter referred to as "%TSAR") and the test results of individual syllables, together with the results of tri-syllable articulation of groups with sound sources having a speech time of approximately 500ms and uttered virtually uniformly (hereafter referred to as "%TSA") and the results of individual syllables, are shown in Fig.5(a)-(c).

In observing this, it can be seen that discrepancies occur in the correct response rate of individual syllables at each sound receiving point in each sound field and, furthermore, that discrepancies occur in test results due to differences in the speech rate even at the same sound receiving point. Since conspicuous echo arrived at 200ms-500ms in the case of sound fields with supplemented echo, influenced by that, there was a conspicuous drop in the correct response rate of tri-syllable articulation at each speech rate of 200ms, 312ms, 500ms and 800ms. However, there is no significant decrease in the correct response rate of %TSA, which was not readily influenced by these echoes. In addition, a tendency becomes apparent for the correct response rate to decrease excessively for the second and third syllables compared to the first in sound fields with supplemented reverberation due to the influence of reflected sound with little energy persisting comparatively into the latter part.

Thus, the use of test sound sources having differing speech rates demonstrated the possibility of being able to express more clearly the differences in the evaluation of articulation due to differing acoustic conditions as the test results, compared to sound sources having only a specific speech rate.

In order to conduct an examination of the correspondence %TSAR results to conversational speech audibility, we examined correspondence of the psychological scale values of the audibility of conversational speech (utterance of sentences 5~5 seconds long without intonation) to %TSAR for a total of ten acoustic conditions, namely, standard sound fields (2 points), echo sound fields (3 points) and reverberation sound fields (5 points), from among the measurement point data under the acoustic conditions described in Section 2.

We had the subjects evaluate which acoustic condition was the most audible with the psychological scale values of conversational speech audibility based on Sheffe's method of paired comparisons).
Meanwhile, %TSAR results were computed for each of four speech rates, however, since conspicuous differences were not confirmed in the spectrum frequency of each of the four speech rates for the conversational speech used in psychological evaluations of audibility, we consider the averaged results of each rate to represent the evaluation results of articulation.

We carried out conversational speech audibility tests and tri-syllable articulation tests with twenty-two adult men and women subjects and attempted to established the correspondence of the psychological scale values of conversational speech audibility to %TSA and %TSAR. The results are indicated in Fig. 6 & 7.

Fig. 6 indicates that, in sound sources in which the speech rate is virtually uniform, %TSA test results are in a range of 40% or more with poor correspondence to psychological scale values, while there is no correspondence of %TSA to the evaluation of audibility. Meanwhile, in the case of results using a sound source taking into consideration the rates of Fig. 7, a favorable correspondence of %TSAR test results to psychological scale values was obtained and a correspondence to the evaluation of speech audibility was also indicated. From these results, it became clear that the evaluation method using tri-syllable articulation, taking the speech rate into consideration, indicated a favorable correspondence to the psychological scale values of conversational speech audibility in standard sound fields with little echo interference, echo sound fields, reverbation sound fields and other sound fields having various properties and that there would be a wide range of applications for this test method.

CONCLUSION

By using tri-syllable articulation test sound sources with variable speech rates, it becomes possible to perceive the influence of reflected sound structure in a room as the influence of the mutual masking of syllables and to obtain a favorable correspondence of tri-syllable articulation, taking the speech rate into consideration, to the results of psychological evaluations of conversational speech audibility. It is, therefore, thought that this method will become a standard method for the subjective evaluation having favorable correspondence to the actual conversational speech audibility.

REFERENCES

EXPLORING TEMPORAL DOMAIN FOR ROBUSTNESS IN SPEECH RECOGNITION

Hynek Hermansky
Department of Electrical Engineering and Applied Physics, Oregon Graduate Institute of Science & Technology, P.O.Box 91000, Portland, OR 97291

SUMMARY
The paper reviews several techniques which are used in conjunction with the short-term analysis and which are reported to be more robust in presence of noise or other non-linguistic factors. We show that one property common to all such techniques is that they are effectively extracting speech features from segments of speech longer than 10-20 ms.

INTRODUCTION TO THE PROBLEM
The communication channel and its noise level remains most often fixed or varies only rather slowly during the conversation. On the other hand, steady configurations of vocal tract are rare and carry only a little of linguistic information.

The description of speech signal as a succession of equally spaced short-term samples originated in speech coding. It assumes that short-term (about 10 - 20 ms) segment of speech are independent samples from different and unrelated stationary processes. Fundamental linguistic unit is likely to be longer than 10 ms and one frame of short-term analysis result provides description of its relatively short (quasi-stationary) part. Since only a short-term "snapshot" of the signal is available at any given time, it is hard to distinguish between the "short-term quasi-stationary" signals (such as speech) and "long-term quasi-stationary" disturbances (such as fixed frequency characteristics of the communication channel or noise).

It appears that the short-term memory of auditory periphery in mammals (exhibited e.g. by the forward masking (see e.g. [22]), firing rate adaptation constant (see e.g. [1]), build-up of the loudness (see e.g. [18])) is at least of the order of about 200 ms, i.e. an order of magnitude longer than the temporal window of the short-term analysis. That means the peripheral human auditory system can effectively integrate rather large (about syllable sized) time-spans of the audio signal.

BEYOND 20 ms
Many speech researchers (see e.g. [3]) do not seem to be aware that some of techniques which are in use in feature extraction for ASR already do consider rather large time-spans of the speech signal. We briefly describe below several techniques (some of them rather well established), for post-processing of short-term speech feature vectors which all claim increased robustness in presence of non-linguistic factors in speech and argue that this increased robustness results from the fact that they do use global knowledge well beyond 20 ms.

DYNAMIC (DELTA) FEATURES
Furui [8] introduced dynamic features of speech to describe time trajectories of speech parameters in the vicinity of a given speech vector. He proposed the first three coefficients of the orthogonal polynomial representation of trajectories of cepstral coefficients within about 50 - 100 ms segment to describe a) the mean value (later substituted by the cepstral vector in the center of the segment), b) the slope, and c) the curvature of trajectories within the given segment of cepstral feature representation.

The higher order dynamic features are invariant to any constant bias within the temporal window used for their derivation. Consequently, since the dynamic features are typically used on the cepstral coefficients, they are invariant to slowly varying linear (convolutive) distortions of the signal introduced e.g. by different frequency characteristics of the communication environment.

The dynamic feature calculation represents finite impulse response (FIR) filtering of time trajectories of cepstral coefficients. The implied band pass filters
are rather selective, emphasizing speech components with a higher (about 8-12 Hz) rate of change [10]. As mentioned above, the dynamic features are invariant to fixed distortions in parameters. Due to emphasis of feature components with a lower rate of change which nevertheless also carry the important linguistic information, dynamic features typically do not perform too well on their own, and are used in conjunction with the original short-term (static) features.

Applebaum and Hanson [5] observed advantage of rather long temporal windows (of the order of 200 ms) was beneficial in isolated-word recognition of noisy Lombard speech, indicating a possible advantage of task-dependent FIR filters in computation of dynamic features.

The dynamic features are used by virtually all state-of-art ASR systems and contribute possibly the only widely accepted significant departure from the plain frame-by-frame short-term feature extraction in ASR. Such success may be attributed to the fact that dynamic features contribute new information which was previously unavailable to the pattern classification component of an ASR system: the information about surroundings of the current short-term segment.

RASTA PROCESSING

Similarly to the dynamic feature calculation, the so called RASTA processing [10] does filtering of time trajectories of speech features. However, it differs from the dynamic feature calculation in: a) allowing for processing sandwiched between two general nonlinearities applied to the feature space (implying nonlinear filtering) and b) typically using rather broad band-pass filter with a relatively flat pass-band which should allow for un-attenuated preservation of most speech components in the feature representation.

The recent work [16] indicates that RASTA processing can simulate forward temporal masking in human hearing.

Ideally, after the appropriate transformation of the features, the disturbing components should combine linearly with the components which need to be preserved. So that the two can be separated by linear RASTA filtering. One of the useful domains appears to be critical-band spectrum with its amplitude compressed by the \( y = \log(\text{const} + x) \) warping function, where the constant is proportional to the noise energy in the signal. The rationale behind such non-linearity is to allow for an approximately linear compression of low-energy spectral components (which contain noise additive in the linear domain) and for an approximately logarithmic compression of high-energy spectral components (which contain speech signal with possible linear distortions additive in the logarithmic domain).

The current RASTA bandpass filter is a fourth-order, single-pole ARMA filter with the pass-band between approximately 1 - 12 Hz and with rather steep slopes. The integration time constant of the filter is about 160 ms. The filter pass-band appears to have some relation to properties of human hearing in perception of FM signals [9]. The steep filter slopes, however, do not seem to be consistent with human auditory perception [16].

Hermsen et al. [11] also investigate use of FIR RASTA filtering of the cubic-root compressed power spectrum trajectories for enhancement of noisy speech. In this technique, the FIR RASTA filters (21 taps with 8 ms sampling, i.e. 168 ms long) are LMS optimized (Wiener-like) to map the 21 samples of time trajectories of a noisy spectrum to a single point of time trajectories of a clean spectrum. Different frequency channels are allowed to have different RASTA filters. For the frequency channels which do convey speech signal (frequencies between 300 and 2500 Hz for the studied analog cellular telephone speech), the RASTA filters have a band-pass character with the pass-band between about 2 and 10 Hz. However, in contrast to the ad hoc designed RASTA filters for ASR, low modulation frequencies are much less suppressed (only about 10 dB down from the maximum).

Since the integrating time constant of RASTA filter is relatively short (spanning about the length of a typical syllable), RASTA processing makes speech features more dependent on a preceding context. Also, speech transitions are typically enhanced. This presents no problem in whole-word ASR but needs to be considered in current sub-word (phoneme) based ASR systems.

CEPSRAL MEAN SUBTRACTION

One of simplest operations which uses global knowledge is the blind deconvolution [19]. In its original formulation of the technique an averaged spectrum of a new signal is matched to an averaged spectrum of some reference. The current reincarnation of this technique, known under the name of cepstral mean subtraction (see e.g. [4]) simply removes means of all time trajectories of cepstral coefficients, thus setting the log spectrum mean to zero. The cepstral mean subtraction is currently quickly becoming a main stream supplementary technique in speech analysis in ASR whenever a real-time operation is not required.

To minimize the inherent processing delay of this technique, the mean can be computed over relatively short amount of future or past data, and windows as short as 50 ms have been used [20]. Rosenberg and his colleagues recently reported that long windows
MULTI-VECTOR INPUT

Makino et al. [13] proposed a so called time-spectrum pattern as a method to classify a short-term speech segment using additional information available in a longer time interval of speech surrounding a given speech instant. In this formulation several short-term analysis vectors (spanning about 100 - 150 ms) are concatenated into one longer vector which is then treated as a basic input feature vector for the subsequent multi-layer perceptron (MLP) classification. This basic idea is currently being used by several ASR groups (see e.g. [21],[14],[12]).

When comparing to the short-term feature vectors, independence and identical distribution of features within a given class may be violated more severely by the time-spectral pattern. Longer timespectral patterns appear to be mostly used in conjunction with nonlinear classifiers. Recent work reports successful use of relatively short (three-frame, i.e. about 35 ms) time-spectral pattern with linear discriminant analysis [20].

The technique relies on the classifier to discover the relative importance of time-advanced and time-delayed speech analysis vectors for the classification of the given speech instant. In principle, the classifier could discover fixed bias in a time trajectory of a given input feature and ignore it during recognition. After the training of the classifier, the weights applied to different time-shifted features form a multi-input multi-output (and in general nonlinear) filter applied to temporal trajectories of speech features.

Even though the time-spectral pattern technique offers the most general form of temporal filtering of speech features, it typically performs better in conjunction with ad hoc designed temporal filters [14],[20].

CONCLUSIONS

The current short-term analysis is an inheritance from speech coding techniques. Looking through a relatively short 10-20 ms temporal window, such analyses cannot distinguish between useful information and disturbance. In extraction of speech features for ASR, it appears that some advantage could be gained by considering speech segments larger than the current 10-20 ms. Several techniques which do so to increase robustness in presence of non-linguistic factors in speech were reviewed in this article.

We argue that one factor which is common in all the discussed techniques is that they perform (generally multi-input multi-output non-linear) filtering of time trajectories of short-term features derived from the speech signal. Additionally, we argue that such temporal processing of speech parameters is consistent with temporal masking in human hearing.
REFERENCES


A ROBUST METHOD OF DETECTING THE PRESENCE OF VOICED SPEECH

Melvyn J. Hunt
Dragon Systems UK Ltd, Millbanks, Stoke Road, Bishops Cleeve,
Cheltenham, England

SUMMARY
This paper describes an attempt to develop a reliable detector of voiced speech that is robust to noise, and in particular works well with noisy telephone material. Methods that look for harmonic structure in the spectrum have been investigated. While the height of the peak in the cepstrum works well with good quality material, it is found to be unreliable with noisy telephone speech. The best technique found applies an autocorrelation analysis to the lowest 1.5 kHz of the power spectrum. It is found to be capable of discriminating voiced speech from voiceless speech, background noise and breath noise in the worst example of telephone material that was tried.

INTRODUCTION
A reliable, robust voicing decision is a necessary component of most vocoders. Indeed, errors in the voicing decision are a major source of problems in making vocoders work in noisy environments. In speech recognition, an ability to locate regions of voiced speech accurately can be important in both the training and recognition phase. Voiceless speech is generally found only in association with voiced speech, and voiceless speech is itself difficult to distinguish from noise. Consequently, location of voiced speech is useful as a means of locating entire speech regions, as well as for segmenting these regions into voiced and voiceless portions.

Voiceless speech is characterized by various features such as its total energy, the pattern of the smooth distribution of energy across the spectrum and the periodicity in the waveform, manifesting itself as harmonic fine structure in the spectrum. Use of this last feature is attractive for various reasons. First, it is largely independent of gain and spectral shaping changes applied to the speech signal. Second, it is robust against additive noise and non-linear distortions. Third, non-speech interfering sounds are less likely to have appropriate harmonic structure than they are to mimic the other features of voiced speech. In addition, since speech recognizers tend to concentrate on smooth spectral features, the detection of harmonic fine structure adds potentially useful independent information that could be used in recognition.

This paper describes work on determining the locations of regions of voiced speech captured over public telephone links, with particular emphasis on material where the SNR is unusually poor.

INITIAL EXPERIMENTS
The experiments described in this paper mainly used recitations of the alphabet over the telephone, which were recorded and kindly supplied by the Oregon Graduate Institute (OGI). Calls from several hundred speakers were examined, and the alphabets with the worst SNR were picked out. These calls turned out to coincide with the ones with the worst recognition performance and the lowest signal level.

A well established technique for analysing the harmonic structure of a signal is to examine the peak occurring in the cepstrum. (The term "cepstrum" is used here to mean the cosine transform of the log power spectrum.) In principle, a peak occurs in the cepstrum for voiced speech at a quefrency
corresponding to the fundamental period. In voiceless speech, no appreciable peak should occur in the range corresponding to a possible fundamental period.

With good quality telephone material, once a suitable lower bound was applied to the power spectrum to prevent spurious fine structure resulting from taking logs of small numbers, the height of the cepstrum peak appeared to be a reliable indicator of the presence of voiced speech. However, with the poor quality telephone calls and with the artificially noisy material, the cepstrum peak height was no longer effective: its height was much reduced, and it correlated with the total energy in the signal rather than with the presence of voicing. Thus, in the letter “h”, for example, the voiceless “ch” portion often had a higher peak than that in the vowel portion that precedes it.

Careful setting of the lower bound of the spectrum as a function of frequency (the “mask”) was helpful but did not solve the problem. Replacing the log function by a cube-root function, which has been reported [1] to give a more noise-robust version of the cepstrum did not help.

THE SPECTRAL AUTOCORRELATION METHOD

Despite the inability of the cepstrum method to detect harmonic structure in the noisy telephone speech, such structure was clearly visible in those parts of high-resolution spectrograms that corresponded to voiced speech — and of course absent from those parts corresponding to voiceless speech or noise bursts.

On reflection, it became clear that the eye was noticing local harmonic structure within the range of a formant. In the regions between formants any harmonic structure was buried in the noise or below the spectral mask. Moreover, the harmonic structure in these poor quality calls was visible only in the lower part of the spectrum, corresponding to the range of the first and second formants.

In contrast to the eye, the cepstrum is just as sensitive to correlations occurring over separations of, say, 2 kHz as it is to correlations between adjacent harmonics, separated by only 100 or 200 Hz. Similarly, while the eye tends to notice structure in the higher energy parts of the spectrogram, the cepstrum is equally sensitive to structure in the lower energy, noise-dominated, parts of the spectrogram, provided that it is above the mask level.

To provide an analysis that would respond in a similar way to that of the eye, we adopted the following method. First, to reduce any broad spectral shape effects the difference between adjacent spectral components is taken (that is, \( y_n = x_n - x_{n+1} \)). Next, an autocorrelation analysis is applied across the spectrum (\( \Sigma y_1 y_{1+k} \)). The order of this analysis is limited so as to provide a sensitivity only to local harmonic structure. Also, it is applied only over the lower-frequency part of the spectrum. The spectral autocorrelation obtained in this way is extended with zeros and its cosine transform is then computed. The height of the peak in this cepstrum-like function is then determined. We call this peak height the “harmonicity” of the signal.
We found that the optimum autocorrelation order was around 14, which corresponds to a frequency range of about 430 Hz, and that the analysis was best confined to the bottom 1.5 kHz. With these values, the method was capable of successfully discriminating between voiced sounds on the one hand and voiceless speech sounds, noise bursts and breath noises on the other.

Since the cepstrum analysis had been carried out over the whole spectrum rather than just the lower portion, we repeated it with the spectrum above 1.5 kHz set to zero. While this proved useful, the performance of the cepstrum analysis remained inferior to that of the autocorrelation method. When the results are normalised such that the peak height is typically the same in equivalent voiced sounds, the (spurious) peak height of the autocorrelation method in voiceless sounds is typically only a half to a third of that of the cepstrum method. Moreover, the peak height in the background noise is only one sixth of that produced by the cepstrum method.

![Figure 4](image_url)  
**Figure 4.** Harmonicity and cepstrum peak height plots for the letters 'g' and 'h' spoken over a particularly noisy telephone connection. The second bump in the 'h' corresponds to the voiceless 'ch' sound and is therefore spurious. Note that the harmonicity and cepstrum peak heights are similar in the voiced regions but the harmonicity has a weaker spurious response to the voiceless sound and a much weaker spurious response to the background noise between the words.

![Figure 5](image_url)  
**Figure 5.** The letter 'z' (pronounced 'zed') from the same noisy telephone call as in fig. 4. The second bump corresponds to breath noise, and the response is therefore spurious. The harmonicity response to the breath noise is much weaker than that of the cepstrum.

**DISCUSSION**

When the analysis is carried out over the same frequency range, the cepstrum and spectral autocorrelation method differ in two obvious ways. First, the autocorrelation method confines itself to local harmonic structure. Second, the autocorrelation method depends on the product of pairs of components in the spectrum, while the cepstrum method depends linearly on the component values. This second difference means that the autocorrelation method tends to put more emphasis on the high-activity regions of the
spectrum, such as the frequencies close to formants. Either or both of these differences could contribute to the superior performance of the autocorrelation method.

Confining the search for harmonic structure to frequency ranges of a few hundred Hz reduces the probability of spurious detection of harmonic structure, to which the cepstrum method is prone. The fact that results with the autocorrelation method get worse when the order of the analysis is increased much beyond 14 confirms that the localisation of the analysis contributes to its superiority. It is not yet clear whether the dependence on the product of component values in the autocorrelation method is also useful. An attempt to increase further the weighting given to large component values did not improve performance.

USE OF THE SPECTRAL AUTOCORRELATION METHOD TO DETERMINE $F_0$

If the spectral autocorrelation method is to be used to make the voicing decision, it is natural to consider its use for determining fundamental frequency ($F_0$) as well. The restricted order of the autocorrelation analysis inevitably limits the resolution of the $F_0$ analysis. A preliminary look at the results of the method suggests that it is usable, apart from a slight shift upwards in frequency resulting from end effects in the autocorrelation analysis causing the amplitude of the autocorrelation function to decrease with increasing delays. While it has not so far shown a clear advantage over the standard cepstrum method as a means of determining fundamental frequency, we believe that a slight variant on the method will be computationally more efficient than the cepstrum.

CONCLUSIONS

Limited-order autocorrelation across the differenced log power spectrum appears to be a robust method of locating regions in a possibly noisy signal that have harmonic structure corresponding to voiced speech. As well as being a robust means of making a voicing decision, the method may have potential for the computationally efficient determination of fundamental frequency.

ACKNOWLEDGEMENTS

Prof. Ron Cole of the Oregon Graduate Institute kindly provided the telephone recordings.

REFERENCES

CHARACTERISTICS OF A HEARING AID FOR TELEPHONIC SPEECH BY SINGLE RESONANT ANALYSIS

Takashi Ikeda, Akira Watanabe*, Yuichi Ueda*, Kouji Tasaki**
Kurume National College of technology, 1232, Komorino-machi, Kurume, 830 Japan
* Faculty of Engineering, Kumamoto Univ., 2-39-1, Kurokami, Kumamoto, 860 Japan
** SHARP Co., Ltd.

SUMMARY
This paper describes a hearing aid which is proposed by a new analysis-synthesis method for telephonic speech, and the characteristics as the real-time system. In this paper the algorithms are introduced first, and then the characteristics of the system are discussed.

The method have the following advantages in comparing other digital hearing aids. One is to use a new analysis-synthesis method. In the method, two resonant components are extracted from telephonic speech by inverse filtering, and re-synthesized by adding them after amplitude compression of the respective components. This provides that each components is compressed at resonant frequencies, so that the formant peaks can always be kept between HTL and UCL of the hearing impaired person. The other is in a way to decide coefficients for amplitud compression. The coefficients are calculated by the rms value of a frame to remove non-linear distortion. A multi-DSP hardware has been developed to realize the system as a prototype. The DSP units are combined by master-slave connection and each takes charge of independent parameters processing. It has been shown that the compressed signals for the system is the same as the simulated ones. Finally the characteristics of the system to transmit telephonic speech is investigated in existence of stationary noise.

PRINCIPLE OF SINGLE RESONANT ANALYSIS METHOD
In the amplitude compression of speech, the components which contribute to speech understanding should be compressed into the residual hearing area of the person. For this purpose, we propose single resonant analysis in which resonant components should be compressed. Compression coefficients in the hearing aids are calculated by the following procedure in general. A function is defined approximately to transfer ordinal speech signal into amplitude compressed one at a frequency \( \omega \). Solid line in Fig.1, is an example for the function. If \( Y(\omega) \text{[dB]} \) represents the hearing level of the normal hearing person, amplitude conversion function \( Z(\omega) \) can be calculated as follows.

\[
Z(\omega) \text{[dB]} = \alpha(\omega)Y(\omega) + \beta(\omega) \quad (1)
\]

\[
= 20 \cdot \log(\hat{X}(\omega)/X_n(\omega)) \quad ; \quad Z(\omega) \leq \text{UCL}(\omega)
\]

\[
Z(\omega) = \text{UCL}(\omega) \quad ; \quad \text{otherwise}
\]

where

\( X_n(\omega) \): HTL of normal hearing person (SPL)

\( \hat{X}(\omega) \): Level of output signal (SPL)

![Diagram showing hearing level and compression](image)

**Fig. 1** A function of amplitude conversion.

Single resonant analysis method is explained by Fig.2. The model includes first and second formant components. In the figure the lower curve represents spectra of input speech. The two single resonant components are extracted by inverse filtering. Those spectra are shown by the dotted lines respectively. This technique makes it possible to decide compression coefficients at the formant frequencies. The compression coefficients are multiplied by the components and added in parallel synthesis. As a result, formant peaks
always exist in the residual hearing area of the listener. Thus, upper spectrum is obtained[1].

![Diagram of a single resonant analysis](image)

Fig. 2 A single resonant analysis.

Instantaneous compression has been used in the conventional hearing aid so far. For the instantaneous compression, the coefficients change very rapidly as the level of input sample changes. It causes non-linear distortions in the output signals. So we propose deciding the coefficients from the rms level of the input samples in a frame. Since the compression coefficients in this system change with rms level, the change of the coefficient is slow and quasi-stationary differing from that of instantaneous amplitude. Therefore the non-linear distortion is little caused. The equation to get for the rms value based on compression coefficient \( \eta(\omega) \) is shown as below. From eq.(I)

\[
Z(\omega_i) = \alpha(\omega_i)Y(\omega_i) + \beta(\omega_i)
\]

\( \omega_i : i \)-th formant freq.

For simplicity of the description, \( Z = \alpha X + \beta \)

where \( Y = 20 \log \frac{X}{X_L} \) (dB) \( Z = 20 \log \frac{\tilde{X}}{X_L} \) (dB)

\( X \) : Formant level of input signal(SPL)

\( X_L \) : HTL of normal hearing person(SPL)

\( \tilde{X} \) : Formant level of compressed signal(SPL)

\[
\therefore 20 \log \frac{\tilde{X}}{X_L} = \alpha \cdot 20 \log \frac{X}{X_L} + \beta
\]

\[
\therefore \tilde{X} = \left( \frac{X}{X_L} \right)^{\alpha} \cdot 10^{\frac{\beta}{10}} = \frac{\tilde{X}}{X} = \eta = \left( \frac{X}{X_L} \right)^{\alpha - 1} \cdot 10^{\frac{\beta}{10}}
\]

Assuming \( \tilde{X} = \tilde{x}(t) \) and

\[
X = k \sqrt{\frac{1}{T_0} \int_0^{T_0} x(t)^2 dt} = k \tilde{X} \] in single resonant waves, the following equation will be derived.

\[
\frac{\tilde{x}(t)}{x(t)} = \eta = \left( \frac{kX}{X_L} \right)^{\alpha - 1} \cdot 10^{\frac{\beta}{10}}
\]

The coefficient multiplier \( k \) is empirically decided to be \( k = 100 \).

\( \eta \) is computed in each frame by \( \alpha, \beta, X_i \) and \( \tilde{X} \), and linearly interpolated between frames.

Finally, we can get the compressed signal by the following equation for \( i \)-th resonant component.

\[
\tilde{x}(t) = \eta_i(t)x(t)
\]

By using the description \( i \) again

\[
\tilde{x}_i(t) = \sum_{i=1}^{N} \tilde{x}_i(t)
\]

Thus, the compression signal can be obtained by the equation,

\[
\tilde{x}_i(t) = \sum_{i=1}^{N} \tilde{x}_i(t)
\]

where \( C_i \) represents a coefficient for parallel synthesis.

The block diagram of the hearing aid for telephonic speech by single resonant analysis method, is shown in Fig.3. In this case, the lowest three formant frequencies (F1, F2 and F3) are used to extract single resonant components. The system is supported by stable speech parameters (formant and pitch frequencies) extraction[2]. The IFn shows an

![Block diagram of the hearing aid for telephonic speech](image)

Fig. 3 Block diagram of the hearing aid for telephonic speech.
inverse filter with a fixed bandwidth, which eliminates the n-th formant component. The extracted three formant frequencies control each zero of inverse filter to cancel the formant resonance. The coefficients for parallel synthesis are calculated with F1 and F2 each. The coefficients for the amplitude compression is decided by two factors. One is the HTL and UCL of a hearing impaired person at the single resonant frequency, and the other is the rms value of the input signal frame which is linearly interpolated between the frames. Since the compression coefficients are calculated by the rms value, it shows quasi-linear characteristics. After multiplying these coefficients by the extracted single resonant components, both are added at last to produce compressed speech signal output. The parameters (F0-F3) are extracted every speech frame, they are interpolated between the frames for each speech samples to control the zero of the inverse filters and to calculate the coefficients.

CHARACTERISTICS OF THE SYSTEM

Assuming the auditory characteristics of sensorineural hearing loss, the new DSP system is examined by speech signals. As an example English word "approach", uttered by a adult male speaker is shown in Fig.4. Speech is filtered by BPF(300-3000Hz) and compressed by the system. Assumed HTL and UCL are indicated in the spectrum of (4-5) in Fig.4. According to the waveforms and sound spectrogram patterns, following three items can be pointed out. First, after the signal processing to compress speech, non-linear distortion can not be observed. Second, energy of formant components increased in the vowel /a/, /o/ and /u/. Third, consonants show similar tendencies to vowels. The spectra of affricate /ts/ is shown in Fig.5. Formants of input signals are compressed, as indicated in the sound spectrogram of Fig. 4. According to the auditory impressions of the samples (66 words in English and 620 /cvc/ patterns in Japanese) sound clearness increases compared with input telephonic speech.

To examine the effectiveness of the amplitude compression, the parameters Cx is introduced. It can be defined the following eq. (2).

\[ C_x = \left( \frac{H_x(f) - \text{HTL}(f)}{\text{R}(f)} \right) \times 100 \quad (2) \]

where

\[ \text{R}(f) \ [\text{dB}] = \text{UCL}(f) - \text{HTL}(f) \]

\[ \text{H}_x(f) \text{ spectrum level of input signal} \]
\[ \text{H}_c(f) \text{ spectrum level of compressed signal} \]

\[ f = \text{F1} \text{ or } f = \text{F2} \]

4-1: Input speech (BPF output)

4-2: Sonogram of input speech

4-3: Compressed speech

4-4: Sonogram of compressed speech

4-5: Spectrum /u/ in the compressed speech

Fig.4. Effect of compression in vowel /u/

Fig.5 Parameter definition in spectra /ts/.
About vowels single resonant analysis method works very well, because of the stable fortant extraction[3]. In all the investigated samples extracted fortant peaks are transferred into assumed hearing area. Then the next interest goes to the consonants. The results of auditory listening of compressed speech which include consonants show the improvements in quality. So we discuss the compression results exactly by the new parameter. The examples are the voiced and unvoiced plosives. In Fig.6 evaluation by parameter Ci, Cc is shown. All the consonants shown in Fig. 6 are picked up from the naturally pronounced English words by a native speaker. (6-1) and (6-2) show the compression result of /p/, /t/, /kt/. (6-3) and (6-4) show the examples of /b/, /d/, /g/. (6-5) and (6-6) are /p/, /t/, /kt/ which includes white noise of the S/N = 15 [dB]. Black bars represent the spectrum level of the input speech, and white bars represent the spectrum level of after the compression, respectively. The figures show the level spreads widely and exist in the area which is lower than HTL before amplitude compression. But the distribution of spectrum level have changed after the compression. As the white bars indicate, levels distribute among the 0 to 100, where 0 means at the level of HTL and 100 means UCL. So this means that all the spectrum level of consonants are lifted into the audible area at F1 and F2, including noise corrupted samples (6-5) and (6-6).

CONCLUSION

The hearing aid using single resonant analysis generates amplitude-compressed signals with little spectral distortion. Because, each single resonant component is amplified linearly and the coefficients of the amplification are decided by rms value of the components which varies slowly. In order to examine how this effect is reflected, a measure which represents the degree of the compression has been introduced. Despite of the small amplitude, it is shown that unvoiced and voiced plosives in 27 words are completely compressed as we considered. The quality of the compressed speech is rather good comparing with the original (telephonic speech) according to preliminary listening.

REFERENCES

A METHOD OF SPEECH QUALITY EVALUATION FOR PATIENTS OF ORTHODONTIC TREATMENT

Ken'iti Kido, Kenji Tanahashi, Yasuhiro Ohuchi, Masako Matsuzaki*, Takayuki Kuroda*

Chiba Inst. Techn., Dept. of Computer Science, Narashino 275, Japan
* Tokyo Med. & Dent. Univ., Dept. Orthodontics and Sch. Dentistry, Yushima, Japan

SUMMARY

This paper describes a vowel display system being developed for the speech quality evaluation. The main purpose of the study is to develop an objective method for evaluation of speech of the patients in orthodontic surgical treatment. The system analyzes input speech every 5 ms using linear predictive technique. The lowest two local spectral peaks P1 and P2 are extracted. The vowel trajectory is displayed using the frequencies of those peaks on the P1-P2 plane. In most cases, P1 and P2 correspond to the lowest two formants, but in the case of indistinct utterance, the correspondence between the local spectral peaks and the formants tends to fall into disorder. And it is considered that the display system can be useful for the speech quality evaluation and the utterance training. The patients' speech differ by the surgical operation from that before the treatment in the hearing and in the spectra and the difference changes by the laps of time. The differences in the vowel trajectories, in the spectra and in the hearing inspection are investigated for the utterances of the same speaker at the different timing using several patients' speech.

INTRODUCTION

This study was begun by a request of dentist who engaged in the orthodontics. The dentist feels that the speech uttered by patients of opposite clench is usually improved by the surgical operation of orthodontic treatment. But, there is no objective way to assert it.

It has already been discovered that the power of high frequency components in fricatives increase by the orthodontic treatment for the opposite bite, but we consider that the quality of vowels should change by the changes in the vocal tract due to the surgical operation.

SUBJECTIVE EVALUATION

The speech samples uttered by the patients before and after the surgical operation were subjectively evaluated by the pair comparison method. Table 1 shows an example of the results. The speaker is a patient of opposite clench. The speech is /fukuraw/. BO, W1, W5 and M13 in the table are abbreviations of "before the operation", "one week after the operation", "five weeks after the operation" and "13 months after the operation", respectively and the figure 93 indicates that 93 % of 30 answers by 15 subjects preferred the speech at M13 than that at BO. The figures in the column K are the mean values of the corresponding lines.
Table 1 Result of pair comparison of speech uttered at four timings

<table>
<thead>
<tr>
<th></th>
<th>B0</th>
<th>W1</th>
<th>W5</th>
<th>M13</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>B0</td>
<td>53</td>
<td>87</td>
<td>37</td>
<td>7</td>
<td>43</td>
</tr>
<tr>
<td>W1</td>
<td>13</td>
<td>53</td>
<td>7</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>W5</td>
<td>63</td>
<td>93</td>
<td>40</td>
<td>23</td>
<td>60</td>
</tr>
<tr>
<td>M13</td>
<td>93</td>
<td>93</td>
<td>77</td>
<td>40</td>
<td>88</td>
</tr>
</tbody>
</table>

ANALYSIS

The speech sample is first A/D converted at 16 bits accuracy and 16 kHz sampling rate. The speech samples are cut out every 5 ms using the Hamming window of 30 ms length. The cut out sequence is analyzed by 18-degree LPC after taking the difference of every successive data. From the result, the speech spectrum is made using 1,024 points FFT.

The local spectral peaks are extracted from the speech spectrum. And, the spectral peaks are called P1, P2, P3, ... from the peak of the lowest frequency.

DISPLAY METHOD

To display the transition of speech spectrum, so far mainly used is so called sound spectrogram or sonagram which is very familiar for us speech scientists. But, it is not easy to get useful information on the speech quality from those spectrograms without rich experience. To use the speech display system in the evaluation or training, easier method to get useful information is strongly required.

**T-P pattern**

It is considered that the detailed spectral information is not necessary and the transition of frequencies of the formants or the local spectral peaks will be more important.

Accordingly, the transition of the frequencies of spectral peaks is displayed as shown in Fig.1 which is called the T-P pattern and displays up to 6 peaks from the peak of the lowest frequency. The frequency scale in the ordinate is logarithmic. Using the logarithmic frequency scale, it becomes easier to observe the transition of significant spectral peaks.

**P1-P2 pattern**

The points are dotted every 5 ms on the P1-P2 plane using the frequencies of P1 and P2. The scales of abscissa and ordinate are the logarithmic frequency as the same as the frequency axis in T-P pattern. To draw the trajectory, the averaged value of successive 8 values of P1 and P2 are computed every 5 ms and the averaged points are connected by a line as shown in Fig.2. Taking such the averaged values, the irregularity of the trajectory due to the absence and the addition of the local peaks in the LPC spectra are much reduced.

The beginning point of trajectory is shown by a small circle and the ending point a rectangle in the P1-P2 pattern. The elapsed time is indicated by the small triangle every 50 ms.
Fig. 1 T-P pattern of normal speech /aiueo/

Fig. 2 P1-P2 pattern of normal speech /aiueo/

Fig. 3 P1-P2 pattern of ventriloquized /aiueo/ by the same speaker as Figs. 1 and 2.

Fig. 4 P1-P2 pattern of normal /niwatori/ by the same speaker as Figs. 1, 2 and 3.

Fig. 5 P1-P2 pattern of patient's /niwatori/ before the surgical operation.

Fig. 6 P1-P2 pattern of patient's /niwartori/ 1 year after the surgical operation.
Figure 1 shows a T-P pattern of a continuous speech /a,i,u,e,o/ uttered by a normal young female speaker. As seen in the figure, T-P pattern gives better perspective than the conventional spectrogram as the transition of the frequency of spectral peaks are clearly displayed and the significant movements of the spectral peaks can be easily observed.

The peaks corresponding to the first formant disappear in some period in /a/ and some discontinuity of peaks corresponding to the second formant are observed at the phoneme boundaries. The peaks higher than P3 seem to give no significant information.

Figure 2 shows the P1-P2 pattern of the same speech. This figure looks like the formant transition pattern on the first and second formant plane.

We investigated also the P1-P3 locus and P2-P3 locus. But we could not find useful feature from those loci.

Figure 3 shows the P1-P2 pattern of the speech /a,i,u,e,o/ ventriloquized by the same speaker. In the ventriloquism, the lips are kept almost closed during the utterance and the speech quality is unfamiliar, but the phonemic information is almost correctly transmitted.

As seen in Fig.3, the deference is well demonstrated by P1-P2 pattern. It is observed that /e/ approaches to /u/ in this case.

EXAMPLES OF P1-P2 PATTERNS OF SPOKEN WORDS

Figure 4 shows P1-P2 pattern of /niwatori/ uttered by a normal young female speaker. The trajectory has just the expected pattern as the first and the second formant pattern except for that P2 is somewhat higher. The speech is very clear in hearing. It is considered from the observation of many utterances that the higher P2 is a special feature of her utterance.

The same speech uttered by a young female patient of opposite clenched is analyzed and displayed on Fig.5. The main difference of this figure from Fig.4 is lower P1 in the initial part /ni/ of the word. The speech uttered by the patient is inferior than that of the normal speakers in hearing.

We analyzed the same speech uttered by the same speaker one year after the surgical operation. Figure 6 shows the P1-P2 pattern of the speech. The substantial difference from Fig.5 can not be observed in Fig.6 only by the visual inspection but the pattern in Fig. 6 is somewhat simpler than that in Fig. 5. Now we are trying to get numerical index which corresponds to the subjective evaluation using such the features.

CONCLUSIONS

A vowel trajectory display system is now under being developed for the purpose of the use in the speech quality evaluation and in the visual feedback for the utterance training for the orthodontic treatment.

The frequencies of the local spectral peaks are used for the display as the main features. The T-P pattern and the P1-P2 pattern are compared each other and the superiority of P1-P2 pattern is demonstrated. There is distinct difference between the clear normal speech and the patients' speech in visual inspection. It remains for the future how to express numerically the difference between them.

REFERENCE

AN EMG STUDY OF THE FEATURE "TENSITY" WITH REFERENCE TO ENGLISH AND KOREAN

Dae-won Kim
Pusan National University, Department of English, College of Education, Pusan 607-735, (South) Korea

SUMMARY
There has been no EMG evidence for the feature tense-lax distinction in English but in Korean. The technique of electro-myography (EMG) was used to see if the existing claims hold true and to determine the effects of stress on the EMG gestures exerted by the orbicularis oris superior muscle for the labial closure in English stops. It was found (1) that in unstressed syllables the peak EMG amplitude was a valid measure for the tense-lax opposition in the bilabial stop consonants of English, with /p/ being tense and /b/ lax, but it was invalid in stressed syllables (2) that in Korean the oral closure interval, the duration of muscle action and the peak EMG activities were valid measure, with /p/ /b/ being tense and /p/ /b/ lax and (3) that in English the effects of stress on the peak EMG amplitude was significant in /b/ but not in /p/. It was hypothesized that in stressed syllables in English /p/ and /b/ may be differentiated by the EMG activities from a muscle other than the orbicularis oris superior muscle, e.g. the respiratory muscles relating to aspiration. The possible reasons for the inconsistent EMG data were discussed.

INTRODUCTION
The existing literature reveals that there are no EMG evidence for the feature tense-lax distinction in English but in Korean. However, there are two types of claims over the EMG activities from the orbicularis oris superior muscle in bilabial stops of English: (1) The orbicularis oris muscle activity is essentially the same for English /p/ and /b/ (e.g. Harris, et al., 1965; Tatham and Morton, 1968; Tatham, et al., 1985; Lubker and Parris, 1970; Sussman, et al., 1973) and (2) in word-final position of the isolated /CVC/ words, /p/ and /b/ were differentiated by the peak EMG amplitude, but not in word-initial position (Fromkin, 1966). Fromkin (1966) used the inconsistent EMG results in rejecting a linguistic hypothesis of tense-lax distinction on the basis of a hypothesis that "if muscular gestures for /p/ and /b/ are either identical or if there is no consistent relationship, a feature other than the tense-lax feature must distinguish these two phonemes, e.g., the action of the glottis" (p. 170). However, in the previous studies (e.g. for CVC/CVCV utterances, Harris, et al.; for CVC utterances, Fromkin, 1996; for CV/CVC/VCV utterances, Lubker and Paris, 1970), EMG data obtained were pulled generally in a single histogram and the possible effects of stress on the EMG activities have been relatively neglected. The isolated monosyllable word of English is considered to be produced with stress on it, and in the /C1VC2/ context, the effect of stress on the muscular activities may be linked more closely to C1V- /C2V (MacNeilage and Declerq, 1967; Tatham and Morton, 1968). If this is the case, the isolated speech items with /CVC/ structures used from Fromkin (1966) appeared to be improper for her claim since in her study she failed to consider the effects of stress on the EMG activities. It appeared from Table I (Lubker and Parris, 1970) that in unstressed syllable there was a clear tendency to a higher EMG activity from the orbicularis oris muscle in /p/ than in /b/ although the difference was not always stable.

This study was designed mainly (1) to see if the EMG activities from the orbicularis oris superior muscle are essentially the same for /p/ and /b/ in English, (2) to determine the effects of stress on the EMG activities exerted by the orbicularis oris superior muscle for the labial closure in English bilabial stops and (3) to examine the claim (Kim, 1980; Kim, 1965) that in Korean the EMG activities from the orbicularis oris superior muscle were significantly greater both in the tense aspirated /pʰ/ and in the tense unaspirated /p/ than in the lax unaspirated /p/. In this study,
the feature tensity has been defined as the amount of muscle action (or energy) used in producing a phoneme. According to this conception, the tense stop should be characterized either by a longer EMG activity or by a greater peak EMG gesture or by both. The articulatory correlates of the feature tensity are time and amplitude. It is hypothesized that the distinctive feature tensity distinguishes /p/ from /b/ in English and /p/, /p/ from /p/ in Korean, with English /p/ and Korean /p/, /p/ being tense, and Korean unaspirated /p/ and English /b/ lax.

**METHOD**

Three (1 American and 2 Korean) of five subjects were used. Two subjects (a British and a Korean) were rejected because of small amplitude of EMG signals throughout trial runs, thus yielding unsatisfactory recordings. The subjects used had no speaking problems.

As seen in Table 1, in order to minimize the effects of coarticulation, rounded and spreaded vowels were avoided. The speech items were produced in isolation rather than in a carrier phrase since the temporal or spatial extent of coarticulation is not yet precisely known.

<table>
<thead>
<tr>
<th>Table 1. English /CVCV(C)/ items</th>
<th>English /CVC/ items</th>
<th>Korean /VCV/ items</th>
</tr>
</thead>
<tbody>
<tr>
<td>/bal:/n/ /pol:/n/ /bɔɹʿʌɛps/ /pɔɹʿʌɛps/ /baɬ/ /paɬ/</td>
<td>/ɑːp/ /ɑːp/ /ɑːp/</td>
<td>/ɑːp/ /ɑːp/ /ɑːp/</td>
</tr>
<tr>
<td>/bɑɬ/ /bɹɑɬ/ /bΛɬ/ /bɬ/</td>
<td>/paɬ/ /paɬ/</td>
<td>(Korean /p/ = a lax unaspirated stop, /b/ = /b/ = /b/ = an aspired stop, /p/ = an unaspirated stop).</td>
</tr>
</tbody>
</table>

The speech materials with the /CVCV/ syllabic structure were produced under controlled experimental conditions with stress being one of them. In order to obtain natural data in unstressed syllables, an attempt was made to construct natural word-like /CVCV/ items in pairs, such as 'ballon', 'balloon', 'perhaps', and 'berth'. The /CVC/ items of English and the /VCV/ items of Korean were used to see if the claims (Fromkin, 1966; Kim, 1980) hold true. Each isolated speech item was produced five times at a normal (or moderately slow) speech rate, yielding a total of 80 utterances in English (/CVCV/ = 20, /CVCV/ = 40, real-word-like /CVCV/ = 20), but ten times in Korean, giving 30 utterances.

For subject training, preparations prior to the test session and measurement, the method used in Tatham (1985) was duplicated. Simultaneous raw voice signals and electromyographic records, using an electro-myograph (QUANITUM 84, CADWELL CO.) and an audio microphone, were obtained from surface electrodes on the upper lip of each of three speakers as they produced bilabial consonants in the isolated speech items.

Three dependent parameters were measured: (1) The peak of the EMG amplitude in μV: the distance from the noise floor to the peak of the EMG amplitude trace, (2) the duration of muscle action in ms: the interval from onset of activity to cessation of activity, and (3) the duration of oral closure in ms: for the unaspirated stops, the interval from the offset of regular pulse to the onset of the regular pulse in the voice signal, but for the aspirated stops, the interval from the offset of the regular pulse to the onset of explosion in the voice signal. Since in word-initial stops of English, the offset/offset of the oral closure was not detectable in the voice signal, the measuring of the oral closure interval was excluded in English. Statworks was used on a Macintosh P.C. for the statistical analysis. This experiment was undertaken at the Pusan National University Hospital.

**RESULTS AND DISCUSSION**

**EMG manifestation of stress in English**

As seen in Figure 1, in isolated /CVCV/ utterances the peak EMG amplitude of the orbicularis oris superior muscle for the labial closure for /b/ was significantly greater (overall average 37%) in stressed syllables than in unstressed syllables, regardless of position, but this difference did not occur in /p/. Thus, the effects of stress on the peak EMG amplitude were phoneme-sensitive. However, the effects of stress on the duration of muscle action were inconsistent in both /b/ and in /p/.

Figure 1. Diagram showing the effects of stress on the duration of muscle action (DMA) in ms and the peak EMG amplitude (PEMG) in μV in /b/ and /p/ of English (n=5, vertical lines = standard deviation).
Similarly, the stress-related aspiration also was phoneme-sensitive but in an opposite direction to the case with the lip muscle gestures. It was reported that VOT in /p.tk/ was significantly greater in stressed syllables than in unstressed syllables while this difference was negligible in /b.d.g/ (e.g., for BrEng intervocalic stops, see Kim, 1987; for AmEng word-initial stops, see Lisker and Abramson, 1967). Considering the phoneme-sensitive phenomena, associated with the peak EMG amplitude and the aspiration, one can presume that in /p/ the manifestation of stress may be carried out by the EMG activity from a muscle other than the orbicularis oris muscle gestures, e.g., the respiratory muscle activities relating to aspiration, and the phoneme-sensitive EMG manifestations of stress may be one of reasons for the inconsistent EMG results within a given speaker.

**Stress-sensitive EMG gestures in English**

Figure 2 shows that in unstressed syllables of the /CVCV/ words and the /CVCVC/ words /p/ and /b/ were significantly differentiated by the peak EMG amplitude exerted by the orbicularis oris superioris muscle, regardless of position (initial/final); the peak amplitude was greater (overall average 32%) in /p/ than in /b/. However, the duration of muscle action for /p/ and /b/ was inconsistent, regardless of the placement of stress, and in unstressed syllables the peak EMG amplitude was a valid measure for the /p/-/b/ distinction. In stressed syllables, on the other hand, both in the /CVCV/ words and in the /CVC/ words, the peak EMG amplitude was essentially the same for /p/ and /b/ as in the existing literature and the duration of muscle action was inconsistent. Accordingly, the hypothesis that the EMG feature: tensility differentiates /p/ from /b/ in English, with the English /p/ being tense and the English /b/ lax, has been verified in unstressed syllables only. This is in contrast with the existing claims (e.g., Fromkin, 1966; Lubker and Parris, 1970; Harris, et. al., 1965; Tatham, et. al., 1985). This difference may be due mainly to the fact that the previous studies have relatively neglected the effects of stress on the EMG parameters. Considering the facts that the greater degree of aspiration in /p/ than in /b/ in stressed syllables, except for the stops in the /st/-stop clusters and word-final position, one may assume that in stressed syllables English /p/-/b/ distinction may be carried out by the EMG activities from a muscle other than the orbicularis oris superior muscle; e.g., the respiratory muscle aspiration.

**Inter-speaker variabilities.**

In an EMG study using /VCV/ words and five subjects, Sussman, et al. (1973) found that /p/ and /b/ were significantly differentiated by the depressor anguli oris muscle, but not either by the orbicularis oris superior muscle or by the quadratus labii superior muscle. Contrasted to this, this study showed that in unstressed syllables the peak EMG activity from the orbicularis oris muscle was significantly greater in /p/ than in /b/. Thus, there are inter-speaker differences in activating certain muscle(s) to distinguish /p/ from /b/, which seems to be one of the main reasons for the inter-speaker variabilities in the EMG data. This assumption

---

Figure 2. Diagram showing the mean duration of muscle action (DMA) in ms, the mean peak EMG amplitude (PENG) in μV, and standard deviation for the English /b/ and /p/ in unstressed and stressed syllables (n=5, vertical lines = standard deviation).

Figure 3. Diagram showing the mean duration of muscle action (DMA) in ms, the mean duration of oral closure (DOC) in ms and the mean peak EMG amplitude (PENG) in μV for the three types of Korean stops /p,.p',ph/ (N=5, subjects=2, vertical lines = standard deviation).
would be supported by the inter-muscle compensation (Sussman et al., 1973) where one of the two subjects did not use his jaw to any great extent during the speech inventory, even for open vowels. He showed the high degree of negative correlation between jaw elevation for the medial consonant and EMG activity in mentalis to elevate the lower lip for the stop.

**Phoneme-sensitive EMG gestures in Korean.** As shown in Figure 3, the timing variables and the peak EMG amplitude from the orbicularis oris muscle were significantly greater in /p/ /p/ than in /p/, regardless of speaker, while there were insignificant differences in the EMG activities between the two tense stops, except for the duration of oral closure. These findings are agreeable with the previous studies (e.g. for intervocalic position, Kim, 1980; for word-initial position, Kim, 1965). Thus, the hypothesis that the EMG feature 'tensity' distinguishes /p, p' from /p/ in Korean, with the aspirated /p/ and the unaspirated /p/ being tense and the unaspirated /p/ lax, has been verified. As pointed out by Kim (1980), however, you need both 'aspiration' and 'tensity' simultaneously to distinguish the three types of Korean stops as follows:

<table>
<thead>
<tr>
<th>Features</th>
<th>Lax unaspirated (/p, t, k/)</th>
<th>Tense aspirated (/pʰ, tʰ, kʰ/)</th>
<th>Tense unaspirated (/p', t', k'/)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspiration</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Tensity</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

This means that neither aspiration nor tensity is the primary feature. This is the same as "both".

**CONCLUSION**

(1) The hypothesis that the distinctive feature tensity distinguishes /p/ from /b/ in English and /p/, p' from /p/ in Korean, with English /p/ and Korean /p, p' being tense and Korean unaspirated /p/ and English /b/ lax, has been verified, except for the case with the stressed syllable in English. (2) In English, the effects of stress on the peak EMG amplitude was phoneme-sensitive. (3) The inconsistent EMG data within a given subject may be due in part to the phoneme-sensitive EMG manifestations of stress. (4) The inter-speaker variabilities in the EMG data would result partly from the inter-muscle compensation. (5) It was hypothesized that the English /p/ and /b/ in stressed syllables may be differentiated by the EMG activities from a muscle other than the orbicularis oris superior muscle, e.g., the respiratory muscles relating to aspiration and that certain speakers use different muscles from others in the articulation of speech sounds, which may have something to do with a regional or personal accent.

**ACKNOWLEDGEMENT**

This research was supported in part by THE NON-DIRECTED RESEARCH FUND, Korea Research Foundation. The author would like to thank Professor Su-Geun Wang, M.D., the Pusan National University Hospital, for technical assistance.

**REFERENCES**


MODELLING THE EAR'S "EFFECTIVE" SIGNAL PROCESSING TO PREDICT PSYCHOACOUSTICS AND SPEECH PERCEPTION IN NORMAL AND IMPAIRED LISTENERS

Birger Kollmeier, Torsten Dau, Inga Holube, Matthias Wesselkamp
AG Medizinische Physik & Graduiertenkolleg Psychoakustik, Universität Oldenburg, Germany

SUMMARY AND INTRODUCTION

A signal-processing model of the peripheral auditory system is proposed that is capable of predicting psychoacoustical experiments (e.g., temporal and spectral effects in masking, modulation perception) as well as speech intelligibility in quiet and in noise for closed-set and open-set speech tests. The "preprocessing" stage of the model consists of a gammatone filterbank followed in each channel by a nonlinear adaptation loop circuit and an integrator or modulation frequency filter bank. Hearing impairment can be simulated by a threshold-simulating additional noise, a broadening of the auditory filter, and an increase of the time constants in the adaptation loops. The "decision device" of the model consists of an adaptive pattern recognizer which is based on a crosscorrelation between the current input signal and a reference signal. For closed-set speech tests, a DTW speech recognizer is employed as detector stage. Although the model only requires a minimum set of assumptions and free parameters, it has successfully been applied to a variety of psychoacoustical experiments and speech intelligibility tests both for normal and impaired listeners. The applications of this model include a better, quantitative understanding of auditory signal processing and auditory functions in normal and impaired listeners, an improved design of front ends for speech recognizers, and an improved design of "intelligent" digital hearing aids.

MODEL STRUCTURE

Fig. 1 outlines the preprocessing performed before the decision device. A gammatone filterbank is used (Patterson et al., 1987), a half-wave rectifier with 1 kHz lowpass filter for envelope extraction (simulating the limiting phase-locking for auditory nerve fibers at high frequencies), and a set of five consecutive adaptation loops according to Püschel (1988). Each of these loops consists of a divider and an RC-low pass. The input signal is divided by the low-passed filtered output signal. Thus, fast modulations of the envelope pass through unchanged, whereas slow variations of the envelope are compressed by a square root law. Combining five consecutive adaptation loops approximates the logarithm of the average input fairly well. The time constants of the five loops are equispaced between 5 and 500 milliseconds for normal listeners. For impaired listeners, the range of time constants can be adjusted to describe the individual's temporal masking data. In addition, the filter widths of the gammatone filterbank can be changed to describe the individual's masking phenomena in the frequency domain in a correct way. At the output of the adaptation circuits, either an envelope low-pass filter with a time constant of 20 milliseconds (Dau et al., 1995) or a
modulation frequency filterbank with a constant relative filter bandwidth is used for each center frequency (Dau and Kollmeier, 1994). While the envelope low-pass filter characterizes the auditory system's sluggishness in following rapid envelope fluctuations and yields an adequate performance of the model for forward, simultaneous and backward masking, the modulation filterbank is required to correctly account for modulation detection and modulation masking experiments. The structure of this modulation frequency analysis in each center frequency band is motivated by physiological findings of a tonotopical organization of modulation tuning on an axis perpendicular to the tonotopical organization of center frequency. This underlying idea has already been successfully applied to speech enhancement in background noise (Kollmeier & Koch, 1994).

Fig. 1 Sketch of the preprocessing performed in the model

The "internal representation" of a given acoustical signal is thus the sum of the preprocessed signal and an "internal noise" which accounts for processing deficiencies in the auditory system. To model psychoacoustical detection experiments, the internal representation of the masker alone is subtracted from the internal representation of the masker plus current test signal as well as from the masker plus the suprathreshold test signal. The respective difference in internal representation is crosscorrelated in order to estimate the similarity between the "actual" change in the masker pattern produced by the current test signal and the "target" change caused by a clearly detectable signal. If this crosscorrelation exceeds a certain criterion, the model predicts a detection of the current test signal. For predicting speech intelligibility in an open-set speech test, a similar decision device was assumed, but the target word with two different versions of masking noise served as the inputs to the model (Wesselskamp, 1994). For predicting a closed-set speech test, the internal representation of the target word in noise is compared with those of the response alternatives (embedded in a different sample of noise) using a dynamic time warping algorithm (Holube & Kollmeier, 1995). While for modelling psychoacoustics only the output of the center frequency channel with the best signal-to-noise ratio and its subsequent decision unit is modelled, a cumulative similarity across all critical bands involved is calculated for the prediction of speech intelligibility. In all cases, the model is used to predict the performance of the individual subject on a trial-to-trial basis. Thus, a simulation of the performance in a test list requires several hours computation time on a modern Sun Sparc S10 workstation. More details are given by Dau et al. (1995), Holube & Kollmeier (1995) and Wesselskamp (1994).

PREDICTION RESULTS

Fig. 2 gives the predicted (filled circles) and observed (open circles) forward masking data for a 1-kHz, 10-ms tone masked by a frozen, white noise sample of 200 milliseconds duration. The threshold of the tone is given as a function of its delay relative to the masker offset. Note
the close similarity between the data and the observation which is only based on very few assumptions about the "effective" signal processing in the ear and the global detection process. Obviously, the model is capable of describing masking data reasonably well. This also holds for several other psychoacoustical paradigms such as simultaneous and backward masking, temporal gap detection, test tone integration and modulation perception (Dau et al., 1995; Dau & Kollmeier, 1994). A prediction of forward masking and notched-noise data in sensorinéural impaired listeners was also achieved by Holube & Kollmeier (1995). They included a threshold-simulating noise, an increased filterbank bandwidth, and increased time constants individually fitted to each patient’s data.

![Graph showing Masked Threshold vs. Probe onset relative to masker-offset (ms)](image)

**Fig. 2** Observed (filled circles) and predicted (open circles) forward masking thresholds (from Dau et al., 1995).

![Graph showing SVI (%) vs. S/N [dB]](image)

**Fig. 3** Observed (solid line, squares) and predicted (dashed lines, crosses and squares) speech intelligibility scores for normal listeners in CCITT noise as a function of signal-to-noise ratio (from Wesselskamp, 1994).

Fig. 3 shows the measured (solid lines) and predicted discrimination functions (dotted lines and symbols for two different model versions) for a closed-set rhyme test employing continuous, speech-spectrum-shaped noise according to CCITT. Both the absolute position and the shape (i.e., the slope) of the discrimination function is predicted by the model in a remarkable way. Note that no further knowledge about the structure of the word was assumed. Hence, the model only reflects the properties of the preprocessing in the auditory system and the "optimal" detection process that might be performed within the cortex. Further details are given by Wesselskamp (1994).

![Graph showing SVI (%) vs. P] and predicted (dashed lines, crosses and squares) speech intelligibility scores in quiet (left panel) and noise (right panel). The lines mark expected values and standard deviations (from Holube and Kollmeier, 1995).

**Fig. 4** Observed (ordinate),

-83-
Fig. 4 gives the comparison between the predicted (abscissa) and observed (ordinate) closed-set speech intelligibility scores for a group of normal listeners (asterisks) and sensorineurally impaired listeners (other symbols) both in quiet (left panel) and in noise (right panel). The predictions in noise are generally very good if only the exactly measured pure-tone threshold of the individual impaired listeners is included in the model computation by an appropriately shaped noise. Introducing the appropriately modified time constants in the adaptation loops as well as the appropriately modified filter bandwidths in the gammatone filterbank only slightly improves the quality of the predictions. In addition, the prediction is generally very poor for speech intelligibility in quiet. However, if the prediction in quiet is modelled by superimposing the threshold-simulating background noise with a random fluctuation in the overall level of this noise, again a good prediction is obtained for normal and impaired listeners' speech intelligibility in quiet (c.f., left panel of Fig. 4).

CONCLUSIONS

The preprocessing model in combination with the decision units described here appear to be an appropriate and reliable way to predict both psychoacoustical experiments and performance in speech intelligibility tasks both in normal and impaired listeners in a satisfactory way. Although it includes already a number of psychophysical facts and physiologically motivated model structures, it yet cannot predict all different psychoacoustical phenomena as well as different speech intelligibility tests. Thus, the current model has to be evaluated and tested further both with a larger number of impaired listeners and different psychoacoustical and speech perception tests. Overall, the model seems to yield a better understanding of the „effective“ signal processing both in normal and impaired listeners. This might be useful for the construction of, e.g., „intelligent“ digital hearing aids and preprocessors for speech recognition systems.

ACKNOWLEDGEMENTS

Work supported by DFG. Thanks to A. Sievers for typing the manuscript.

REFERENCES


PREDICTION OF SPEECH TRANSMISSION QUALITY OF TELEPHONE HANDSETS USING SHORT-TIME SUBBAND ANALYSIS AND PSYCHOACOUSTIC MODELS

Winfried Kreberger and Jörg Becker
Institut für Elektrische Nachrichtentechnik, Aachen Univ. Technology, Germany

SUMMARY

A predictor is presented which estimates the mean opinion score (MOS) for linear distorted speech in noisy environments, e.g. speech transmitted via a telephone handset applied to the human ear. The critical band rate excitation pattern is calculated in 50 ms time segments for the original speech signal, the speech signal measured in the ear canal, and the ambient noise. For each time segment three psychoacoustic parameters are calculated. An intelligibility index (AI) is evaluated using SNR analysis considering simultaneous masking effects. Naturalness (NI) is estimated by spectral distance between original and distorted speech. The loudness index (LI) is defined by a trapezoid function of the 10%-percentile value of speech loudness which is calculated similar to ISO 532. The MOS is predicted as a weighted sum of AI, NI and LI. The prediction results were validated by opinion tests. Correlation coefficients between $p = 0.93$ and $p = 0.96$ were achieved.

INTRODUCTION: PREDICTION OF MOS

In this contribution a model is presented which predicts the mean opinion score (MOS) for linear distorted speech in noisy environments. A typical application is the estimation of speech transmission quality of telephone handsets [Kre95]. Three main aspects of speech quality - intelligibility, naturalness and loudness - are considered. In contrast to the assessment of codecs, especially loudness is very important for the assessment of handsets. Bandwidth and level of the speech signal in the ear canal are influenced by the position of the handset referring to the ear. In addition, ambient noise is transmitted to the ear canal via the acoustic leakage between handset and ear.

Fig. 1 shows a block diagram of the prediction model. Critical band rate excitation patterns, approximated by one value per critical band [ZF90], are calculated in 50 ms time segments of the original speech signal, the (linear distorted) speech signal measured in the ear canal of an artificial ear, and the additive room noise measured in the ear canal.

Based on the time-dependent excitation patterns three parameters are calculated: Intelligibility index AI, naturalness index NI and loudness index LI. They are described in the following sections. The MOS value is estimated by a weighted sum:

$$\text{MOS} = w_0 + w_{AI}AI + w_{NI}NI + w_{LI}LI(N_{10}, AI).$$ \hfill (1)
INTELLIGIBILITY INDEX

The intelligibility index (AI) used in this contribution is derived from the articulation index [FS47] and is defined as

$$AI = \sum w_i r_i,$$

where $w_i$ is a subband weighting coefficient, $r_i$ denotes a monotonic increasing function of the effective SNR$_i$ in each subband $i$ [FS47]. For the calculation of the effective SNR$_i$ the following masking effects are considered:

- Residual masking, modelled by an additive internal noise [Pav87],

- Masking by external additive noise,

- Self masking: The effective SNR$_i$ is limited to 24 dB,

- Interband masking,

- Excessive masking for high speech levels.

French and Steinberg [FS47] used 20 subbands which cover the frequency area 250 - 7000 Hz with $w_i = 0.05$ for each subband. The band limits were defined empirically. The presented model uses 21 critical bands (No. 2 - 22 according to [ZF90]). The weighting coefficients $w_i$ are simplified as proposed by Pavlovic [Pav87] for normal speech. The use of total excitation levels includes the effects of interband and excessive masking automatically.
NATURALNESS INDEX

The naturalness index NI reflects only those aspects of naturalness which depend on a kind of spectral distance. For the assessment of speech codecs often spectral distances between original and distorted speech signal are measured in each time segment [QBC88]. The total distance is defined as time average over all time segments (speech pauses excluded). A simple, but effective naturalness index NI can be derived from the absolute distances $d_i$ (in dB) between the total excitation levels of original and distorted speech signal in each subband $i$:

$$\text{NI} = \frac{1}{2} \sum_{i=2}^{2L} w_i \max(1 - d_i / d_{\text{max}}, 0).$$

(3)

This measure can be enhanced if $w_i$ prefers symmetrical transmission characteristics:

$$w_i = \begin{cases} 0.5 & : i \geq 10 \quad \text{and} \quad d_{20-i} > 10\text{dB} \\ 1 & : \text{otherwise} \end{cases}$$

(4)

LOUDNESS INDEX

Loudness describes a sensation, which depends on the level and the frequency spectrum of sound. Psychoacoustic models and algorithms for the calculation of loudness are defined in ISO 532 (see also [PZ72]). The calculation of loudness used for this investigation is very similar to ISO 532, with some modifications introduced by Aures [Aur84].

For speech signals a time-averaged loudness evaluation does not reflect the human signal processing [ZF90], because the perception of speech loudness is dominated by the time segments with maximum loudness. A good approximation for the perceived loudness of speech is given by that loudness value which is exceeded in 10% of time, called $N_{10}$.

In auditory speech quality assessment the MOS decreases for low and high loudness values, because low loudness reduces intelligibility and high loudness reduces comfort. These effects are considered by the loudness index $\text{LI}(N_{10})$ using a trapezoid function (fig. 2). LI is set to 0 for loudness values $N_{10}$ between $N_1 = 15$ sone and $N_2 = 45$ sone. $\text{LI}(N_{10})$ decreases for $N_{10} < N_1$ or $N_{10} > N_2$. The steepness of $\text{LI}(N_{10})$ is set to $s_L = 0.1\text{sone}^{-1}$ for $N_{10} < N_1$. Since high loudness values are less disturbing in connection with low intelligibility, the steepness $s_L$ depends on the intelligibility index AI for $N_{10} > N_2$, (see fig. 2).

![Figure 2: Trapezoid function LI($N_{10}$)](image_url)
COMPARISON TO OPINION TESTS

The prediction results were compared to the results of an opinion test using totally 442 speech samples. Speech samples of two speakers were filtered simulating typical transfer characteristics of wideband and narrowband telephones and presented in a noisy environment. Speech level and SNR were varied in a wide range\[1][Kre94].

For the prediction the following parameter values are used: \( w_0 = 0.00, w_{ai} = 2.78, w_{SS} = 2.43, w_{LB} = 2.00, d_{max} = 40 \text{ dB}, N_1 = 15 \text{ sone}, N_2 = 45 \text{ sone}. \)

Over all the comparison proves a high correlation between measurement and prediction. A correlation coefficient \( p = 0.93 \) for all samples and \( p = 0.96 \) for speech levels below 76 dB SPL was achieved. 67\% (resp. 80\% for speech levels below 76 dB SPL) of all predicted values are within the confidence interval of the measured MOS values.

CONCLUSIONS

The algorithm presented in this paper considers three main aspects of speech quality: intelligibility, naturalness and loudness. The MOS can be predicted for linear distorted speech in noisy environment, achieving a high correlation to results of opinion tests.

To get quality criteria for handsets only the acoustic environment has to be defined. If speaker and ambient noise are well-defined the prediction values can be regarded as measure for the speech transmission quality of handsets.

REFERENCES


\[1\] The following parameters were used: Speech levels: 52, 64, 76, 88, 100 dB SPL, SNR, 12, 24, 36 dB, totally 13 combinations. Filters: lowpass 3.4, 4.8, 7.0, 11.0 kHz, ideal bandpass 100, 150, 190, 240, 300 – 3400 Hz, bandpass (18 dB/octave) 400, 500, 630, 790, 1000, 1260 – 3400 Hz, 500, 1000 – 7000 Hz. Speakers: 1 male, 1 female.
DETECTION D’ÉVÉNEMENTS IMPULSIONNELS EN PAROLE

Francois Malbos, Marc Baudry
Laboratoire d’Informatique de l’Université du Maine (LIUM)
Avenue O. Messiaen, B. P. 535, 72017 Le Mans Cedex, France.

ABSTRACT
The aim of this work is the detection of impulsive events in speech. In order to minimise the temporal localization error, the wavelet transform is used. Elaborated and validated for the burst detection of the unvoiced stop consonants, the detection algorithm is sensitive to several discontinuities: the glottal pulses of vowels and the occlusion relaxation of nasal consonants. Apart from this, our system is interesting for two reasons: in opposition to the Short Time Fourier Transform, the localization error only depends on the frequency structure of the discontinuity and it allows the measurement of impulsive discontinuity characteristics.

INTRODUCTION
Le but d’un module de segmentation relatif à un système de décodage acoustico-phonétique est la localisation de discontinuités acoustiques qui correspondent à la transition entre deux entités élémentaires (phonème, diphonème, syllabe, …) ou à une transition infra-phonémique (barre d’explosion des occlusives). Dans ce cadre et dans une étude précédente, nous avons élaboré un système de détection des occlusives en modélisant leur barre d’explosion par une impulsion (Malbos[94]). Le but de cet article est de présenter d’une part, la diversité des discontinuités relatives au signal de parole que l’on peut mettre en évidence à l’aide de cet algorithme et d’autre part, les intérêts de notre travail en termes de localisation temporelle de la discontinuité et de l’évaluation de son caractère plus ou moins impulsionnel.

PRINCIPE DE LA METHODE
Un des buts de cette étude étant la localisation temporelle des discontinuités, nous avons choisi la transformée en ondelettes. En effet, possédant une résolution fréquentielle relative constante, elle est adaptée pour la détection de discontinuités (Montréal[91], Petropulu[92]). Cette transformée permet de décomposer linéairement un signal en une superposition de signaux élémentaires déduits les uns des autres par translation et dilatation d’une ondelette mère ψ(t) (Flandrin[93]). Notant respectivement a (a>0) et b les facteurs d’échelle et de translation temporelle, les coefficients d’ondelette S(a,b) sont calculés en projetant le signal sur la famille de fonctions ψ_{a,b}(t) telles que :

\[ \psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \]  \hspace{1cm} (1)

Afin de détecter et de mesurer le caractère impulsionnel des discontinuités présentes au sein du signal de parole, nous avons évalué à l’aide de fonctions de corrélation et pour chaque voie de la décomposition fréquentielle, l’indice de similarité entre le module de la transformée en ondelettes du signal de parole et celui de l’ondelette analysante. Parmi les ondelettes décrites dans la littérature, celle de Morlet notée \psi_M(t) est à la base de notre algorithme. Elle s’écrit de la façon suivante :

\[ \psi_M(t) = e^{-i\nu t^2} e^{-i\omega_0 t} \hspace{1cm} \text{avec} \hspace{1cm} \omega_0 = 5,5 \]  \hspace{1cm} (2)
Utilisée plus largement dans un cadre de détection et d'identification des occlusives, le choix de cette ondelette est légitime. En effet, elle minimise le principe d'incertitude de Weyl-Heisenberg (Chui[92]). De plus, en raison de sa forme analytique, elle permet l'accès à l'information présente au sein du module de la transformée en ondelettes. Pour le calcul des coefficients d'ondelette, nous avons choisi l'algorithme rapide de Barrat (Barrat[92]). Il apprivoise l'ondelette de Morlet par une fonction \( \psi_B(t) \) telle que :

\[
\psi_B(t) = (1 + \sigma || t || e^{-\frac{c^2 || t ||^2}{2}} e^{i \omega t})
\]

avec \( c' = 5 \) et \( c = 1.5 \) \( \text{(3)} \)

Une telle approximation permet de calculer les coefficients d'ondelette à l'aide d'un filtre à réponse impulsionnelle infinie. Se référant à la figure 3, il apparaît que dans le cadre de la détection de discontinuités, les incertitudes temporelles des deux ondelettes développent des équivalences en haute fréquence. Afin de s'affranchir de la différence énergétique entre les différents phonèmes, les fonctions de corrélation \( C_{ij} \), où l'indice \( i \) est le numéro de la voie d'analyse, sont normalisées en énergie. Chaque fonction de corrélation \( C_i(t) \) présente un maximum \( M_j \) synchronisé avec la discontinuité. De plus, ce maximum est précédé d'un minimum \( m_i \). Cette propriété résulte de la normalisation énergétique et de la forme gaussienne de l'ondelette analysante. L'étude des fonctions de corrélation dans le cadre de la barre d'explosion des occlusives a permis de mettre en évidence que ces deux extrêmes sont séparés temporellement par une loi proportionnelle au facteur d'échelle de la transformée en ondelettes (Malboe[95]).

D'une façon identique à celle adoptée pour la détection des occlusives, cette propriété a permis de calculer le signal \( SC(t) \) défini par :

\[
SC(t) = \frac{1}{N} \sum_{i=1}^{N} C_i(t - \tau_i)
\]

où \( N \) est le nombre de voies de la décomposition fréquentielle et \( \tau_i \) une loi proportionnelle au facteur d'échelle \( \sigma \) de la transformée en ondelettes.

L'étude menée dans le cadre de la détection des occlusives a mis en évidence que les minima locaux \( m \) de la fonction \( SC(t) \) présentant un fort écart d'amplitude avec la valeur moyenne du signal \( SC(t) \) sont pertinents. Si ces extrêmes ne permettent pas une localisation précise de la discontinuité, celle-ci peut être atteinte par le calcul de la somme synchrone de l'ensemble des fonctions de corrélation \( (\tau_i = 0) \).

**DIVERSITE DES RUPTURES DETECTEES**

Il est important de souligner que les minima décrits précédemment permettent de localiser différents types de ruptures : discontinuité entre un silence et le début d'un phonème, relâchement de l'occlusion d'une consonne nasale, instant de rupture du contact alvéo-dentale du phonème \( N \), barre d'explosion des occlusives sourdes et sonores. Quelques exemples sont présentés à l'aide des figures 1 et 2. Pour chacun d'entre eux, les figures A, B et C sont respectivement associées à la représentation temps-amplitude du signal de parole, au module au carré de la transformée en ondelettes (scalogramme, quart d'octave sur 6 octaves entre 93 Hz et 5 kHz) et la fonction la fonction \( SC(t) \) calculée pour 16 voies d'analyse réparties au quart d'octave sur quatre octaves. Pour la figure 1, trois types de discontinuité apparaissent nettement. Elles correspondent à :

- une série de bruits impulsionnels de très faible amplitude relatifs à un silence (minima (a)),
- le relâchement de l'occlusion de la consonne nasale /\( m/\) (minimum (b)),
- les impulsions glottiques relatifs à la voyelle /\( a/\) (minima (c)).

Pour la figure 2, la barre d'explosion de l'occlusive sonore apparaît nettement à l'aide du minimum (d).
LOCALISATION TEMPORELLE

L'incertitude de localisation temporelle du module de détection est relative à la structure de la discontinuité. Plus précisément, elle correspond à celle de la voie de la transformée en ondelettes de plus haute fréquence pour laquelle la discontinuité présente une énergie significative. Une telle propriété justifie le choix de la transformée en ondelettes par rapport à la transformée de Fourier à court terme. En effet, pour celle-ci, l'incertitude temporelle est relative à la taille et à la forme de la fenêtre de pondération (Coulon[84]). L'utilisation de notre algorithme pour la détection des occlusives a permis d'estimer l'erreur à [0,2-1] ms pour les labiales, [0,2-0,8] ms pour les alvéolaires et [0,2-0,3] ms pour les dentales (Malbos[95]). Ces valeurs sont inférieures à celles des systèmes basés sur l'analyse de Fourier pour lesquels l'incertitude est supérieure à 1,2 ms. Pour une discontinuité et une fréquence d'échantillonnage déterminées, l'incertitude peut être estimée à l'aide de la figure 3.
Mesure du caractère impulsionnel

L'amplitude du minimum m du signal SC traduit l'augmentation d'énergie dans l'ensemble des voies de la décomposition fréquentielle. Ainsi, une discontinuité sera d'autant plus impulsionnelle que l'amplitude du minimum m sera faible. Cette propriété a été mise en évidence dans le cadre de la détection des occlusives (Malbos[95]). En effet, respectivement 98,9% et 67,9% des occlusives sourdes et sonores présentent un minimum m au sein du signal SC dont l'amplitude est inférieure à un seuil égal à 0,30. Une telle évaluation du caractère impulsionnel apparaît nettement pour les vibrations laryngées des voyelles /a/ et /u/ des figures 1 et 2.

CONCLUSION

Cette étude permet de généraliser l'utilisation d'un détecteur conçu pour la détection de la barre d'explosion des occlusives à d'autres discontinuités relatives au signal de parole. Outre la diversité des événements détectés, l'intérêt du système réside d'une part, dans l'incertitude temporelle de localisation de la discontinuité. D'autre part, la méthodologie proposée permet de mesurer le caractère plus ou moins impulsionnel de l'événement détecté. L'orientation de notre travail pour la segmentation du signal de parole nécessitera de générer l'éventuelle sous-segmentation de certains phonèmes. Une telle étape peut être envisagée à l'aide d'un critère entropique (Coifman[92]). En raison de la sensibilité de notre module, il serait particulièrement intéressant de l'évaluer dans d'autres domaines : la restauration d'enregistrements anciens, l'analyse de signaux sismiques et de signaux médicaux.

REFERENCES BIBLIOGRAPHIQUES

AN IMPROVED SPECTRAL SUBTRACTION WITH SMOOTHING FOR NOISY SPEECH RECOGNITION

Hiroshi Matsumoto, Yoshinori Nimura
Faculty of Engineering, Shinshu University, 500 Wakasato, Nagano-shi, Japan

ABSTRACT

This paper proposes a spectral smoothing method to improve the spectral estimate obtained by a standard spectral subtraction method. The smoother intends to interpolate the low SNR spectral components using high SNR components by fitting cosine series based on a weighted minimum mean square error criterion. This method is evaluated through both DTW- and HMM-based word recognition tests using the NOISEX-92 database after down sampling to 8 kHz. The results show that the smoother is effective under very noisy conditions, and that the combination of this method and the frequency-weighted HMM achieves the highest overall performance, for example, giving 96% in accuracy at 0dB SNR of car noise.

1. INTRODUCTION

In order to develop reliable speech recognition systems in noisy environments, it is important to suppress noise components prior to recognition phase and also to make patterns matching itself robust to interfering noise. While noise adaptation such as Parallel Model Combination (PMC) [1] is another useful approach, additive noise reduces discrimination among models in very noisy conditions. Therefore, this study concerns with the former approach.

Among speech enhancement techniques, the spectral subtraction (SS) has proved to be effective and efficient noise reduction methods [2],[3]. However, estimated spectra tend to have discontinuity since each spectral component is independently processed. In order to improve this drawback, this paper proposes a spectral smoothing method which fits cosine series to an estimated log-power spectrum based on a minimum mean square error weighted by SNR dependent function. This smoother intends to interpolate low SNR spectral components from high SNR ones. Furthermore, this noise reduction method is combined with a robust HMM and a distance measure based on a frequency weighting technique [4]. These recognizers are evaluated on the NOISEX-92 database.

2. SPECTRAL SUBTRACTION WITH SMOOTHING

In this study, we use a p-channel filter bank in speech analysis. A standard spectral subtraction is applied to noisy speech power spectrum \( \mathbf{z} = [z_1, z_2, ..., z_p] \). The clean speech power spectrum \( \mathbf{\hat{z}} = [\hat{z}_1, \hat{z}_2, ..., \hat{z}_p] \) is estimated by

\[
\hat{z}_k = \max \{ z_k - \alpha \hat{N}_k, \gamma \hat{N}_k \}
\]

where \( \hat{N}_k \) is the estimated noise power spectrum in the \( k \)-th channel, and \( \alpha \) the overestimation factor, and \( \gamma \hat{N}_k \) the spectral floor. This spectral subtraction doesn’t take account of the correlation between neighboring spectral components. Therefore, the estimated spectra tend to have discontinuity in low SNR components due to variation of noise.
To improve this drawback in spectral subtraction, unreliable estimates \( \hat{x}_k \)'s in low SNR are interpolated in the log spectral domain using higher SNR components by fitting a smoothed log spectrum \( \log \hat{\mathbf{z}} = \mathbf{C}^{-1} \mathbf{z}^c \), where \( \mathbf{C} = [c_{ij}] \) represents the discrete cosine transform (DCT) and \( \mathbf{x}^c \) is an unknown cepstral vector. The \( \mathbf{x}^c \) is estimated to minimize

\[
J(\mathbf{x}^c) = \sum_{k=1}^{p} (\log \hat{x}_k - \log \hat{\mathbf{z}}_k)^2 W_k.
\]  

(2)

The weighting coefficient \( W_k \) is a monotonically increasing function of the estimated SNR and is bounded between 0 and 1. In this study, we choose the following Wiener filter as such weighting coefficient:

\[
W_k = \max \left\{ \frac{\hat{x}_k - \beta \hat{N}_k}{x_k}, 0 \right\}
\]  

(3)

where \( \beta \) is the overestimation factor for weighting. Thus, lower SNR components are interpolated depending on their SNR using higher SNR components. The estimated \( \hat{\mathbf{x}}^c \) is given by the normal equation \( \mathbf{A}\hat{\mathbf{x}}^c = \mathbf{b} \), where \( \mathbf{A} = [a_{ij}]_{p \times p} \) and \( \mathbf{b} = [b_i]_{1 \times 1} \) are given by

\[
a_{ij} = \sum_{k=1}^{M} c_{ik} c_{jk} W_k \quad \text{and} \quad b_i = \sum_{k=1}^{M} c_{ik} W_k \log \hat{g}_k.
\]  

(4)

As the global SNR becomes lower, the effective rank of \( \mathbf{A} \) decreases. Thus, we constrain \( \hat{\mathbf{x}}^c \) to minimize the norm. The \( p \times p \) real symmetric matrix \( \mathbf{A} \) is decomposed as \( \mathbf{S} \mathbf{A} \mathbf{S}^T \) with \( \mathbf{S} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_p] \) and \( \mathbf{A} = \text{diag}[\lambda_1, \lambda_2, ..., \lambda_p] \), where \( \mathbf{u}_r \) is the eigenvector of \( \mathbf{A} \) associated with the \( r \)th largest eigenvalue \( \lambda_r \). The minimum norm least square estimate \( \hat{\mathbf{x}}^c \) is obtained by

\[
\hat{\mathbf{x}}^c = \mathbf{S}^T \mathbf{A}^{-1} \mathbf{S} \mathbf{b}
\]  

(5)

where

\[
\mathbf{A}^{-1} = \text{diag}[\lambda_1^{-1}, ..., \lambda_p^{-1}, 0, ..., 0]
\]  

(6)

and the rank \( r \) of \( \mathbf{A} \) is determined by a threshold \( \theta \) such that \( \lambda_r > \theta \lambda_1 > \lambda_{r+1} \). Therefore, the smoothness of \( \log \hat{\mathbf{x}} \) can be adjusted by the rank reduction threshold \( \theta \).

3. ROBUST DISTANCE METRICS AND HMMs

3.1 Distance Metrics

In the DTW-based recognition system, a standard cepstral distance (CEP) and a weighted-group delay difference (WGD) are used [4]. The WGD, which is originally defined in LPC-spectra, is extended to filter bank spectra. Both distance measures are defined in Table 1.

In this table, the subscripts, \( f \) and \( g \), signify the reference and test patterns, and the vector \( \mathbf{x}^d \) represents the pseudo-group-delay spectrum defined by

\[
\mathbf{x}^d = \mathbf{C}^{-1} \mathbf{Q} \mathbf{x}^c,
\]  

(7)

where \( \mathbf{Q} \) is the quefrency weighting matrix \( \text{diag}[1, 2, ..., p] \).

<table>
<thead>
<tr>
<th>Dist.</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEP</td>
<td>(</td>
</tr>
<tr>
<td>WGD</td>
<td>(\mathbf{x}_f^d - \mathbf{x}_g^d)^T (\mathbf{x}_f - \mathbf{x}_g))</td>
</tr>
</tbody>
</table>

Table 1 The distance measures.

3.2 Robust HMMs

In the HTMM-based recognition, the pseudo-group-delay spectrum \( \mathbf{x}^d \) is used as an observation vector to utilize its robustness to noises. The two types of single mixture continuous density HMMs with a fixed diagonal covariance are used: a grand variance HMM (HMM-GT) and the frequency weighted HMM [5].

In HMM-GT all the covariances are fixed to the 'grand diagonal covariance' \( \Sigma_G^d \) over all the training speech [6]. In the frequency-weighted HMM all the covariances are replaced to
a frequency-weighting matrix \( W = \text{diag}[w_1, w_2, \ldots, w_n] \). The frequency-weighting coefficients \( w_k \)'s are derived from the grand mean \( \mu_0^d \) of the pseudo-group delay spectra as follows:

\[
w_k = \eta \exp\{\nu \tilde{\mu}_k^l\}, \tag{8}\]

where \( \tilde{\mu}_k^l \) is the \( k \)th component of the smoothed log power spectrum given by

\[
\tilde{\mu} = C^{-1}LC \mu_0^d \tag{9}\]

with \( L = \text{diag}[1, 1/2, \ldots, 1/q, 0, \ldots] \), and \( \nu \) is a compression factor and \( \eta \) a scale factor. Therefore, the covariance in this model is not estimated statistically, but instead is derived based on a prior knowledge on the perceptual importance in frequency domain and/or expected variance due to degradation of speech.

In training both HMMs, the state means and the transition probabilities are re-estimated with the fixed variances.

4. Evaluation

4.1 Data Base and Speech Analysis

In speech analysis, a 15-channel uniform filter bank system with a flat composite spectrum was implemented at a sampling rate of 8kHz using the same design procedure as described by Dautrich, Rabiner, and Martine [7]. The spacing of channel center frequencies and the band width of each bandpass filter were set to 250Hz (8/32kHz) and 300Hz, respectively. First, the speech signal was preemphasized with \( (1 - 0.98 \tau^{-1}) \). The output of each bandpass filter was followed by a square law detector and a moving average filter of 12 points. The channel outputs were sampled at every 10ms.

Two stationary background noises were used: car and white noises. The degraded speech by car noise was used from the NOISEX-92 database. On the other hand, the white noise was generated in a computer, and was added to clean speech so that the global SNR for each word is equal to a predetermined value. Each word was represented by a left-to-right HMM with 26 states. The HMM models were trained using 10 repetitions of noise-free samples. Another set of ten utterances was used for testing. The Viterbi algorithm was used for testing. The beginning and end points were fixed to those in the label files. Thus, only substitution errors were scored.

4.2 Optimization of Baseline System

First, in order to optimize the parameters in the frequency-weighted HMM, the scale factor \( \eta \) and the compression factor \( \nu \) were examined. The smoothing order of \( q \) in equation (9) was set to 8 throughout this study according to a preliminary experiment.

Table 2 shows the effect of the normalized scale \( \tilde{\eta} \) with \( \nu = 0 \) at 12 dB SNR of white noise and at 6 dB SNR of car noise. The normalized scale \( \tilde{\eta} \) is defined by \( \{W^{-1}/[\Sigma_0^d]\}^{1/\nu} \). It is clear that the larger scale significantly improves recognition accuracy and the optimum value of \( \eta \) is between 50 and 150 depending on SNR and the types of noises. In the case of \( \nu \), it was found that the larger \( \nu \) tends to slightly improves the recognition accuracy for white noise, but to reduces it for car noise. Thus, in the subsequent experiments, the \( \eta \) is set to 100 for all the noise conditions, and the \( \nu \) is set to 1.0 for white noise and to 0.0 for car noise.

4.3 Smoothing Effect

Before the experiments on the spectral subtraction with and without the smoother, the parameters involved were examined. As the results of the preliminary experiments, the overestimation factor \( \alpha \) was set to 2 over all the noise conditions, and the rank reduction threshold was set to 0.001 and \( \beta \) was set to 2. Table 3 (a) to (d) compare the recognition rates obtained by each of DTW- and HMM-based recognizers with or without the spectral subtraction (SS) and the smoother (SM) in both noise conditions. In the case of the simplest
Table 3 Effects of the SS with and without the smoother in the four recognizers.

(a) CEP

<table>
<thead>
<tr>
<th>SNR [dB]</th>
<th>White</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>w/o SS</td>
<td>25</td>
<td>31</td>
</tr>
<tr>
<td>SS</td>
<td>42</td>
<td>77</td>
</tr>
<tr>
<td>SS+FS</td>
<td>74</td>
<td>94</td>
</tr>
</tbody>
</table>

(b) WGD

<table>
<thead>
<tr>
<th>SNR [dB]</th>
<th>White</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>w/o SS</td>
<td>16</td>
<td>38</td>
</tr>
<tr>
<td>SS</td>
<td>68</td>
<td>99</td>
</tr>
<tr>
<td>SS+FS</td>
<td>88</td>
<td>98</td>
</tr>
</tbody>
</table>

(c) HMM-GT

<table>
<thead>
<tr>
<th>SNR [dB]</th>
<th>White</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>w/o SS</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>SS</td>
<td>48</td>
<td>78</td>
</tr>
<tr>
<td>SS+FS</td>
<td>56</td>
<td>83</td>
</tr>
</tbody>
</table>

(d) HMM-PW

<table>
<thead>
<tr>
<th>SNR [dB]</th>
<th>White</th>
<th>Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>w/o SS</td>
<td>50</td>
<td>67</td>
</tr>
<tr>
<td>SS</td>
<td>80</td>
<td>99</td>
</tr>
<tr>
<td>SS+FS</td>
<td>80</td>
<td>99</td>
</tr>
</tbody>
</table>

distance measure CEP, the smoother is very effective at very low SNR's and increases the recognition rates by 20% to 30% as compared with the standard spectral subtraction. For the WGD and the HMMs, since these recognizers themselves achieve higher recognition rates without the smoother, the smoother is less effective but still better than the standard SS. For instance, the smoother increases the recognition rate for the WGD from 68% to 88% at 0dB SNR of added white noise, and for HMM-PW from 68% to 78% at -6dB SNR of added car noise. The highest overall performance was attained by the HMM-PW combined with the SS followed by the smoother.

5. CONCLUSIONS

This paper has presented the spectral smoothing method to remove the spectral distortion caused by the spectral subtraction. The frequency-weighted HMM combined with the spectral subtraction followed by the proposed smoother achieved the highest recognition performance under most of noise conditions. For non-stationary noises, although further evaluation is necessary, it might be expected that the proposed method together with the frequency-weighted HMM would be robust to such noises.

REFERENCES

4) H.Matsumoto and H.Mitsui, "A robust distance measure based on group delay difference weighted by power spectra", in Proc. ICSLP-90, Kobe, pp.267-270, Nov. 1990.
MEASUREMENT AND ANALYSIS OF 3D SHAPES OF VOCAL TRACT, DENTAL CROWN AND NASAL CAVITY USING MRI

Masafumi Matsumura, Takuya Niikawa, Koji Shimizu*, Yasuji Hashimoto*, Tatsuya Morita**
Osaka Electro-Communication University, Neyagawa, Japan
* Medical Systems Division, Shimadzu Corporation
** Faculty of Engineering, Osaka University

SUMMARY

This paper deals with the measurement of three-dimensional (3D) shapes of vocal tract, dental crown and nasal cavity using magnetic resonance imaging (MRI). 3D MR images of the vocal tract and nasal cavity of 3 adult males were obtained during the steady-state production of Japanese vowels and fricatives. Since it was difficult to observe profiles of dental crowns that contained a small amount of water using MRI, profiles of these were obtained using a dental crown plate that was shaped so as to tightly attach to the subject's dental crown by thermoforming and that contained a contrast medium for MR imaging. The measurement error of the dental crown from the 3D MR image was less than 5.7%. Vocal tract area functions are estimated from the 3D vocal tract shapes, and then formants of the vowels are computed. The results indicated that formant frequencies of the vowels computed from the 3D vocal tract shapes showed good agreement with the subject's original productions. The 3D shapes of the vocal tract during the production of fricatives /s, sh, shu/, the frontal sinus, the sphenoidal sinus and the maxillary sinus were observed.

INTRODUCTION

Magnetic resonance imaging (MRI) capable of noninvasively obtaining transverse, coronal and sagittal sections of the body is widely used in medical examinations. Rokkaku et al. applied MRI to measure a three-dimensional (3D) vocal tract shape [1]. Recently, the MR system has been performed high-quality imaging by superconductive magnet, and high-speed imaging by advanced image processing. Analysis and modeling of the vocal tract shape and dimensions obtained from MR images have been reported [2-4]. Also, a computer algorithm for shape estimation of arbitrary sections of the vocal tract and nasal cavity from 3D MR images has been proposed [5].

However, it is difficult to observe profiles of the dental crown that contains a small amount of water using the MRI. Due to this disadvantage, there are insufficient data on 3D shapes of the vocal tract during the production of consonants. The volume occupied by the teeth was modified using MR images of dental impressions [6]. However, no data on 3D shapes of the vocal tract during the production of consonants such as fricatives /s/ have been obtained. Also, few data on 3D shapes of the vocal tract and nasal cavity have been obtained for the analysis of acoustical coupling of the vocal tract and nasal cavity at the velum.

This paper deals with the measurement of 3D shapes of the vocal tract, dental crown and nasal cavity during the production of Japanese vowels and fricatives /s/, using MRI. A method of simultaneously obtaining MR images of the dental crown and vocal tract using a dental crown plate is proposed. 3D shapes of the vocal tract, dental crown and nasal cavity were obtained from 3D MR images using a previously reported algorithm [5]. The validity of the method is confirmed by comparison of the profile of the dental crown obtained from MR images with dimensions of the subject's dental crown impressions. Vocal tract area functions are estimated from the 3D vocal tract shapes, and then formants of the vowels are computed. The 3D shapes of the vocal tract during the production of fricatives, the frontal sinus, the sphenoidal sinus and the maxillary sinus were observed.
METHODS

All MR images in this study were collected using a 1.0 Tesla superconductive MR system (MAGNEX100HP, Shimadzu Corp., Japan). Transverse and coronal MR images were measured by multislice T1-weighted spin echo imaging methods. Each image was acquired with the repetition time TR=1300ms and the echo time TE=20ms, using the image matrix of 256×256 over a field of view of 25cm. 3D MR images consisted of MR images of 41 transverse sections from the larynx to the nasal cavity at 4-mm intervals and 32 coronal sections from the tip of the nose to the atlas at 4-mm intervals. Measurement time was 142 seconds for each 3D MR image. Section thicknesses of each excited plane for the transverse and coronal MR images were 5 and 3.5 mm, respectively. The 3D MR images and the vowel output uttered by a supine subject in the MR system were obtained simultaneously during steady-state production of Japanese vowels and fricatives. Experiments were performed on 3 Japanese adult males with no history of speech disorders. The subjects' voices were recorded using a high-sensitivity condenser microphone placed inside the magnet set 20 cm from the subjects' lips. Vowel and fricatives utterances were sustained for about 2-3 seconds and intermittent utterances were sustained for about 142 seconds.

Data on dental crown shapes are required for the analysis of fricative production and for precise estimation of the acoustical characteristics of the vocal tract. MR images of a dental crown containing a small amount of water were obtained using a dental crown plate that was shaped so as to tightly attach to the subject's dental crown by thermoforming and that contained a contrast medium for MR imaging. The thickness of the dental crown plate was 0.6mm. By attaching the dental crown plate to the subject's upper and lower teeth, it became possible to obtain profiles of the vocal tract and dental crown simultaneously using MRI. 3D shapes of the vocal tract and nasal cavity are obtained from the 3D MR images, using a previously reported algorithm [5]. Tracing errors generated by the algorithm are corrected by hand tracing.

RESULTS

Figure 1 shows an example of a sagittal MR image. The results show that a profile of the dental crown is clearly obtained using the dental crown plate. Distance W between both of the first molar teeth and H between the first molar tooth and the incisor are measured from a dental crown shape of transverse section. Measurement errors of W and H obtained by comparing with dimensions of subject's dental impression were 4.7% and 5.7%, respectively. Because the thin dental crown plate is low in thickness and tightly attached to the upper and lower teeth, there was essentially no effect of the dental crown plate in place on the speech production. The results show that the method using the dental crown plate to measure upper and lower teeth profiles is useful.

Figure 1. A sagittal MR image obtained during the production of Japanese fricatives /s/.  
- 98 -
Vocal tract area functions of the Japanese vowels /a,i,u,e,o/ are obtained from the 3D vocal tract shapes. Formant frequencies of the vowels are computed from the vocal tract area functions without sinuses by a method proposed by Sondhi and Schroeter [7]. The formant frequencies computed from the 3D vocal tract shapes and the ones estimated from the subject's original productions are listed in Table 1. Except for F1 and F3 of /i/ (22Hz, 7.8% ; 124Hz, 5.8%), F1 of /e/ (18Hz, 6.5%) and F1,F2 and F3 of /o/ (15Hz, 5.3% ; 44Hz, 5.4% ; 125Hz, 6.9%), differences between the values computed from the 3D vocal tract shapes and the ones estimated from subject's utterances are all less than 5%.

Table 1. Formant frequencies of vowels in Hz.
(Subject's original utterances / 3D vocal tract shapes)

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>546 / 568</td>
<td>1135 / 1098</td>
<td>2202 / 2297</td>
<td>3586 / 3415</td>
</tr>
<tr>
<td>i</td>
<td>279 / 257</td>
<td>1440 / 1393</td>
<td>2141 / 2265</td>
<td>3544 / 3441</td>
</tr>
<tr>
<td>u</td>
<td>320 / 336</td>
<td>1348 / 1393</td>
<td>2235 / 2299</td>
<td>3560 / 3537</td>
</tr>
<tr>
<td>e</td>
<td>278 / 296</td>
<td>1267 / 1308</td>
<td>2131 / 2171</td>
<td>3572 / 3549</td>
</tr>
<tr>
<td>o</td>
<td>284 / 299</td>
<td>811 / 855</td>
<td>1808 / 1683</td>
<td>3558 / 3557</td>
</tr>
</tbody>
</table>

Analysis of the frontal region during the production of fricatives /s/, /sh/ and /shu/ is based on coronal sections of the vocal tract. Figure 2 shows a 3D shape (a) and area functions (b) of the vocal tract during the production of Japanese fricatives. The 3D shape shows a narrow oral cavity that is formed by the upper and lower teeth and the frontal tongue body during fricative production. The vocal tract area functions broadly show the expected patterns such as narrow oral cavity with constrictions at the tongue tip. Locations of minimum vocal tract constriction measured from the lips are 8.6, 34 and 43mm, and areas of minimum vocal tract constriction are 42, 48 and 38 mm² for fricatives /s/, /sh/ and /shu/, respectively.

Figure 2. 3D shape and area functions of the oral cavity in coronal sections from the lip to posterior wall of the pharynx during the production of Japanese fricatives /s,sh,shu/.
Analysis of the paranasal cavities is based on transverse sections from the maxillary sinus to the frontal sinus. Figure 3 shows a 3D shape of the paranasal cavities in transverse sections from inferior to superior sections for subject 1. Figure 4 shows an area function of the maxillary sinus represented by relationships between positions along the superior-inferior axis and area. Across sinuses, the area functions show the following patterns: large area for the sphenoidal sinus, wide maxillary sinus, and nonsymmetric shapes of the maxillary sinus with respect to the mid-sagittal line. The validity of the results is confirmed by comparison of the profiles of the vocal tract and nasal cavity with anatomical data.

![Figure 3. 3D shape of the paranasal cavities in transverse sections from superior to inferior sections](image)

![Figure 4. Area function of the maxillary sinus in transverse sections.](image)

**CONCLUSIONS**

This study investigated the feasibility of the simultaneous measurements of 3D shapes of the vocal tract, dental crown and nasal cavity using MRI. A method of simultaneously obtaining MR images of the dental crown and vocal tract using a dental crown plate was developed. The 3D shapes and area functions of the vocal tract during the production of fricatives were investigated.

Acknowledgment. We would like to thank Messrs. Takeshi Wada and Takashi Tachimura of the Division for Oral-Facial Disorders, Faculty of Dentistry, Osaka University for fabrication of the dental crown plates used in this study.

**REFERENCES**

KNOWLEDGE FROM SPEECH PRODUCTION USED IN SPEECH TECHNOLOGY: ARTICULATORY SYNTHESIS

Richard S. McGowan
Haskins Laboratories, 270 Crown Street, New Haven, CT 06511, USA

INTRODUCTION

There appears to be a continuing trend toward incorporating knowledge of speech production into speech technology—text-to-speech synthesis (e.g. Parthasarathy & Coker, 1992; Bickley et al., 1994), low-bit rate coding (see Schroeter & Sondhi, 1992), and automatic speech recognition (e.g. Shirai & Kobayashi, 1986; Rose et al., 1994). For automatic speech recognition using knowledge of the coordination of the vocal tract articulators and the resulting acoustics can reduce apparent token-to-token variability so that general pattern recognition algorithms have less work to do. Using articulatory representations in speech coding has the potential of greatly reducing bit rate because the articulators move relatively slowly and may be described by a few parameters by using an underlying dynamical model or by using simple curve fitting. Finally, text-to-speech synthesis can be improved using articulator control parameters, because the laws of physics can be used to produce the correct bundle of acoustic features with a comparatively limited parameterization—the acoustic output is constrained by the laws of physics. All these applications that depend on articulatory representation of speech production, can be grounded in what is called an articulatory synthesizer. An articulatory synthesizer is a device that produces speech output from a set of articulatory parameters (an articulatory representation). These devices are usually implemented in software on a digital computer.

The production of speech using an articulatory synthesizer (the "forward" mapping) can be divided into two major components: that of finding the mapping from the linguistic units to the articulatory movement, and that of finding the mapping from the articulatory movement to the aerodynamic state and acoustic output. The forward problem is solved with the composite mapping. The first mapping is the domain of people interested in the control of human movement and coordination as it relates to the vocal tract during speech, which includes some linguists and some experimental psychologists. The second mapping from articulatory movement to aerodynamic state and acoustics is the domain of acousticians. It is easily seen that both components of the forward mapping are important for the named technical applications. To perform text-to-speech synthesis from an articulatory point-of-view, the composite mapping is constructed, and to perform low bit rate coding or automatic speech recognition, one would need to find the inverse of the composite mapping, if the approach is to use articulatory information. If an analysis-by-synthesis procedure is used to construct the inverse composite mapping, then it is necessary to construct each component forward mapping.

TASK DYNAMICS

Only one example of one part of mapping from linguistic units to articulatory movement will be discussed here. This example is the model of articulatory coordination for articulators in performing speech gestures used at Haskins Laboratories, known as task dynamics (Saltzman & Munhall, 1989). This model describes the formation and breaking of constrictions in the vocal tract using a set of independent, linear, second-order differential equations: one equation for each constriction (e.g. labial, tongue body, or tongue tip). Because constrictions can be made with the coordinated activity of vocal tract articulators, such as lips, jaw and tongue, these equations are transformed into a model for articulator geometry. In the articulator coordinate system, the equations for constriction dynamics become coupled and nonlinear, and the pseudo-inverse of the Jacobian is used in their solution because there are more articulator than constriction degrees-of-freedom. This means that several articulators can be used to attain the same constriction target (upper lip, lower lip, or jaw can be used to attain lip closure).
Task dynamics models phenomena that are observed in real speech behavior. This includes articulatory compensation, where one articulator compensates for another that cannot move. (e.g. the lips can increase their total movement to close the mouth when the jaw cannot move.) The other pervasive phenomenon in speech that task dynamics models is that of coarticulation. For instance the jaw can be used to close the mouth or it can be used to lower the body of the tongue for certain vowels, such as /a/. When there is mouth closure, say for /b/, followed by an /a/ both goals influence the jaw, so that the mouth closure is probably attained by more lip involvement than would be without the presence of the /a/. This is so the jaw can be lower for the following /a/.

There are some real advantages to using task dynamics in the technical applications to be considered. The first is that constrictions of the vocal tract and the output acoustics are closely related so that the analysis-by-synthesis that recovers task dynamic parameters from speech is facilitated. Further, phonology based on articulatory gestures and instantiated in task dynamics is being constructed by linguists (Browman & Goldstein, 1990). This kind of work is necessary to finally map the task-dynamic parameters, such as the natural frequency of a lip closure, to linguistic units. This, of course, is required if automatic speech recognition is to be done using an articulatory representation.

Task dynamics takes the approach of finding the appropriate coordinate system to define speech behaviors (currently, constriction dynamics) and a means of transforming this coordinate system into a physical coordinate system (vocal tract articulators). This approach is extremely valuable in attempting the man-machine applications that are aimed above. However, there is room for an evolution in the details of this approach. For instance, the aerodynamics of the vocal tract appear to be controlled in a task specific way, and thus these must be included in some way (McGowan & Saltzman, in press). It is not clear whether all vocal tract gestures use constriction targets, and, in particular, vowels may need a more spatially global specification (Mattingly, 1990). Also, even where constriction targets are appropriate, there may be a region of targets rather than a point target (Guenther, 1994). The extension of this model should be undertaken to account for a variety of individual vocal tract shapes and for the sequencing of gestures, which is important for the rhythm mechanisms for rate and stress.

ARTICULATION-TO-ACOUSTICS

Where are we now in terms of the mapping from articulation to acoustics in articulatory synthesis, which is the second mapping that has to be constructed? What is the relation between the physics of fluid flow in the vocal tract and the propagation models that we are currently using? All the articulatory synthesizers known to the author use one-dimensional models of wave propagation (some with corrections for large area changes). The voice source is generated in a variety of ways, including simulations of self-oscillating vocal folds. The noise sources in the vocal tract are modeled as point sources, and their amplitude and frequency characteristics depend on aerodynamics in various degrees of sophistication.

Some synthesizers are time-domain synthesizers (e.g. Maeda, 1982), so that the waves created by the sources are propagated on a space-time grid. Other synthesizers use a frequency-domain transfer function to represent the wave propagation in the vocal tract (e.g. Sondhi & Schroeder, 1987; Davies et al., 1993). The output speech can be calculated by mapping the transfer function to the corresponding time-domain transfer function via an inverse discrete Fourier transform (DFT). An alternative is to find the poles and zeros of the transfer function and to use a formant synthesizer to produce the output speech (e.g. Lin, 1994). The remainder of the paper will suggest two research directions for articulation-to-acoustics mapping. The first is a proposal to use a set of orthonormal bases functions, other than circular functions, to represent the vocal tract transfer function. These bases functions are from what Coifman (1991, p 881) calls a "library of wavelet packets", including wavelets, used in multiresolution analyses. The other proposal is to provide a four-parameter, articulatory model for the control of the voice source.

MULTIRESOLUTION SYNTHESIS

When a frequency domain transfer function is used to represent vocal tract wave propagation, a time domain transfer function is calculated as an inverse discrete Fourier transform (DFT), and the sources convolved with the resulting transfer function. To obtain reasonable frequency resolution it is necessary that the transform window be of reasonable duration (25.6 ms for Sondhi & Schroeder [1987] for a 20kHz sampling rate). However, in performing the inverse transform, the vocal tract is assumed to be unchanging within the duration of the transform window; thus invoking the quasi-steady (stationarity) approximation. There are speech environments for which this approximation may be inappropriate, including the closure and release of stops, fricatives, affricates, and approximants. Specifically, there are two possible problems in these speech environments. First, the filtering properties of the vocal tract may be rapidly varying, and, second, the source properties may be changing rapidly because of vocal tract changes. The voice source is affected by the configuration of the upper vocal tract in what is known as source-tract interaction. Also, the aerodynamic noise source properties of amplitude and spectral content are directly affected by the vocal tract configuration because of changes in constriction areas and pressure distributions. Thus, the quasi-steady assumption is suspect in certain phonetic environments. In fact, this has been a problem in using DFTs for the analysis of speech.
A multiresolution decomposition may help in this regard (Meyer, 1993). In such a decomposition, the high-frequency components can be more localized in time than the low-frequency components. Thus, the high-frequency components can change more rapidly than the low-frequency components without violating stationarity. Recent work has been done in multiresolution decomposition and its generalizations for analysis and compression of speech signals (e.g. Wickerhauser, 1993). These decompositions make use of wavelets, wavelet packets, and other orthonormal bases to find decompositions suitable for a given application. In the case of data compression Shannon entropy can be minimized (Coifman & Wickerhauser, 1993). For purposes of speech analysis, the multiresolution decompositions allow the analyst to tile the time-frequency plane tailored to the physical situation. While there is still a trade between frequency and time resolution because of the Heisenberg uncertainty principle, the duration of the time window can be tailored to the analysis frequency. Thus, a spectrogram would consist of time slices depending, not only on the time coordinate, but also the frequency coordinate. In the particular case of a wavelet transform, one obtains an octave-band decomposition. However, there are more general orthonormal bases that allow a more irregular tiling of the time-frequency plane, with a lower bound set on the area of a tile by the Heisenberg uncertainty principle.

It is proposed here that a multiresolution form of decomposition be used for articulatory synthesis, as well as, analysis. This would involve decomposing the time-dependent part of the equations of motion for air in the vocal tract into a general orthonormal basis. The vocal tract could still be divided into small tube sections for the spatial discretization, if desired. The matrices describing the transformation of pressure and volume velocities from one section to another (Sondhi and Schroeter’s chain matrices) would be written in new orthonormal coordinates. In one possible implementation of a multiresolution synthesis area functions would be sampled at a fast rate, and the area function averaged over different intervals depending on the frequency scale of interest, with the higher frequencies requiring less duration for quasi-steady conditions than the low frequencies. The time derivatives, including fractional derivatives, would be written in terms of the chosen orthonormal basis. While Fourier analysis transforms the derivative operator to a diagonal operator, the wavelet decomposition transforms the derivative into a sparse matrix. This sparse matrix can be used for fast computation of derivatives in the wavelet basis (Beylkin, 1993). Further, any noise sources can be can be shaped by bandpass filters composed of wavelets.

FOUR PARAMETER VOICE SOURCE

Articulatory synthesizers can have voice sources that are not controlled using articulatory parameters (e.g. Rubin, et al. 1981). Or they have voice sources that are continuum mechanical models of the self-oscillating folds and air flow in the laryngeal region (Ishizaka & Flanagan, 1972). The latter simulations can require too much computation time or produce poor voice quality in running speech. While the former voice sources can produce natural sounding voice, they are not controlled by articulatory parameters.

The cover-body model of the vocal folds is the starting place for a four parameter model of the voice source (Hirano, 1974). It is supposed that all aspects of voice quality can be determined by "...the relationship between the body and cover of the vocal fold" (Hirano, 1974, p 91). The cover body picture of phonation has been expanded by others, most notably Titze (1994), who has constructed muscle activation plots (MAPs) for fundamental frequency control. (While these MAPs have largely been based on canine data, Titze's group has recently measured stress-strain relations for the human vocal ligament (Titze et al., 1994).) In these plots isofrequency contours are plotted against cricothyroid (CT) muscle activation and thyroarytenoid (TA) activation. However, for purposes of controlling the properties of the of the cover and body of the folds, the CT activation could be thought to represent any factor, intrinsic or extrinsic that controls the length of the cover and body, and stiffening both structures when they are lengthened. This change could be due to factors such as raising and lowering the larynx so that the larynx changes position along the spine, thus rotating the thyroid and cricoarytenoid relative to one another (Honda, 1995). Also, muscles who's primary effect is thought to be additive and aductory motion of the folds can have an effect the length of the cover-body in ways analogous to the CT. On the other hand, TA activity, reduces the length of the cover-body complex, but by stiffening the body and relaxing the cover. There are many ways of attaining the same fundamental frequency using different combinations of CT and TA activation. However, these different combinations can often, if not always, be distinguished in other acoustic dimensions, including source amplitude and spectral content.

There are other ways to control fundamental frequency, source amplitude and spectral content. These include the degree of adduction and the transglottal pressure. The latter parameter is partly determined by what is happening in the upper vocal tract independent of the larynx. A tight constriction in the upper vocal tract and an open glottis will mean that transglottal pressure decreases. Thus, the four parameters in the proposed model of voice source control are the transglottal pressure, degree of abduction, (generalized) CT activity, and TA activity. These parameters should provide enough degrees of freedom to produce just about any voice quality. This would not be true if one of these parameters were omitted, and so this set could be considered minimal. Also, while it is not a detailed anatomical model of the larynx and its vibratory modes, it is sufficiently articulatory given the state of the art in articulatory synthesis.

- 103 -
ACKNOWLEDGMENTS

This work was supported by grant NIH grant DC 01247 to Haskins Laboratories. The description of task dynamics has benefited by discussions the author has had with Elliot Saltzman. Thanks to Phil Rubin and Doug Whalen for reviewing this work.

REFERENCES


UNIVERSAL ADAPTATION METHOD BASED ON HMM COMPOSITION

Yasuhiro Minami, Sadaoki Furui
NTT Human Interface Lab., 3-9-11. Midori-cho, Musashino-shi, Tokyo, 180 Japan

SUMMARY
This paper proposes an adaptation method for universal noise (combination of additive noise and multiplicative distortion) based on the HMM composition (compensation) technique. Although the original HMM composition can be applied only to additive noise, our new method can also estimate multiplicative distortion by maximizing the likelihood value. The signal-to-noise ratio is automatically estimated as part of the estimation of multiplicative distortion. Phoneme recognition experiments show that this method improves recognition accuracy for noisy and distorted speech.

INTRODUCTION
Background noise, channel noise, and channel distortion are crucial problems in speech recognition. They are usually modeled by combining additive noise and multiplicative distortion in the linear spectral domain. If both additive noise and multiplicative noise can be simultaneously estimated, that is, if universal noise adaptation can be achieved, it should be very useful in speech recognition applications.

Various methods for removing estimated noise and distortion have been proposed, including spectral subtraction for additive noise and cepstral normalization for multiplicative distortion. However, since these methods utilize the average values of linear spectra and cepstra as noise and distortion, they cannot be simply extended to remove combinations of additive noise and multiplicative distortion.

An accurate noise adaptation method using a noise model, called HMM composition or compensation, has recently been proposed for additive noise [1][2]. This method creates HMMs for noisy conditions using speech HMMs and a noise HMM. It uses the mean and covariance of the noise distribution to adapt the speech distribution. However, it does not consider multiplicative distortion.

A different approach has been proposed to estimate either the multiplicative distortion or additive noise spectrum by maximizing the likelihood value. Rahim et al. removed telephone line bias in the cepstrum domain (multiplicative distortion) [3]. Rose et al. formulated the maximum likelihood parameter estimation procedure for additive noise or multiplicative distortion [4].

This paper extends the HMM composition method to accommodate both additive and multiplicative noise by using a maximum likelihood estimation criterion. In this framework, the S/N estimation is performed as part of the estimation of multiplicative distortion.

NOISY AND DISTORTED SPEECH MODELING
The model for producing speech signals under most noisy conditions is shown in Figure 1. Speech signal S is produced by speech HMMs and noise signal N is produced by a noise HMM. Both S and N are defined in the linear-power spectral domain. First, S is multiplied by multiplicative distortion G, which includes speaker characteristics. Then additive noise N is added to speech signal SG. Finally, the speech signal is multiplied by multiplicative noise H, which includes line and channel distortion. We thus obtain the final noisy and distorted speech signal as \( X = H(GS + N) = HGS + HN \). By setting \( W = HG \), we get \( X = WS + HN \); therefore, the basic noisy speech model can be converted into the model shown in Figure 2.

The HMM for HN can be trained by using the signal recorded for a period without speech. The HMMs for S can be made from noise-free speech. The problem is how to estimate W. Since W is multiplicative distortion, it can be written as

\[
W = \{ w_0, w_1, w_2, ..., w_p \}
\]

in the linear spectrum domain, where \( p+1 \) is the number of power spectral components.

FORMULATION OF ADAPTATION
To estimate the value of W, we model X by combining the HMMs for HN and WS using the HMM composition method. Here we assume that W is a fixed vector. Then W is estimated by maximizing the likelihood score \( P(O, \Lambda | W) \) or \( P(O, A | W) \), where \( P(O, \Lambda | W) \) is the joint likelihood score, \( P(O, A | W) \) is the Viterbi likelihood score, \( O = \{ x_1, x_2, ..., x_T \} \) is a time sequence of input vectors, \( M(W) \) is a set of phoneme models as functions of W, and \( \Lambda = \{ s_1, s_2, ..., s_T \} \) is the time sequence of states.

To maximize \( P(O, \Lambda | W) \) or \( P(O, A | W) \), we propose the following two methods [5].
(1) Exact method

The $P(O|AIM(W))$ can be maximized by using the steepest descent method, using the following iterative equation.

$$W(k) = W(k-1) + \epsilon \frac{\partial \log(P(O|AIM(W)))}{\partial W},$$

(2)

where $\epsilon$ is the step size.

If the output probabilities are represented by mixture Gaussian densities, it becomes complicated to maximize Eq. (2) directly. Therefore the maximum single Gaussian density in each mixture is used instead of multiple mixtures.

$$b_{s,t}(x_t) = \frac{1}{(2\pi)^{d/2} |\Sigma_{s,t}|^{1/2}} \exp \left( -\frac{1}{2} (x_t - \mu_{s,t})^T \Sigma_{s,t}^{-1} (x_t - \mu_{s,t}) \right),$$

(3)

Viterbi decoding is used to obtain $P(O|AIM(W))$, $\mu = (\mu_1, \mu_2, ..., \mu_n)$, and $\Sigma = (\Sigma_1, \Sigma_2, ..., \Sigma_n)$, where $\mu$ and $\Sigma$ are the sequences of mean vectors and covariance matrices selected by the Viterbi algorithm.

To deal with the many random variables in the equations, the following conventions are used. $R_c^e$ represents source $R$ in domain $c$, where $c=(cep, lg, lin)$. For instance, $X_{cep}$ is the random variable associated with noisy speech in the linear spectrum. The corresponding Gaussian distribution is $N(\mu_{cep}, \Sigma_{cep})$. The main notations for random variables are as follows:

- $HN$, $HN^*$: Noise in the cepstrum, the logarithm spectrum, and the linear spectrum domain.
- $S$, $S^*$: Speech in the cepstrum, the logarithm spectrum, and the linear spectrum domain.
- $X_{cep}$, $X_{lin}$: Noisy and distorted speech in the cepstrum, the logarithm spectrum, and the linear spectrum domain.

From the definition for HMM composition (compensation), the following equations are obtained.

$$\mu_{cep} = \Gamma \mu_{cep},$$

(4)

$$\Sigma_{cep} = \Gamma \Sigma_{cep} \Gamma^T,$$

(5)

$$\mu_{cep} = \exp(\mu_{cep}^2) 2 \mu_{cep},$$

(6)

$$\sigma_{cep} = \mu_{cep} \sigma_{cep} \exp(\sigma_{cep}^2),$$

(7)

$$\mu_{cep} = \log\left( \frac{w_{cep} \mu_{cep} + \mu_{cep}}{w_{cep} \mu_{cep} + \mu_{cep}} \right),$$

(8)

$$\sigma_{cep} = \log\left( \frac{w_{cep} \sigma_{cep} + \sigma_{cep}}{w_{cep} \sigma_{cep} + \sigma_{cep}} \right) \mu_{cep},$$

(9)

$$\mu_{cep} = \Gamma^{-1} \mu_{cep}.$$

(10)

Figure 1. A production model for noisy and distorted speech.

Figure 2. A converted production model for noisy and distorted speech ($W=HG$).
\[ \sum_{i} X_{ce}_{pi} = \Gamma^{-1} \sum_{i} X_{i} g \Gamma^{-1} \mathbf{T} \]  

(11)

Cosine transform, \( \mathbf{T} \): transpose, and \( u, v \): parameter indices, \( 0 < u, v < p \).

Ignoring the differential coefficients in Eq. (2) calculated from the covariance matrix (considering only the differential coefficients calculated from the mean vector), we obtain the following equation, with \( 0 \leq u, v \leq p \):

\[ w_{p}(i) = w_{p}^{-1}(i-1) + \sum_{i} \alpha \frac{\partial X_{ce}_{pi}^{(2)}}{\partial w_{p}} \alpha \frac{\partial X_{ce}_{pi}^{(2)}}{\partial w_{p}} - \frac{1}{2} \frac{\partial X_{ce}_{pi}^{(2)}}{\partial w_{p}} \alpha \frac{\partial X_{ce}_{pi}^{(2)}}{\partial w_{p}} + \frac{1}{2} \frac{\partial X_{ce}_{pi}^{(2)}}{\partial w_{p}} \alpha \frac{\partial X_{ce}_{pi}^{(2)}}{\partial w_{p}} + \frac{1}{2} \frac{\partial X_{ce}_{pi}^{(2)}}{\partial w_{p}} \alpha \frac{\partial X_{ce}_{pi}^{(2)}}{\partial w_{p}} \]  

(12)

By differentiating Eq. (8), we obtain

\[ \frac{\partial X_{ce}_{pi}}{\partial w_{p}} = \mu_{p} w_{p} + \mu_{p} w_{p} + \mu_{p} w_{p} + \mu_{p} w_{p} - \mu_{p} w_{p} - \mu_{p} w_{p} \]  

(13)

Finally, Eq. (12) can be calculated by using Eq. (13).

This framework can be extended to a noisy speech model, as shown in Figure 3. In this model, the S/N ratio in training the noise HMM and that in adapting the HMMs can be different. Since \( X \) can be written as \( X = H(g(i) + k, N) = H(g + k, N) = WS + k, HN \), Eqs. (8) and (9) can be modified as follows:

\[ \mu_{k} = \log \left( \frac{w_{k} X + k, N}{w_{k} X + k, N} \right) \]  

(14)

\[ \alpha_{k} = \log \left( \frac{w_{k} X + k, N + kH}{w_{k} X + k, N + kH} \right) \]  

(15)

In Eq. (14), since \( \log(k) \) is a bias in the spectrum domain and, hence, does not occur in pdf in the cepstrum domain, it can be ignored in the adaptation process. Equations (8) and (9) are obtained by setting \( W = [ w, w, w, w, ..., w ] \) in Eqs. (14) and (15). This means that the model in Figure 3 can be converted into the model shown in Figure 2.

(2) Parallel method

The parallel method (Figure 4) is mathematically simpler but computationally more costly than the exact method. In this method,
several sets of models having different $W$, $\varsigma$ are prepared. Using these models, the likelihood scores, $P\text{XIM}(W, \varsigma)$, are calculated for all $\varsigma$, and the set of models having the maximum likelihood is selected.

This method is especially useful when only the S/N ratio is estimated. Estimating the S/N ratio is a special case of estimating $W$ where $W = (k, k, \ldots, k)$. Various $k$, $\varsigma$ are prepared at several intervals in the S/N ratio to estimate the S/N ratio.

**EXPERIMENTS**

We tested our method in terms of the phoneme recognition rate. Noisy speech data were artificially created on a computer. Noise recorded in a computer room was added to clean speech data at 12 dB S/N ratio. The data were then passed through a distortion filter whose characteristic was set to 1.097. The input data were sampled at 12 kHz. The input feature vector consisted of 16 cepstra, 16 delta cepstra, and 1 delta power.

The speaker-independent HMMs were trained with speech data recorded from 64 speakers under noise-free conditions. One sentence with the transcription was used for adaptation. The training sentence was the first sentence in the phoneme-balanced sentence set. The S/N ratio was first estimated using the parallel method. This value was then used as the initial value for the next estimation. Then, $W$ was estimated using the exact method. Using these two steps speeds up the convergence of the likelihood value.

In the exact method, Viterbi decoding is performed first to obtain $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ and $\Sigma = (\Sigma_1, \Sigma_2, \ldots, \Sigma_k)$. Then a new $W$ is calculated. These steps are repeated until the likelihood value converges, at which point the HMMs for noisy and distorted speech are obtained.

Using these HMMs, we performed phoneme recognition experiments. Fifty-one evaluation sentences were uttered by one male speaker. Since our database has only phoneme descriptions and no precise phoneme labeling, we used an evaluation algorithm for the phoneme recognition rate that does not need precise phoneme labeling [6].

**RESULTS**

Table 1 shows the results. In the "Clean HMM" experiment, the phoneme HMMs without HMM composition were used. In the "HMM composition with ideal S/N" experiment, the highest recognition rate was selected, varying the S/N ratio every 3 dB in HMM composition. In the "HMM composition with estimated S/N" experiment, the S/N ratio estimated by the parallel method was used in HMM composition. In "HMM composition with estimated $W$", the value of $W$ estimated by the exact method was used in HMM composition.

The recognition rates increased from 44.7% to 56.5% when the estimated S/N ratio was used and the recognition rates in "HMM composition with ideal S/N" and "HMM composition with estimated S/N" were almost the same. This indicates that the parallel model can estimate the S/N ratio. The recognition rate greatly increased from 56.5% to 67.7% when the estimated $W$ was used. This means that our method can effectively estimate the filter characteristic.

<table>
<thead>
<tr>
<th>Clean HMM</th>
<th>HMM composition with ideal S/N</th>
<th>HMM composition with estimated S/N</th>
<th>HMM composition with estimated $W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>44.7%</td>
<td>56.7%</td>
<td>56.5%</td>
<td>67.7%</td>
</tr>
</tbody>
</table>

**CONCLUSION**

Our proposed noise adaptation method estimates both additive noise and multiplicative distortion in a single framework by maximizing the likelihood value of a training sentence based on the HMM composition technique. Phoneme-recognition experiments confirmed that this method greatly improves the recognition rate for noisy and distorted speech.

**ACKNOWLEDGMENTS**

We thank the members of the Furui Research Laboratory of the NTT Human Interface Laboratories for their valuable discussions. We used the Acoustic Society of Japan's speaker-independent continuous-speech database to train the speaker-independent HMMs.

**REFERENCES**


15th International Congress on Acoustics
Trondheim, Norway 26 - 30 June 1995

APPROACHES TO ENVIRONMENT COMPENSATION IN AUTOMATIC SPEECH RECOGNITION

Pedro J. Moreno, Bhiksha Raj, Richard M. Stern
Department of Electrical and Computer Engineering and School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 152213, USA

ABSTRACT

This paper describes a series of cepstral-based compensation procedures that render the SPHINX-II continuous speech recognition system more robust with respect to acoustical changes in the environment. The first two algorithms, SNR-based MultiVariate Gaussian-based cepstrum normalization (SNR-based RATZ) and STAsitical Reestimation of HMMs (STAR), compensate for environmental degradation based on comparisons of simultaneously-recorded data in the training and testing environments ("stereo data"). They differ in that RATZ modifies the incoming feature vectors to a recognition system while STAR modifies the internal representation of speech by the system. We also describe N-CDCN, an improved version of codeword-dependent cepstral normalization (CDCN) which does not require stereo training data but nevertheless achieves performance levels comparable to RATZ and other algorithms that require stereo training. Use of these compensation algorithms significantly reduces the error rates for SPHINX-II. The algorithms are tested in a variety of databases and environmental conditions.

INTRODUCTION

Robustness with respect to environmental variability remains a continuing problem for speech recognition technology (e.g. [3]). For example, the use of microphones other than the ARPA standard Sennheiser HM-414 headset (CLSTLK) severely degrades the performance of speech recognition systems like the SPHINX-II, even in relatively quiet environments [1, 4].

Traditional algorithms to compensate for environmental variation have either relied on the availability of simultaneously-recorded data in the training and testing environments ("stereo data"), or have utilized structural models to define the degradation. For example, multiple fixed codeword-dependent cepstral normalization (MFCDCN) [4] uses stereo data to compute correction vectors to compensate for the effects of the environment. Dual-channel codebook adaptation (DCCA) [4], on the other hand, modifies the statistical representation used in the HMMs for speech on the basis of comparisons obtained from stereo data. The other approach to environmental compensation is through the use of structural models of degradation. On the other hand, codeword-dependent cepstral normalization (CDCN) [1] assumes that speech is degraded by unknown additive noise and unknown linear filtering. It makes use of expectation-maximization (EM) techniques to determine the parameters characterizing these distortions.

In this paper we describe three new cepstral-domain compensation strategies, SNR-based MultiVariate Gaussian-based cepstrum normalization (SNR-based RATZ), STAsitical Reestimation of HMMs (STAR), and new CDCN (N-CDCN). SNR-based RATZ and STAR both make use of stereo data. They differ in that RATZ modifies the incoming feature vectors to a recognition system while STAR modifies the internal representation of speech by the system. RATZ and STAR are similar in philosophy to MFCDCN and DCCA, respectively [4], but they achieve improved performance through the use of better mathematical models which introduce strong structural constraints into the assumed distribution for speech. N-CDCN is a modification and improvement of the original CDCN algorithm, which is based on a structural model of degradation.

EFFECT OF THE ENVIRONMENT ON SPEECH STATISTICS

In this section we describe how even well-behaved environments, such as those that can be modeled by unknown linear filtering and additive stationary noise, modify the statistics of "clean" speech in very unpredictable ways. Even though we can formulate equations that analytically describe how the pdfs of clean speech change, the solutions for these equations are mathematically intractable.

For analytical purposes, we adopt the simple model of degradation proposed by Acero [1]. In this model, degraded speech is characterized by passing high-quality clean speech through a linear filter and contaminating the filtered output by additive stationary noise. For simplicity, we will also assume that the feature vector is unidimensional,
although all conclusions developed can be easily extended to an arbitrary $N$-dimensional space such as the log spectral domain. The degraded speech can be characterized as:

$$Z(\omega) = X(\omega)H(\omega) + N(\omega)$$

where $Z(\omega)$ represents the power spectrum of the degraded speech, $X(\omega)$ is the power spectrum of the clean speech, $H(\omega)$ is the transfer function of the linear filter, and $N(\omega)$ is the power spectrum of the additive noise. In the log-spectral domain this relation can be expressed as:

$$z = x + q + \log (1 + e^{z - q}) \quad r(x, n, q) = \log (1 + e^{z - q})$$

where $z, x, q,$ and $n$ represent the logs of $Z(\omega)$, $X(\omega)$, $|H(\omega)|^2$, and $N(\omega)$, respectively, for some particular $\omega$.

Assuming knowledge of the pdf of the clean speech, $p(x)$, with mean $\mu_x$ and variance $\sigma_x^2$, degradation will affect the mean and variance of $z$ in the following manner:

$$\mu_z = E[z] = \mu_x + q + \mu_{r(x,n,q)} \quad \sigma_z^2 = E[(z - \mu_z)^2] = \sigma_x^2 + \sigma_{r(x,n,q)}^2 + 2E[r(x,n,q)] - \mu_x \mu_{r(x,n,q)}$$

For simplicity we assume that $x$ is Gaussian and that the power spectrum of the noise and the transfer function of the filter are known and deterministic. In this simplified special case the new equations for the mean and variance are:

$$\mu_z = \mu_x + q + \frac{1}{x} \int N(\mu_x,\sigma_x) \cdot \sigma_{r(x,n,q)} dx \quad \sigma_z^2 = \int N(\mu_x,\sigma_x) \cdot (x + q + \sigma_{r(x,n,q)})^2 dx - \mu_x^2$$

These equations become difficult to solve. In fact, we are not aware of any analytical solutions for these equations. Equations (4) were obtained under the unrealistic assumption that $N(\omega)$ is known a priori and deterministic. In practice, $N(\omega)$ must be estimated, producing a random estimate for $n$ to which we assign the pdf $p(n)$. Assuming that $n$ and $x$ are statistically independent, the new expression for $\mu_z$ becomes:

$$\mu_z = \mu_x + q + \frac{1}{x} \int N(\mu_x,\sigma_x) \cdot \sigma_{r(x,n,q)} dx \cdot \int N(\mu_n,\sigma_n) dx$$

This equation is more difficult to solve than equations (4). The computation for $\sigma_z$ becomes even more complicated.

We conclude that when we assume that the corrupted distributions have a Normal shape, the effects of the environment on signal statistics can be modeled by additive correction terms to the mean of $z$ (thus shifting its pdf), and the variance of $z$ (thus compressing its pdf). This is the approach that is followed in the SNR-based RATZ and STAR algorithms described in this paper.

**COMPENSATION ALGORITHMS**

In this section we briefly describe the three new algorithms RATZ, N-CDCN and STAR. SNR-based RATZ and STAR assume the availability of a database of "stereo" training sentences, with one channel containing speech recorded using a high-quality microphone and a second channel containing degraded speech samples.

**SNR-based RATZ**

SNR-based RATZ uses frame energy information, represented by the zeroth component of the cepstral vector $(x_0)$, to model the statistics of the cepstra of the clean speech hierarchically: the $x_0$ cepstral coefficients are defined to have a Gaussian mixture distribution, and the cepstral vectors corresponding to each $x_0$ Gaussian are further assumed to have a mixture Gaussian distribution:

$$x = [x_0 \ x_1 \ \ldots \ x_{p-1} \ x_p]^T \quad p(z) = \sum_{k=0}^{N-1} \ P[k] \ N(z_0, \mu_{x_0}, \sigma_{x_0}) \cdot \sum_{j=0}^{N-1} \ P[j \| k] \ N(z_p, \mu_{x_p}, \sigma_{x_p})$$

where $P(k), \mu_{x_0}$, and $\sigma_{x_0}$ represents the a priori probability, mean, and variance of each energy component, and $P(j \| k), \mu_{x_p}$ and $\sigma_{x_p}$ represent the a priori probability, mean vector and covariance matrix of each multivariate Gaussian mixture conditioned on the frame energy. These parameters are learned through traditional EM methods.

The effects of the environment on the statistics of clean speech are modeled as shifts in the means and variances:

$$\mu_{x_0} = \mu_{x_0} + R_{x_0} \quad (\sigma_{x_0})^2 = R_{x_0} \quad (\sigma_{x_0})^2 = R_{x_0} + R_{x_0} \quad \mu_{x_p} = \mu_{x_p} + R_{x_p} \quad (\sigma_{x_p})^2 = R_{x_p} + R_{x_p}$$

resulting in a new set of statistics describing the degraded speech vector $z$.

The shift parameters are learned using a traditional maximum likelihood approach that attempts to maximize the probability that the observed noisy data set is generated by the transformed statistics. When stereo data are available the reestimation formulas use it by modelling the a posteriori probabilities using the stereo clean vectors.
The degraded speech is compensated by using an MMSE technique to shift them back to the clean speech statistics:
\[ r_{jk} = [r_{0,k} + r_{k,j}]^T \]

\[ \hat{x} = E(x|z) = \int_x p(x|z) dx = x - \int_x r(x)p(x|z) dx = x - \sum_j \sum_k p(j,k|z)r_{jk} \]  

(8)

**STAR**

STAR assumes that any effect of noise, channel, or environment can be accurately formulated as shifts in the means and corrections to the variances of HMMs:

\[ \mu_{x,k} = \mu_{x,k} + r_k \quad \Sigma_{x,k} = \Sigma_{x,k} + \hat{R}_k \]  

(9)

This is approach is similar to that in [2]. This model is applied to all of the four streams of data that SPHINX-II uses, that is, cepstral, delta-cepstral, double delta-cepstral coefficients, and a fourth three-dimensional stream that contains the cepstral component \( C_0 \), its difference \( \Delta C_0 \), and its double difference \( \Delta^2 C_0 \).

Correction factors for the statistics of each of these four streams are computed using the following modified Baum-Welch estimation formulas:

\[ \hat{r}_{x,k} = \frac{\sum_{i=0}^{T-1} \gamma_i(k) (x_i - \mu_{x,k})}{\sum_{i=0}^{T-1} \gamma_i(k)} \]  

(10)

where \( \gamma_i(k) \) the probability of being in state \( k \) at time \( i \) is conditioned on statistics \( \lambda \) of the clean speech channel of the stereo training data. We assume that the a priori probabilities do not change due to the effects of noise or environment. We note that \( \gamma_i(k) \) is computed using clean speech and clean models. As a result, it does not change from estimation to estimation, and no iteration is needed.

The estimation formulas for corrections to the covariance matrices are similar:

\[ \hat{R}_k = \left( \sum_{i=0}^{T-1} \gamma_i(k) (x_i - \mu_{x,k} - \hat{r}_k)^T (x_i - \mu_{x,k} - \hat{r}_k) \right) \left( \sum_{i=0}^{T-1} \gamma_i(k) \right)^{-1} - \Sigma_{x,k} \]  

(11)

**N-CDCN**

New Codebook-Dependent Cepstral Normalization (N-CDCN) is an improved version of its predecessor, CDCN. The original CDCN has been able to achieve a respectable amount of error reduction [1] for many types of acoustical degradations without requiring the use of stereo training data. Nevertheless, it has not found much use in current CMU speech systems because it requires empirical information and retraining of the HMMs. N-CDCN alleviates these problems while retaining CDCN's ability to compensate for the combined effects of additive noise and distortion produced by unknown linear filtering. As before, compensation is performed on a sentence-by-sentence basis.

N-CDCN assumes a structural model of the degradation where speech is contaminated by additive stationary noise after being filtered by an unknown linear filter. This can be expressed as:

\[ \epsilon[k] = x[k] + h[k] + n[k] \]

A statistical description of the cepstral space characterizing "clean" speech, as parametrized by a mixture of multivariate Gaussians, is estimated by EM methods. Given a noisy utterance, parametrized by a sequence of cepstral vectors, and given the previously learned statistics describing the clean speech, N-CDCN iteratively estimates noise and linear filtering vectors that maximize the likelihood of observing the degraded speech. The initial values of the filter and noise estimates are the values that maximize the likelihood in the case of no estimation error.

Finally, an MMSE technique is used to estimate the unobserved clean speech vectors given the observed degraded speech, the previously-estimated noise and filter vectors, and the statistics describing clean speech.

**PERFORMANCE ANALYSIS**

In this section we describe the results of a series of experiments that compare the recognition accuracy of the algorithms described in Sec. with previous algorithms developed at CMU. The experiments measure recognition accuracy obtained using speech from the CENSUS database [1] that was recorded using the omnidirectional desktop Crown PZM GFS microphone. The CENSUS database consists of strings of letters and digits stereo that were simultaneously recorded using the close-talking Sennheiser HMD-414 microphone and the Crown PZM6FS desktop microphone.

Table 1 shows the results of these experiments. The system was trained on clean speech from the close-talking Sennheiser HMD-414 microphone and tested using noisy speech from the desktop Crown PZM-6FS microphone. The word error rate of the SPHINX-II system when training and testing on clean speech was 14.7%. It can be seen that SNR-based RAITZ and STAR provide slightly better performance than FCDCN. The performance of N-CDCN is bet-
ter than that of its predecessor, CDCN, and it has the additional advantage of not requiring that the system be retrained.

<table>
<thead>
<tr>
<th>Compensation Method</th>
<th>Stereo Data Needed?</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMN</td>
<td>No</td>
<td>29.5%</td>
</tr>
<tr>
<td>FCDCN</td>
<td>Yes</td>
<td>21.9%</td>
</tr>
<tr>
<td>STAR</td>
<td>Yes</td>
<td>21.5%</td>
</tr>
<tr>
<td>SNR-based RATZ</td>
<td>Yes</td>
<td>21.0%</td>
</tr>
<tr>
<td>CDCN</td>
<td>No</td>
<td>24.3%</td>
</tr>
<tr>
<td>N-CDCN</td>
<td>No</td>
<td>20.7%</td>
</tr>
</tbody>
</table>

Table 1: Word error rates obtained using speech from CENSUS database using different compensation algorithms in conjunction with Sphinx-II.

In a separate set of experiments we evaluated the performance of the various compensation algorithms in low signal-to-noise ratio (SNR) conditions. Clean speech from the CENSUS database was contaminated with artificially-produced additive Gaussian noise at different global SNRs. Figure 2 compares recognition accuracy obtained using the census database for the STAR, SNR-based RATZ, and N-CDCN algorithms, as a function of global SNR. For comparison purposes, we also provide the recognition accuracy obtained when the system was completely retrained at each SNR using the speech with added noise, as well as the baseline accuracy obtained using cepstral mean normalization (CMN) alone. In contrast, the experimental results described in Table 1 are based on data recorded at an SNR of 23.0 dB, which is indicated by the vertical dashed line in Figure 2.

As can be seen in Figure 2, the STAR algorithm outperforms all other algorithms, and the difference becomes especially evident at low SNRs. Algorithms that attempt to correct the noisy cepstra using MMSE techniques introduce additional classification errors over the optimal classifier based on noisy cepstra. Hence, a technique like STAR, that attempts to "classify" the noisy speech based on noisy cepstral statistics can outperform techniques that attempt to clean noisy data using MMSE methods, and classify the data using clean statistics.

**SUMMARY AND CONCLUSIONS**

In this paper we described three new procedures, SNR-based RATZ, STAR, and N-CDCN, that improve speech recognition accuracy in unknown acoustical environments. SNR-based RATZ and STAR are based on the availability of stereo training data, while N-CDCN performs blind compensation.

We also presented a brief analytical study with simulations to support our modelling of how noise and filtering can affect clean speech statistics. We showed how the addition of a hierarchical structure to data driven methods like SNR-based RATS improves performance compared to previous methods developed at CMU, especially at low SNRs. Finally, we also show how methods that attempt to modify the statistics of the HMMs can perform even better, especially at low SNRs.

**ACKNOWLEDGMENTS**

The authors thank Uday Jain, Matthew Siegler and Evandro Gouvea for useful suggestions. Pedro J. Moreno has been supported by a Fulbright fellowship awarded by the Ministerio de Educacion y Ciencia, Spain. This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

**REFERENCES**

A NEW REAL TIME HELIUM SPEECH UNSCRAMBLER USING SINGLE DSP

Takayoshi Nakai, Naoki Imagawa, Hisayoshi Suzuki
Department of Electric and Electronics, Faculty of Engineering, Shizuoka University,
3-5-1 Johoku, Hamamatsu-shi, 432, Japan

SUMMARY
This paper describes a newly designed helium speech unscrambler based on digital signal processing of line spectrum pair (LSP) parameters. The unscrambler can process the whole operations in real time by single DSP chip with small number of 36 subsidial ICs. In order to cope with the change in frequency bandwidth of helium speech spoken by divers at the different depths from 30 meters to 300 meters, analysis sampling frequency are changed by three steps. It is 20kHZ for 30m-90m, 22.5kHz for 90m-180m, and 25kHz for 180m-300m. Delay time between the input signal and the output processed signal is approximately 30ms.

INTRODUCTION
When a diver works at more than tens meters in depth, he breathes hyperbaric helium-oxygen gas mixture. Speech spoken in this environment is amazingly distorted and the divers confront to serious difficulty on communication with another diver as well as the people on a ship. Several researchers have been engaged in developing the helium speech unscrambling[1] - [8].

Authors have developed and evaluated a helium speech unscrambler based on analysis-synthesis system including LSP parameter conversion[6]-[8]. This instrument was composed of 210 digital IC elements including five DSPs (Fujitsu MB8764) and worked in real time pace. This instrument showed sufficient performance as far as it was used on a ship or a manned diving simulator that supports a diver working at depth of 180m to 300m range. It was needed, hopefully, to develop a carryable system for each diver. The new system is designed for this goal, and to be used at shallower sea diving for non-professional people who dive to engage in a research in the sea. In following sections, we describe implementation of the system using single high-performance DSP chip for unscrambling the helium speech spoken at 30 meters to 300 meters in depth.

UNSCRAMBLLING HELIUM SPEECH
The main difference between helium speech and normal speech is in the frequency characteristics of the vocal tract. Relation of formant frequencies between helium speech and normal speech is represented by Fant and Lindqvist equation as follows[9].

\[ F_a = \sqrt{(F_h/k)^2 - (r - 1)F_{wa}^2} \]  

where  
- \( F_h \): formant frequency of helium speech,  
- \( F_a \): formant frequency of normal speech spoken in the same articulatory configuration of the vocal tract in helium-oxygen environment,  
- \( F_{wa} \): the lowest resonant frequency of the vocal tract,
The main process for unscrambling helium speech is to convert formant frequencies $F_k$ of input speech to $F_k$ of output speech according to the equation (1). Since the changes in pitch frequencies of helium speech and normal speech are at most 30%, it is not necessary to convert pitch frequency. We proposed previously[6]-[7] the unscrambling method that satisfies the specifications mentioned above.

Figure 1 shows the block diagram of our unscrambling system. The system consists of five sections: pre-processing, PARCOR analysis, inverse filtering & down sampling(IFDS), LSP conversion, and LSP synthesis sections.

First, at the pre-processing section, the input speech data sampled at $F_s$ kHz are pre-emphasized and windowed. Next, at the analysis section, PARCOR coefficients are extracted from the processed data using linear predictive method. At the LSP conversion section, the PARCOR coefficients are converted to LSP parameters, and they are converted to those in normal air. On the other hand, at the IFDS section, sound sources used by LSP synthesis are extracted by inverse filtering with PARCOR and by down-sampling from $F_s$ kHz to 10 kHz. At the LSP synthesis section, unscrambled signals are synthesized from the converted LSP coefficients and down-sampled sound sources.

Ratio of sound velocity in helium-oxygen to that in normal air, $k$, extends from 1.72 at 30 meter depth to 2.64 at 300 meter depth and the ratio of density of helium-oxygen to that of normal air, $r$, extends from 1.50 at 30 meter depth to 5.24 at 300 meter depth. Since the frequency bandwidth of helium speech is almost $k$ times wider than that of normal speech, $k \times 10$ kHz is adequate for the analysis sampling frequency provided the synthesis sampling frequency 10kHz. In order to reduce the number of calculations of down-sampling at the IFDS section, however, we selected that $F_s$ is 20kHz for 30m-90m depth, 22.5kHz for 90m-180m depth, and 25kHz for 180m-300m depth.

FIGURE 1: Block diagram of unscrambling system.

HARDWARE CONFIGURATION

We used a ADSP-2101 DSP(Analog Devices Co.) as a main processor unit. It executes a 16bit fixed point calculation in high speed by multi-instructions at the same time, and is easy to input/output via two serial ports by internal/external interrupt.

In fixed point calculation, round off or truncation error occurs in multiplication or division. To cope with this problem, the maximum absolute value in each frame is calculated at the pre-processing section, and then, data in each frame are so normalized as to have always almost the same maximum value.

Fig. 2 shows the block diagram of the hardware system. Since the DSP starts the processing after storing the data of three frames, there are delay time between the input signal and the output processed signal of 30ms at the analysis sampling frequency 20kHz and 25kHz, and of 33ms at the sampling frequency 22.5kHz. In this system, every process in all sections including I/O stage is
executed by single DSP chip. All program data at three analysis sampling frequencies are stored in the boot memory EPROM and the program data of the analysis sampling frequency corresponding to the diving depth are booted in the internal main memory of the DSP when the DSP starts. The system includes 2K words RAMs of the external data memory, 1K words RAMs of the external program data memory, and 8K words ROMs of the external data memory for tables of the LSP conversion. There are 36 subsidial ICs in the system. The power dissipation is 12 watts.

![Diagram](image)

**Figure 2: Hardware of the system.**

Table 1 shows the number of calculations per one frame and program size at the analysis sampling frequencies 20kHz, 22.5kHz and 25kHz. At all the analysis sampling frequencies, program sizes are less than 2K words of the internal program memory and all the processes in each frame can be finished within one frame time.

**PERFORMANCE**

In order to evaluate performance of the system, we tested the five short sentences spoken by two divers on three conditions: in normal air, at 180 meter depth in this case he cannot listen to the unscrambled speech (NOHP), and at 180 meter depth while he listen to their unscrambled speech through headphones as well as their own helium speech (UNHP). We observed pitch frequency and formant trajectory. The results are summarized as follows:

1. The unscrambled speech at up to 300 meter depth is almost perfectly understandable: the intelligibility scores of words are 95% at 200m depth and 83% at 300m depth, respectively.

2. Divers can speak with a fairly distinct articulation: Fig. 3 shows 1st and 2nd formants trajectory when diver A spoke a Japanese short sentence /niowa amai (desuka?)/, which means

<table>
<thead>
<tr>
<th>Table 1: Software specifications of the system</th>
<th>m&amp;a: multiplication and addition, add: addition, div: division</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>sampling frequency</strong></td>
<td><strong>20kHz</strong></td>
</tr>
<tr>
<td>number of calculation per 1 frame</td>
<td>m&amp;a 20744</td>
</tr>
<tr>
<td></td>
<td>add 5700</td>
</tr>
<tr>
<td></td>
<td>div 76</td>
</tr>
<tr>
<td>program size</td>
<td>1178 words</td>
</tr>
<tr>
<td>MIPS</td>
<td>10.2</td>
</tr>
</tbody>
</table>

-115-
"(Is) salt taste sweet?". The trajectory of (c) in condition of UNHP moves more steady and smoothly than that of (b) in condition of NOHP, but does not move so smoothly and dynamically as that of (a) in normal air.

3. Pitch frequency becomes closer to normal condition: for diver A, his average pitch frequency on condition of UNHP is 2% lower than that on condition of NOHP, and is 3% higher than that in normal air. For diver B, it is 18% lower than that on condition of NOHP, but 18% higher than that in normal air.

CONCLUSIONS

Implementation of the new real-time helium speech unscrambler is described. The system consists of 37 ICs including single DSP and is possible to be redesigned as compact as a carryable size system. It is found that, when a diver speaks while he hears his unscrambled voice simultaneously, his pitch frequency is unchanged from normal speech and he can speak with a fairly distinct articulation.

ACKNOWLEDGEMENTS

The authors would like to thank the staff of the Japan Marine Science and Technology Center for cooperating with us for recording helium speech and Mr. Keizo Ishida, the technician of our laboratory, for managing our computers.

REFERENCES


Figure 3: 1st and 2nd formants trajectory of /siowa amai/, (a) spoken in normal air, (b) spoken at 180m depth, (c) spoken at 180m depth wearing headphones. Symbols A → G, are correspondent to /i/, /o/, /a/, /æ/, /m/, /æ/, and /i/, respectively.
DISTINGUISHING APICAL AND LAMINAL STOPS: A STUDY OF RELEVANT ARTICULATORY AND ACOUSTIC PROPERTIES

Paroo Nihalani
National University of Singapore

INTRODUCTION

In the Chomsky & Halle (1968) feature system, dental and alveolar consonants share the feature [Coronal]. Ladefoged (1968) observes that if a language has a contrast between dental and alveolar stops, then the principal articulatory difference between these two sounds is in the part of the tongue that is used. Dental stops are typically laminal being articulated with the blade of the tongue against both the teeth and the front part of the alveolar ridge, whereas alveolar stops are produced with the tip of the tongue against the alveolar ridge. Catford (1977) observes that an apical versus laminal articulation will have acoustic effects within the same place of articulation category. In particular, the size of the sublingual cavity will vary with the position of the tongue blade, and this in turn will affect the resonance frequencies of obstructive noise. It is therefore hypothesized that these two articulatory strategies (an apical versus laminal articulation) will have audibly different acoustic effects.

Jongman et al. (1985:248) claim that the contrast between dental and alveolar consonants is a marked one linguistically. First the contrast seems to occur relatively infrequently in natural languages and is rather limited to the African (Katcha and Kadugli), Dravidian (Malayalam) and Australian language families (cf. Dixon, 1980; Maddieson, 1980, 1981). Second when the contrast does occur, its distributional properties are very restricted. For example, the Malayalam contrast between dental and alveolar stops occurs only in intervocalic position in voiceless geminates. Moreover, according to the UPSID data (Maddieson, 1980, 1981), a four-way contrast between voiced and voiceless dental and alveolar stops is extremely rare, it is attested in only three Australian languages (Temein) and two East African languages (Katcha and Kadugli), and the data for these languages is extremely sketchy.

Sindhi is a member of the Indo-Aryan family of languages and is spoken in India and Pakistan roughly by about 14 million speakers. Sindhi has the fullest stop system of all Indo-Aryan languages and is reported to use dental and alveolar stop consonants contrastively. Nihalani (1974) has undertaken an extensive investigation of the physiological features of stop consonants in Sindhi and his instrumental records show clear evidence in support of the apical versus laminal distinction in Sindhi. These contrast syllable-initially (i.e. in the word-initial and word-medial positions). Thus there are no special restrictions about their distribution.

The goal of the present study is to explore the question of what exactly is meant articulatorily by 'apical' and 'laminal' articulations in the following two Indian languages; Malayalam (which belongs to Dravidian family of languages) and Sindhi (which is a member of the Indo-Aryan family of languages); and then to examine the corresponding acoustic differences. Another point to be determined is whether these differing articulatory strategies are consistent within a language or vary only according to speaker-specific idiosyncracies.
INSTRUMENTATION

Articulatory Measurements

Palatograms and linguagrams were made of 6 Malayalam and 2 Sindhi speakers uttering phrases containing the relevant segments. While palatogram and linguagrams were being made, the words were also recorded on a Marantz PMD 360 cassette recorder with a Sennheiser MD 419 microphone in order to be certain that each given acoustic signal corresponded to an articulation with known characteristics.

The use of palatography has been described by Ladefoged (1957). It is a technique used to investigate contact between the tongue and the roof of the mouth in the production of speech sounds. In the current experiments, the procedure described by Dart (1991) was used; palatograms and linguagrams were made using a liquid medium consisting of one part charcoal to two parts olive oil which was painted on either the upper or lower articulator (palate or tongue) with a small brush. After the utterance of the test word, the area where the medium wiped off on the opposing articulator was photographed with a Polaroid Land camera modified for dental use. This solution was easily rinsed away with a little water between each photograph. Dart (1991) points out several advantages of this method over the alternative dry spray method. In the first place, the area contact is much more clearly delineated, leaving fewer ambiguities for interpretation. Besides, it is undoubtedly more pleasant for the speaker since the solution is directly painted on to the relevant area only during which time the speaker may breathe normally, thereby avoiding the inhalation of the powder and irritation of the nasal cavities which is often a problem with the dry method. More importantly, there is no problem with the creation of excess saliva as has been generally reported when sweetened chocolate powder is sprayed directly into the mouth.

In order to make measurements of the constriction length, an impression was made of each speaker's palate to give a permanent record of the shape of the speaker's palate and dentition. Since this parameter is crucial to the definition of the feature [distributed], all tokens were measured for constriction length. Measurements were made by mapping the place of constriction shown in the palatograms on the casts of the palate and measuring the distance from front to back at the central area in line with the central incisors.

Acoustic Measurements

Using a Kay DSP Sonagraph Workstation, wide band spectrograms were made and the following measurements were determined from the spectrograms of the test utterances:

(a) closure duration of stop consonant

(b) voice onset time (VOT)

(c) formant frequencies (noted from wide band spectrograms) for all tokens both immediately before and after the consonant closure

(d) FFT power spectra were made of the release transients of stop consonants. The spectra were made on a Kay DSP Sonograph Workstation using a sampling rate of 20,480 samples/sec, and a bandwidth of 300 Hz. A full Hamming window of varying length was put over the transient, and the peak magnitude average was calculated.
### SUMMARY OF ARTICULATIONS USED BY MALAYALAM SPEAKERS

<table>
<thead>
<tr>
<th>Articulator</th>
<th>Laminal</th>
<th>Apical</th>
<th>Sublaminal</th>
<th>Postalveolar</th>
<th>Alveolar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. PA</td>
<td>laminal</td>
<td>apical</td>
<td>sublaminal</td>
<td>postalveolar</td>
<td>alveolar</td>
</tr>
<tr>
<td></td>
<td>den-alv</td>
<td>alveolar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. GK</td>
<td>laminal</td>
<td>apical</td>
<td>sublaminal</td>
<td>postalveolar</td>
<td>alveolar</td>
</tr>
<tr>
<td></td>
<td>den-alv</td>
<td>alveolar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. SK</td>
<td>laminal</td>
<td>apical</td>
<td>apical</td>
<td>postalveolar</td>
<td>alveolar</td>
</tr>
<tr>
<td></td>
<td>den-alv</td>
<td>alveolar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. GA</td>
<td>laminal</td>
<td>laminal</td>
<td>sublaminal</td>
<td>postalveolar</td>
<td>alveolar</td>
</tr>
<tr>
<td></td>
<td>den-alv</td>
<td>alveolar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. DN</td>
<td>laminal</td>
<td>apical</td>
<td>sublaminal</td>
<td>postalveolar</td>
<td>alveolar</td>
</tr>
<tr>
<td></td>
<td>den-alv</td>
<td>alveolar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. PV</td>
<td>laminal</td>
<td>apical</td>
<td>sublaminal</td>
<td>postalveolar</td>
<td>alveolar</td>
</tr>
<tr>
<td></td>
<td>den-alv</td>
<td>alveolar</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Articulator</th>
<th>Laminal</th>
<th>Apical</th>
<th>Sublaminal</th>
<th>Postalveolar</th>
<th>Alveolar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tongue body</td>
<td>lower</td>
<td>higher</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sublingual cavity</td>
<td>smaller</td>
<td>larger</td>
<td>larger</td>
<td>smaller</td>
<td></td>
</tr>
<tr>
<td>Length of constriction</td>
<td>greater</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **VOT**: shortest longest VOT
- **F locus**: F2 lower, F2 higher, F2 even higher (BUT F1 lower)
- **F3 and F4**: higher, lower, lower (steeper), F3 & F4 higher
- **Burst**: Lower peak, Higher peak

### SUMMARY OF ARTICULATIONS USED BY SINDHI SPEAKERS

<table>
<thead>
<tr>
<th>Articulator</th>
<th>Laminal</th>
<th>Apical</th>
<th>Laminal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. MD</td>
<td>laminal</td>
<td>apical</td>
<td>laminal</td>
</tr>
<tr>
<td></td>
<td>denti-alveolar</td>
<td>alveolar</td>
<td>alveolar</td>
</tr>
<tr>
<td>2. NI</td>
<td>laminal</td>
<td>apical</td>
<td>laminal</td>
</tr>
<tr>
<td></td>
<td>denti-alveolar</td>
<td>alveolar</td>
<td>alveolar</td>
</tr>
<tr>
<td>Both speakers</td>
<td>laminal</td>
<td>apical</td>
<td>laminal</td>
</tr>
<tr>
<td>---------------</td>
<td>---------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>denti-alveolar</td>
<td>alveolar</td>
<td>alveolar</td>
</tr>
<tr>
<td>Closure duration</td>
<td>longest</td>
<td>shortest</td>
<td>longer</td>
</tr>
<tr>
<td>Tongue body</td>
<td>lower</td>
<td>higher</td>
<td>higher</td>
</tr>
<tr>
<td>Sublingual cavity</td>
<td>smaller</td>
<td>larger</td>
<td>smaller</td>
</tr>
<tr>
<td>Constriction length</td>
<td>greatest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VOT</td>
<td>longer</td>
<td>shortest</td>
<td>longest</td>
</tr>
<tr>
<td>F locus</td>
<td>F2 lower</td>
<td>F2 higher</td>
<td>F2 even higher</td>
</tr>
<tr>
<td></td>
<td>F4 &amp; F3 higher</td>
<td>F4 &amp; F3 lower</td>
<td>F4 &amp; F3 higher</td>
</tr>
<tr>
<td>Burst</td>
<td>higher peak</td>
<td>lower peak</td>
<td>higher peak</td>
</tr>
</tbody>
</table>

**CONCLUSIONS**

(a) Laminal stops have longer closure duration.
(b) Laminal stops have longer VOT than apicals.
(c) Laminal articulations generally have long constriction and therefore justifiably called [+distributed]. Note that such a generalisation cannot, however, be made for alveolar laminal [c] tokens; difference in constriction length was not significant.
(d) Laminal stops are characterized by higher F3 and F4 presumably because of the smaller sublingual resonance cavity than the apicals which have comparatively an increase in the size of the sublingual cavity.
(e) Laminals are characterized by higher peak magnitude than apicals as reflected in the burst transient spectra.

**REFERENCES**

TONE RECOGNITION OF MANDARIN SPEECH

Yasuhisa Niimi, Xiao-Yang, Gian-Geng Ke, Yutaka Kobayashi
Kyoto Institute of Technology

SUMMARY

This paper describes a method for recognizing tones of continuous Mandarin speech. The basic part of the tone recognition system consists of two main components. The first component computes fundamental frequencies every 5 msec, performs the logarithmic conversion of them, and then normalizes them by subtracting a speaker-dependent constant. The normalized Fo-contour is quantized, and fed to the second component, that is, the tone recognition component, which is composed of a set of HMM's, each being trained for one of four Chinese tones and their variations. This approach is new in the following points. (1) Instants in which fundamental frequencies are not extracted, are coded by a special code, while the normalized Fo-contour is conventionally quantized. (2) An input speech is divided into segments by using the amplitude envelope. The isolated tone recognition is performed for short segments, and the continuous tone recognition is conducted for long segments. The tone recognition rate of 72.4% was obtained for five male speakers.

INTRODUCTION

Mandarin (standard Chinese) is a tonal language, in which the same phonetic syllable when pronounced in different tones gives quite distinct meanings. Each Chinese character is a syllable in pronunciation. Although more than 50,000 characters are used in the writing system, there are only 1300 distinguishable syllabic sounds. Each isolated syllable is pronounced in one of four lexical tones. If the tones are ignored, the number of distinct syllables is reduced to about 410. Therefore, the tone recognition plays an important role in automatic speech recognition of Chinese.

![Diagram of four tones](image)

Figure 1. An example of the four tones.
Figure 1 illustrates an example of the typical fundamental frequency (Fo) patterns of the four tones, in reference to which we define the tone 1 (high and smooth tone), the tone 2 (rising tone), the tone 3 (falling and rising tone), and the tone 4 (falling tone). In spite of speaker's age and sex, those patterns are observed clearly in isolated syllables. However, the tone patterns of the syllables in continuous speech are heavily influenced by the surrounding syllables. For example, a reduced pattern (short and falling) is observed especially in de-emphasized syllables, which we refer as to the tone 0. The regular changes of tones are also observed, for example, the first of two successive tone 3's changing to the tone 2.

Several attempts have been made to the application of the hidden Markov modeling technique to the lexical tone recognition of Mandarin speech. Chen, et al. [1] conducted tone recognition for Chinese phonetic alphabets and Chinese digits (monosyllabic words). Yang, et al. [2] carried out a series of experiments to evaluate the offsets of pitch reference base, code book size, and tonal modeling topology, using isolated monosyllabic utterances. These two approaches were based on a delta modulation of pitch sequence, and a combination of vector quantization and HMM techniques. Liu, et al. [3] expanded this method to polysyllabic words.

This paper presents an HMM approach for the tone recognition of continuous Chinese. We have defined six pattern classes, of which the four indicate the four lexical tones, the one represents the reduced tone (tone 0), and the rest is the 'no-pitch' class to represent silent portions or unvoiced consonants. We introduce a new coding scheme for \( \Delta F_0 \)-contour, and a segmentation scheme by using the amplitude envelope. The new coding scheme for \( \Delta F_0 \)-contour is introduced to deal with continuous speech, in which intervals in which Fo's cannot be extracted are coded by a special code, while the normalized Fo's are conventionally quantized. Secondary in order to ample acoustic information in addition to theFo-contour, the amplitude envelope is used to divide an input speech into segments. Then the isolated tone recognition is performed for short segments, and the continuous tone recognition is performed for long segments. The following will describe the outline of the tone recognition system and the experimental results.

**TONE RECOGNITION SYSTEM**

Figure 2 shows the block diagram of the tone recognition system proposed in this paper. The basic part of the system consists of two components: Fo-extraction component and tone recognition component.

1. **Fo-extraction component**  
   The Fo-extraction component computes fundamental frequencies every 5 meec by using the cepstral analysis. The fundamental frequency contour is smoothed with a median smoother, logarithmically converted, and normalized by subtracting a speaker-dependent constant. This constant is decided for every utterance by the formula \( \bar{F}_0 - kt \) where \( \bar{F}_0 \) is the average of fundamental frequencies in the utterance, \( t \) indicates the discrete time at which Fo's are sampled, and \( k \) is a constant empirically determined. This time-dependent constant is introduced to compensate Fo-contour gradually declining at the end of utterances. Intervals in which Fo cannot be extracted, corresponding to pauses like short silences before stop consonants or breath group, or unvoiced consonants, are also detected by the cepstral analysis. Those intervals are coded by a special code denoted by 'c'. Thus an input speech is transformed into a sequence of normalized values of Fo and the special code, which is denoted by \( X_i (i = 1, 2, \ldots) \). A sequence of two dimensional vectors, denoted by \( (u_i, v_i) (i = 1, 2, \ldots) \), is derived from this sequence as follows.

\[
\begin{align*}
  u_i &= \begin{cases} 
  X_{i+1} + X_i & (X_{i+1} \neq c \text{ or } X_i \neq c) \\
  c & (X_{i+1} = c \text{ or } X_i = c)
  \end{cases} \\
  v_i &= \begin{cases} 
  X_{i+1} - X_i & (X_{i+1} \neq c \text{ or } X_i \neq c) \\
  c & (X_{i+1} = c \text{ or } X_i = c)
  \end{cases}
\end{align*}
\]
The two dimensional vectors other than \((c,c)\) are conventionally vector-quantized, and then input to the tone recognition component.

In addition to coding the Fo-contour, this component conducts the segmentation of utterances using the amplitude envelope in order to disambiguate ambiguous tonal patterns which are often observed in two or more successive voiced syllables. For example, a long falling pattern followed by a short rising pattern can be interpreted as the tone 4 followed by the tone 3 or the tone 2 depending on the location of the syllable boundary. Syllable boundaries are determined by the two steps; deep and wide dips of the envelope are first detected as the candidates of syllable boundaries, and then filtered out by the criterion that the interval between two adjacent boundaries be longer than a threshold value.

(2) Tone recognition component We have defined six tone classes for this study. Each pattern class has an HMM whose structure is a left to right type with four states. Exceptions are the classes of the tone 2 and the tone 4, for which two different HMM's are prepared, because if syllables with those tones follow the consonants like /m/, /n/ and /l/, their tonal patterns are quite different from those in the other case. This means that we have eight subclasses.

In the tone recognition phase we used two methods: isolated tone recognition and continuous tone recognition. The isolated tone recognition is applied to a short segment of which the duration is nearly equal to the average length of Chinese syllables. The segment of that length is fed to each of eight HMM's to compute the conditional probability \(P(M/F)\) where \(M\) indicates one of eight subclasses and \(F\) a sequence of codes for the segment. The tone of the segment is the tone class that maximizes \(P(M/F)\). For the continuous tone recognition, we build a large and ergodic HMM in which the eight HMM's are interconnected with equal transition probabilities from one pattern class to another. The initial states of the large HMM are those of its component HMM’s. The initial probabilities for those states are defined to be equal, while those for the rest states are defined to be zero. The final states of the large HMM are also those of the eight small HMM’s. In the tone recognition experiments described later the continuous tone recognition was applied to the long segments or utterances as whole, for which the most likely sequence of labels is decided by Viterbi algorithm.

EXPERIMENTS AND RESULTS

We conducted the following four experiments, each using a different recognition scheme. The first one did not use the segmentation, but the rest used it. (1) Continuous tone recognition. The whole sequence of codes for an utterance is input to the ergodic HMM explained in the previous section. The tone sequence with the highest probability is determined by Viterbi algorithm. (2) Isolated tone recognition 1. Utterances were first divided into several segments...
Table 1. Rates of tone recognition(%) 

<table>
<thead>
<tr>
<th>No. of experiments</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>70.5</td>
<td>74.4</td>
<td>77.7</td>
<td>82.1</td>
</tr>
<tr>
<td>insertion</td>
<td>4.0</td>
<td>0.4</td>
<td>1.7</td>
<td>1.2</td>
</tr>
<tr>
<td>deletion</td>
<td>5.5</td>
<td>9.3</td>
<td>4.1</td>
<td>3.8</td>
</tr>
</tbody>
</table>

by using the amplitude envelope. Then the isolated tone recognition was applied to each segment. (3) Isolated tone recognition 2. Utterances were first segmented in the same way as in the second experiment, but the segments of which the duration was longer than the predetermined threshold (250ms, one and a half of the average length of Chinese syllables[4]) were further divided into shorter segments. Then the isolated tone recognition was applied to each segment. (4) Combination of isolated and continuous tone recognitions. In this scheme the isolated tone recognition was applied to shorter segments than the threshold used in the third experiment, and the continuous tone recognition was applied to longer segments.

We used 92 utterances spoken by a male speaker to compare those four methods. Table 1 illustrates the experimental results. The fourth method, that is, a combination of the isolated and continuous tone recognitions, showed the highest performance, the rate of tone recognition being 82%. Table 1 also shows the use of segmentation is useful in comparison with the simple continuous tone recognition. However, the segmentation tends to increase the rate of missing tones because of its incompleteness. We further tested the fourth method for other five speakers, the two of which are female. The rate of the tone recognition was 72.4% on the average, degraded due to the noisy environment in which the utterances were spoken.

CONCLUSION

This paper has presented an HMM approach for the tone recognition of continuous Chinese. The approach includes the two new ideas. First, a coding scheme for Fo-contour is introduced to deal with continuous speech, in which non-pitch intervals are coded by a special code, while the normalized Fo-contour is conventionally quantized. Second, the segmentation by using the amplitude envelope is introduced. To the resulting segments either isolated tone recognition or continuous tone recognition is applied depending on their length. The tone recognition rate of 72.4% was obtained for five speakers.

In this study some contextual effects caused by the coarticulation were taken into accounts, but not sufficient. Especially modeling the tones 0 and 3 in detail is essential for the better result, left as the future work.

REFERENCES

SPEECH ENHANCEMENT BASED ON THE WAVELET TRANSFORM USING DECAYING SINUSOIDAL WAVELET

Ryouichi Nishimura, Futoshi Asano, Yōiti Suzuki, Toshio Sone
Research Institute of Electrical Communication, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-77, Japan

Summary

Speech enhancement using wavelet transform is studied in this paper. As a fundamental wavelet, we propose the use of decaying sinusoid, which has properties that correlate with those of speech. With this fundamental wavelet, the averaged wavelet transform of noise is subtracted from the wavelet transform of the input signal as a natural extension of Spectral Subtraction method. The results of the computer simulation are also shown.

Introduction

Speech enhancement based on the short-time Fourier transform such as “spectral subtraction[1]” is the most widely used technique. In the short-time Fourier transform, time-frequency resolution is constant over the entire frequency range, and it is uniquely determined by the frame length. On the other hand, an alternate method of time-frequency analysis, wavelet transform, has higher frequency resolution in lower frequency region. This is consistent with the human auditory system, and thus expected to be suitable for speech enhancement. Another advantage of wavelet transform is that the basis of the transform can be chosen from among various functions according to the purpose.

Decaying sinusoidal wavelet

The wavelet transform can be interpreted as the cross-correlation between the input signal and the dilation of fundamental wavelet \( \psi \),

\[
\hat{f}(a, b) = |a|^{-1/2} \int_{-\infty}^{\infty} f(x) \psi \left( \frac{x - b}{a} \right) \, dx,
\]

where \( a \) and \( b \) are called a scale parameter and a shift parameter, respectively.

Voiced speech is decomposed into several decaying sinusoids excited at every pitch period.
Thus, if the decaying sinusoid is used as a fundamental wavelet, its correlation with speech signal is expected to be high, resulting in contrast between the transform of speech and that of noise. A decaying sinusoid with \( Q \approx 5 \) which is used in the later simulation is depicted in Figure 1, where \( Q \) denotes the sharpness of resonance.

**Application to speech enhancement**

In this section, the spectral subtraction method is applied to the wavelet transform with proposed fundamental wavelet. In the original spectral subtraction, the power spectrum of the input signal is estimated from averaged power spectrum of noise and that of the input signal, i.e.

\[
\hat{S}(\omega) = Y(\omega) - E_w[D(\omega)],
\]

where \( \hat{S} \) denotes the estimated short-time power spectrum of speech, and \( Y \) and \( D \) are those of the input signal and noise, respectively. \( E_w[ \cdot ] \) means the ensemble average over frames.

We propose to apply the spectral subtraction to wavelet transform as follows: The wavelet coefficients for the speech are estimated by

\[
\hat{S}^\omega(\omega) = \hat{Y}(\omega) - E_d[\hat{D}(\omega)],
\]

where \( \hat{S}^\omega \) corresponds to the estimate of the wavelet coefficients of speech and \( \hat{Y} \) and \( \hat{D} \) correspond to those of the input signal and noise, respectively. \( E_d[ \cdot ] \) denotes the average over shift parameters.

Temporal information appears in phases of the Fourier transform. In the original spectral subtraction, the estimated power spectrum should usually be supplemented by phase of the input signal. As real fundamental wavelet being employed, the manipulation for adding
Figure 2: The block diagram of speech enhancement system with wavelet transform.

phase is not included in the proposed method. A block diagram of a speech enhancement system using wavelet transform is shown in Figure 2.

Performance evaluation

The performance of the proposed method is evaluated via simulation. The input signal is a Japanese word “sakura” degraded by adding white noise to S/N of 10 dB. The sampling frequency was 20 kHz. The input signal is processed with both the traditional spectral subtraction and the proposed method. The input and output signals are depicted in Figure 3(a) and 3(b). The signal enhanced by the original spectral subtraction and by the proposed method are also shown in Figure 3(c) and 3(d), respectively.

As compared with (c), the amplitude of noise is smaller in (d) while the shape of the speech wave form is close to the original. Table 1 shows the improvement in S/N defined as

\[ P_i = P_s - P_e, \]  

where \( P_i \) denotes the improvement of S/N, and \( P_s, P_e \) are residual power of the input signal and the enhanced signal, respectively. As can be seen from this table, S/N is improved by 2.2 dB for the proposed method.

Table 1: The improvement of S/N in the proposed method over the original signal.

<table>
<thead>
<tr>
<th></th>
<th>2.0 dB</th>
<th>4.2 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>original method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>proposed method</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CONCLUSIONS

Wavelet transform is applied to spectral subtraction, where a decaying sinusoid is used as a fundamental wavelet. The results of the simulation show the improvement in performance.

REFERENCES

SPEECH SYNTHESIS OF TELETEXT

Georg E. Ottesen, Berit Horvei, Sverre Stensby
Acoustics Research Center, SINTEF DELAB, N-7034 Trondheim

SUMMARY
A demonstrator system synthesises Norwegian teletext and subtitles sent on teletext. The user interface is designed for dyslexics and the visually impaired. This paper presents an overview of the system and discusses some of the requirements for a practical system. The system will be demonstrated in a poster session at the conference.

INTRODUCTION
SINTEF DELAB, Telenor Research and the Department of Linguistics at the University of Trondheim have been developing a high quality Norwegian text-to-speech system during the last five years. This project group has also participated in the ONOMASTICA project [1] producing pronunciation rules and a pronunciation dictionary for proper names in Europe. The following results have been implemented in a Norwegian text-to-speech demonstrator system:

- A Norwegian pronunciation dictionary
- A pronunciation dictionary for proper names
- An intonation model for Norwegian
- Synthesis based on sampled waveforms (diphone synthesis)

Two students at NTH have made two systems using synthesis of teletext as an aid for disabled people. The first system was designed with a user interface for blind people [2]. The second one was designed for dyslexics and the visually impaired [3]. The second system was further developed as a demonstrator system for the Norwegian Broadcasting Cooperation (NRK) in 1994.

In Norway about 80,000 persons are visually impaired, and about 150,000 have dyslexia. They constitute a large group who could benefit from text-to-speech conversion of teletext.

SYSTEM OVERVIEW
The main components of the system are a teletext receiver, a text-to-speech synthesis system, and a user interface for dyslexics and the visually impaired.

This demonstrator uses two PC's, a PC board for receiving teletext and a PC board with a digital signal processor (DSP), as shown in figure 1.
Figure 1: Overview of the teletext synthesis system

The hardware requirements can be reduced considerably in a practical system. The extensive use of hardware allows us to build a highly modular system with a fast response and most of the programming in high level language. The response time will be crucial in many applications, and the effect of the response time can be evaluated by adding a delay to this fast prototype system.

THE USER INTERFACE
The user interface is a windows based system and some of the commands are:

- Select TV-channel and teletext page
- Store and retrieve teletext pages
- Scan page by page
- Synthesise whole page
- Synthesise words pointed at
- Synthesise subtitles
- Evaluate reading ability

Dyslectics and the visually impaired can interpret the structure of a page of teletext and visually decode the graphic information. The synthesis is then used as an aid to read the text.

The synthesiser can read fluently page by page. It can also be used as a reading support to read only difficult words. The reader selects the words to be read by pointing with the mouse.

The reading ability is evaluated by registering the time used to read a page and counting the number of words, the number of long words and the number of synthesised words on the page.

SYNTHESISING NORWEGIAN
The four main parts of a speech synthesiser which must be developed for each language are the pronunciation lexicon, the pronunciation rules, the intonation model and the acoustic units for speech generation. The effort on these different tasks varies a lot from one language to another.
The correspondence between writing and pronunciation is rather irregular in Norwegian and hard to cover by rules. Thus a large pronunciation dictionary is needed for Norwegian, as discussed by Sverre Stensby [4]. The dictionary at present consists of 48,000 full form words and 42,000 proper names. Abbreviations are also included in the dictionary.

Compound words are frequent in Norwegian. Inclusion of all the compound words in a lexicon is unfeasible, and rules have been developed to decompose compounds into words found in the lexicon. Words which are not found in the lexicon even by decomposing must be treated by the 750 text-to-phoneme rules.

A real challenge for the text-to-phoneme conversion of Norwegian is the large number of ambiguous words in normal text. This problem is hardly noticed by the intelligent reader, who chooses the right interpretation from the context. In the synthesis program a statistical parser disambiguates the words in a sentence by using probabilities of sequences of parts of speech.

The production of speech sounds is carried out by processing short records of sampled speech. The recorded units are diphones, starting in the middle of one phoneme and ending in the middle of the next one. A high quality recording was carried out in an anechoic chamber, and the speech was digitised at a sampling frequency of 16 kHz. The diphone library consists of 1330 diphones extracted from nonsense testwords by an automatic segmentation procedure [5]. Pitch and duration of the recorded diphones are manipulated by the PSOLA (Pitch Synchronous OverLap Add) algorithm [6].

**SYNTHESISING TELETEXT**

Teletext contains a wide variety of information, e.g. table of contents, news, program overviews, advertisements, weather forecasts and even jokes. This information is presented as running text, tables and graphics.

Texts from Norwegian teletext with about 13,000 words have been analysed by the synthesis system to evaluate the performance of the pronunciation lexicon. This is compared to the performance for a newspaper text of about 10,000 words in figure 2.

This illustrates the complexity of synthesising Norwegian teletext. Both proper names and compound words are frequent. Even with a pronunciation dictionary of 90,000 words and decomposition of compounds, 12% of the words are not found in the lexicon. Many abbreviations are used in the foreign exchange and stock-exchange quotations. English words are frequently used in film and record titles.

Foreign programs are normally subtitled in Norway, only children's programs are dubbed. The subtitles can also be sent on teletext. Synthesising subtitles from teletext will give reading disabled people access to the foreign films. The foreign language can be audible in the background to convey emotions. The synthesiser acts as a reader of the text.

Normally the subtitles appear when people start to speak. In this case, the synthesis must have a fast response to synchronise with the pictures and the foreign voices. This implementation of the speech synthesis system needs only about 140 ms to process two lines of teletext before starting to speak.
CONCLUSION
The PC-based demonstrator system synthesises Norwegian teletext and subtitles sent on teletext. The user interface is designed for dyslectics and the visually impaired.

Teletext is a difficult type of text to synthesise, compared to newspaper texts. Even with a pronunciation lexicon of 90.000 entries and decomposition of compound words, still about 12 percent of the words in teletext are unknown to the synthesis system and must be treated by rules. Thus additional lexicons need to be developed for specific applications.

REFERENCES


SPANISH TEXT-TO-SPEECH, FROM PROSODY TO ACOUSTICS

Jose Manuel Pardo, Francisco M. Giménez de los Galanes, José Angel Vallejo, Miguel Angel Berrojo, José Manuel Montero, Emilia Enriquez*, Ángeles Romero Cortés
* Dep. Lengua Espanola, Univ. Nacional de Educacion a Distancia, Madrid, Spain

INTRODUCTION

We present our Spanish text-to-speech system, which was first developed back in 1984. This first system was a formant synthesizer running on the PROSE2000 TSC board [7]. After that release, a multispeaker system was developed [9] which included some improvements on the duration assignment rules. The intonation model for the intonation generation was modified later in 1989 [6] including a complete description of the Spanish intonational patterns, which made use of the non-orthographic pauses generated by a basic heuristic parser [8].

Our latest work on prosody includes improvements on the categorization module and a better analysis of the syntactic structure of the phrase as well as new rules for duration and f0 assignment. On waveform generation we have increased the number of phonetic rules for the formant synthesizer and we have developed a new acoustic module based on waveform concatenation.

We have also worked on different platforms for real-time implementations, including a version of the Klatt synthesizer running on a DSP board for PC computers and a software implementation of the waveform concatenation system that only needs an additional D/A converter.

LINGUISTIC ANALYSIS AND SYNTHESIS

The first step in every TTS system is to normalize the unrestricted input text. As the system is able to accept not only Spanish words but also special signs like abbreviations, digit strings and control commands, all these items are converted into their equivalent extended sequences.

As Spanish is a language with a rather small set of strict rules for reading a text (including stress position if it is correctly written), the next phase of the linguistic processing, the grapheme-to-phoneme conversion, is greatly simplified. In order to achieve a better prosodic performance, a table of unstressed function words is used. Inter-word sound homologations (taking into account the influence of adjacent words into the one currently being processed) are also applied.

After being preprocessed, the text is labeled by means of a set of morpho-syntactic rules developed for Spanish. According to their endings and the linguistic context, each word is classified into one of the 41 categories available (including several types of conjunctions, adjectives, verbs,...). Although the quality of the labeling has been maximized for this local-context method (analyzing a large corpus of text and adding new rules) and the results are highly satisfactory, we are presently developing new modules for categorization using statistical techniques and phrase-structure syntax. All this new information will be the input to more complex prosodic modules for a better duration and intonation modeling.
The final step in this text processing is the insertion of non-punctuated pauses. Any Spanish reader will not read a text exactly as it is written, stopping only at punctuation marks. In order to imitate this behaviour, the system predicts the position of this user-dependent stops carefully analyzing the string of categories coming from the categorization module. As a wrongly inserted mark is worse than the lack of several, the algorithm is rather conservative, adding only the less conflicting pauses. The heuristics rules applied in this module have been developed iteratively by a group of linguistic experts. The breath groups size is computed taking into account the relation between the number of syllables and the number of the stressed ones, for every two consecutive orthographic marks. If a sentence is larger than 25 syllables or includes more than 6 stressed syllables the parser inserts a pause, so we avoid long and unnatural sentences. The type of the pauses and its duration depend on the so obtained breath group.

Using the structural information resulting from the syntactic processing of the previous module, a more accurate breath group parsing will be achieved.

PROSODY MODULE

In order to improve intelligibility and to give naturalness to synthetic voice we include a prosodic module. This determines the breath groups size, inserts pauses and controls the duration and the F0 contour of every utterance.

Duration is assigned to every phoneme following a modified Klatt's model [5]: inherent phoneme values are modified according to parameters such as stress, syllable structure, within word and sentence position, postvocalic context, word length, etc. Rules and initial phoneme values have been taken from a continuous speech database, using values normalized by the articulatory rate. Speech rate can be changed too, by modifying phonemes and pauses duration in different degree.

The generated F0 contour is a stylized curve made of straight lines between peaks and valleys [6]. It is computed by dividing every breath group into three parts. The initial one, from the beginning to the first stressed syllable, assigns a starting value to the first syllable depending on the type of pause at the end of the breath group, and interpolates linearly to the first stressed syllable F0 value. The central one, from the first stressed syllable to the last one, computes peaks and valleys values upon number of stressed syllables, and interpolates linearly among them. The last one, which includes the final part of the breath group, uses the type of sentence and the position of the last stressed syllable to obtain F0 values. Peaks, valleys and F0 declination rules have been estimated on statistical values extracted from a continuous speech database recorded from four male speakers.

FORMANT SYNTHESIS

The final step of the system is the generation of the waveform. In a formant synthesizer it is generated by a set of dynamic filters (representing the behaviour of the human vocal tract) excited by a source modeling the glottis opening and closure.

In our current implementation up to 31 parameters are available for updating every 10 ms. Pitch, formant frequencies, bandwidths and amplitudes are the most relevant parameters for the production of high quality synthesized voice.

For each Spanish sound the target values for the parameters have been extracted from natural recordings. A set of rules have been developed in order to model the influence of the phonetic context (C-V, V-C-V and V-C, with different transitions if the vowel is frontal).
WAVEFORM CONCATENATION

We have recently incorporated a new acoustic module to our synthesis system based on waveform concatenation. This module overrides the Klatt synthesizer and the corresponding parameter generation module and is based on concatenation of stored diphone-type units. The corpus has been designed under a memory-economy criterion and it consists of 455 units of different length including about 200 triphones and about 20 sulphone units.

This corpus is hand segmented from a recording of logatomes that include a unit surrounded by not strongly coarticulating phones. This set of units is pitch marked following a semiautomatic method which produces marks at the instant of the glottis closure based on the module of the analytical signal [2].

The units are modified before concatenation to adjust the prosody to that of synthesis specified by the prosodic module. We have experimented two different approaches: a PSOLA like approach and another one with spectrum modifications by resampling techniques.

a) PSOLA prosody modification.

We have developed two methods: a Time Domain PSOLA and a Linear Prediction PSOLA. While the first is really cheap in computation, the second one allows a big reduction in the memory requirements as it handles a CELP coded version of the acoustic database. Both systems can achieve really high quality under the assumption of relatively small changes in fundamental frequency and duration.

b) Resampling-based prosody modification.

This method is based on resampling the LPC residual signal (excitation signal) allowing a certain degree of controlled aliasing [3]. After the residual is modified, the spectral envelope is added up by LPC synthesis filtering. This system performs, at least, as the PSOLA systems do, but it allows a bigger range of frequency modification, while the bandwidth is limited to a maximum of 4 to 5 kHz. This upper bound is due to speech spectral characteristics, since a harmonic spectrum can rarely be found beyond that limit.

The concatenation problem is solved differently in the time domain approach than in the LP-PSOLA and the resampling approaches. While a simple averaging of short-time signals (windowed periods) is carried out in the time domain system, a complex spectral treatment has been integrated in the LPC based methods. It involves explicit formant detection and an LPC spectrum modification algorithm [4]. Even when the complexity of the method could not be worthwhile if used only for the smoothing of concatenation discontinuities, this implementation is interesting in terms of arbitrary spectrum modification for speaker variation, joint f0-spectrum modification, etc.

REAL TIME IMPLEMENTATIONS

The latest version of the formant synthesizer is running on a PC based platform which makes use of the VISHA DSP board (based on the AT&T's DSP32C). The analysis as well as the parameter generation are carried out on the X86 processor while the DSP implements the Klatt synthesizer.

About the waveform concatenation methods, we have developed a low cost, real time TD-PSOLA synthesizer. The whole system runs on a i486 PC that makes use of an audio facility. It has been developed under DOS, and it makes use of 2 Mbytes of XMS memory for the corpus storage.
Specifically, we make use of the VISHA PC board (developed in our department) for the D/A subsystem [1] and of the DAC-12 module (also developed in our department) which is a D/A module attachable to the parallel port of the PC.

REFERENCES


MIDSAGITTAL DISTANCES AND AREA FUNCTIONS – BASIC CONSIDERATIONS FOR AN ARTICULATORY MODEL

Bernd Pompino-Marschall
Forschungsschwerpunkt Allgemeine Sprachwissenschaft
Jägerstr. 10/11, D-10117 Berlin, Germany

SUMMARY
The paper describes some problems in data acquisition for constructing an articulatory model that could be used to study articulatory coordination in speech production using an analysis-by-synthesis approach.

In an introductory section different available methods of measuring lingual articulation are compared and discussed with respect to the reconstructability of vocal tract area function. It is argued that most probably a physiologically oriented model of tongue behavior could help in analysing the differently measured tongue movements in a functionally realistic way.

Following a short discussion of the modelling of the articulator-acoustic transformation a sketchy description of our planned procedure is given.

INTRODUCTION
According to the generally accepted view of the source-filter theory of acoustic speech production it is the cross sectional area function along the middle axis of the vocal tract that specifies the acoustic output. Therefore, articulatory data especially of the behavior of the tongue is most important in phonetic research.

With the systems for electromagnetic midsagittal articulometry (EMA; cf. Perkell et al. 1992; Schöne 1988) available today we have an excellent tool for studying lingual articulatory movements without the risks connected with the technique of cineradiography used in the past. But there are some limitations of this technique. Parallel with the x-ray microbeam system (1) it only supplies us with information about the location of single points on the tongue surface and (2) these points have to be in the midsagittal plane. Furthermore, (3) the measurements are restricted to the upper vocal tract. As Lindau-Webb & Ladefoged (1989) have shown, it is nevertheless possible to reconstruct the entire tongue contour from only two measurement points with the aid of further data about the vowel production of the tested subject. Regarding the most front part of the vocal tract the limitation to midsagittal data can also be compensated for in part by parallel electropalatographical (EPG) recordings. But an accurate transformation of sagittal distance of cineradiographic measurements or midsagittal distance measured by EMA (which often will differ from the latter due to lateral tongue grooving) to cross sectional area is still missing (cf. e.g. the modification of Heinz & Stevens’ (1965) power functions by Beaufremps et al. 1995).

Apart from the highly difficult and extremely expensive technique of magnetic resonance imaging (MRI; cf. Bear et al. 1991; Foldvik et al. 1993) there is no possibility of direct measurement of area functions. Therefore we want to develop a 3D articulatory model that will allow us to study lingual articulation in an analysis-by-synthesis fashion.
The model shall be implemented in collaboration with the Technical University of Berlin along the lines of their 'fuzzy head' model for lipreading (cf. Bothe 1994a, b).

**BASIC DATA FOR AN ARTICULATORY MODEL**

The basic data for the model is derived from the measurement of electrode positions of the EPG palate of the subject supplemented by measurements of oral casts (cf. Fig. 1) as well as tongue contour data.

![Diagram](image)

**Fig. 1**: Reconstruction of the subject's palate by measurement of EPG electrode position (black) and oral casts (grey); the grid representing the occlusion plane.

The tongue contour in isolated vowels is determined by oral casts and midsagittal tracings with an EMA coil.

To derive a measure of surface tongue deformation including longitudinal and lateral stretch/compression the oral castings are made with pellets spaced regularly over the oral tongue surface. For these measurements a casting with the tongue pressed against the palate is drilled along a 1 cm grid of the tongue surface. The small drilling holes are filled with dye and the grid this way is applied to the tongue for pellet positioning. Figure 2 illustrates measurements taken from the castings.

![Diagram](image)

**Fig. 2**: Tracings of the midsagittal slice of a casting during the articulation of /i/ (left) and of two coronal slices at different points of the vocal tract (right; palate and tongue contour only).

For the EMA tracings the ruler normally applied to the forehead and the mandible coil to check the correct midsagittal alignment of the EMA helmet was modified to a guide for a slightly bended elastic plastic slat. With a receiver coil glued to the top of this slat we are able to trace the palate as well as the tongue surface during the production of a steady state vowel even
further back than the back coil position under normal recording conditions. The experimental data to be accounted for by the articulatory model will be parallel acoustic, EMA and EPG recordings of sustained articulations as well as connected speech.

**ACOUSTIC CONSIDERATIONS**

![Diagram](image)

Fig. 3: The model of neutral tube disturbance: The length of the arrows represent the influence (ΔFN) of local (1/30 vocal tract length) area (d) increments (light grey) or reductions (dark grey) of a specific degree with respect to the neutral tube (thick black) on the frequency of the first three resonances (shown as standing waves of the neutral tube; after Pomponio-Marschall (1995)). Superimposed the distinctive regions of Carpe & Mrayani (1990) are shown.

As implied already in Ungeheuer (1962) and explicitly stated in Tillmann (1980, pp. 255ff), starting with Webster's horn equation the acoustics of vowel articulation can be expressed as acoustic consequences of local deviations from neutral cross sectional area. These acoustic consequences (expressed as deviations from neutral &-resonances) depend on the energy distribution of the standing waves for the different resonant frequencies within the neutral tube as show in Figure 3. Tube sections of different length along the vocal tract in this way form so-called 'formant shifters' (Tillmann 1980): An increase in area of the glottal half of the vocal tract exerts the same influence on the frequency of F1 as an area reduction within the front half, i.e. a downward shift. Equivalently, for F2 the tube of 1/6 of the entire vocal tract length starting

- 139 -
at the glottis has the reciprocal effect of the most front 1/6 section at the lips, as do the respective neighboring back and front 1/3 tube sections. For F3 this results in a vocal tract partition of 1/10, 1/5, 1/5, 1/5, 1/5, 1/10. The nomograms (e.g. Fant 1960) or the distinctive regions of Carré & Mrayati (1990) reduce to this general principle.

PROSPECT

The area functions to be fed into the acoustic model shall be computed from the articulatory model at variable sections of 1/30 total vocal tract length perpendicular to the middle axis formed by the centers of gravity of the cross sectional areas. With this procedure we hope to arrive at a realistic model that is able to mimic EMA movement data and to produce the same acoustic result.

REFERENCES


VOWEL SPECTRAL CHARACTERISTICS TO DETECT VOCAL STRESS

Robert Ruiz, Claude Legros
Laboratoire d'Acoustique, Université Toulouse-Le Mirail, 5 allées Antonio Machado, 31058 Toulouse Cedex, France

SUMMARY:
The need to achieve knowledge of acoustic signs of stress in speech has increased with the multiplication of recognition and synthesis systems. The first must operate in any "voice situations" and the second should produce speech with emotional content. In this work, the effects of emotions over speech are studied from a cockpit voice record.

Previous studies have indicated that spoken stress seemed to be related with the distribution of energy in the spectrum [2]. In this paper some vowel spectral characteristics are tested in order to detect possible changes. The results indicate high frequency modifications.

INTRODUCTION:
It has been shown that vowel spectrum is modified in stressful situations. Qualitative and quantitative observations have been done. The first are issued from sonagrams [4] and the second ones deal with the frequency and the amplitude of the first formants [1,4,6]. Although the results indicate spectral modifications, formant values do not take into account the whole spectrum. Moreover, in our previous studies we have essentially observe high frequency modifications not only located at resonances but rather for a frequency band.

Therefore, and because high-order acoustical formant characteristics seem to have a more intra and inter speaker variability than low order ones, we have chosen to study the energy distribution in the spectrum.

SPECTRAL CHARACTERISTICS:
Five spectral vowel characteristics have been studied [5].

- The energetical balance frequency (noted Fe) is the frequency for which lower and upper bands energies are equal.
- The spectral balance frequency (noted Fs) is the frequency for which the surface between Cumulative Spectral Probability Diagrams for lower (0 - Fs) and upper (Fs - Fmax) bands is minimal.

Let us consider a discrete spectrum (issued from an FFT), cumulating the number of discrete frequencies in sound level classes of equal width (1dB in our case) leads to a cumulative histogram. The CSPD is obtained by computing one minus this one and applying a polynomial regression [2,3].

When the whole frequency band of the vowel spectrum is divided into two complementary ones the two corresponding CSPD plotted on the same graph are distant (Fig.1). Indeed, sound level decreases with frequency. The spectra studied are normalized to a zero mean in order to avoid the influence of mean sound level on spectral modifications.

From an automatic measure of the surface Δ (Fig.1) between them it is possible to vary the sharing frequency in order to obtain a minimal Δ. This frequency is called the spectral balance frequency Fs.

For Fs the sound level distribution in the two bands is similar. Fs and Δ (for the mean frequency of the second antiresonance of vowels uttered in a resting state) are the characteristics issued from C.S.P.D.
In a second approach, energetical ratios are measured at the second antiresonance frequency. This choice of frequency limit between lower and upper bands comes from previous observations indicating noticeable spectral modifications in the upper band for stressed speakers [3].

- For the frequency of the second antiresonance of each vowel spectrum, the ratio of high frequency energy to the total one is calculated and noted \( \frac{E_h}{E_t} \)

- The same ratio is calculated but for the mean frequency of the second antiresonance of vowels uttered in a rest state. It is noted \( \left( \frac{E_h}{E_t} \right)_{(2)} \).

With these five characteristics the problem is to observe and discuss their sense of variation with increasing stress.

**EXPERIMENT AND RESULTS:**

Data are issued from the magnetic speech record of the Cockpit Voice Recorder (C.V.R) of a crashed plane. French steady-state vowels [a] are analysed by the software I.L.S during their stable period with a cut-off frequency of low-pass filter of 5 kHz; a sampling frequency of 10 kHz; an overlapping of 50%; a hamming window and a pre-emphasis.

A mean spectral envelope is obtained by averaging the Auto-Regressive coefficients (order 15) of each frame of the signal and then applying an FFT (order 10) to the set of mean coefficients. Characteristics are calculated from this mean envelope normalized to a zero mean for Fs and \( \Delta \) estimations. A set of 24 monophongs has been selected from a stressful phase at the end of the flight where the degree of seriousness was increasing. The results are shown on figures 2 to 6.

**DISCUSSION:**

First of all, one can notice a great dispersion of the values of Fs (Fig.2). At the actual state of the research, not enough vowels have been analysed to conclude about the variations of Fs in stressed and normal voice.

On the other hand, Fs variations have a steady increase (Fig.3). Fs and Fe are separate indicators. Straightforward comparison of the energy in the lower and upper bands would not provide the same information about the presence or absence of stress than CSPD comparison. It is more easy to encounter two different spectra which have the same energy in the two bands than the same CSPD. Let us consider the two envelope spectra of Fig.7: their respective energy in the two bands is approximately the same. So Fe is the same for the two spectra. But for the second, Fs and \( \Delta \) are less than for the first one. The simple comparison of energy in the two bands fails to detect any spectral change of this kind.
Fig. 2: variation of $F_s$ versus vowels analysed.  
Fig. 3: variation of $F_e$ versus vowels analysed.  

Fig. 4: variation of $\Delta$ versus vowels analysed.  

Fig. 5 and 6: variation of first $\left(\frac{E_b}{E_t} \right)_{(1)}$ and second $\left(\frac{E_b}{E_t} \right)_{(2)}$ energetical ratios versus vowels.

Secondly, we note that $F_e$, $\Delta$ and the two energetical ratios all indicate an increasing quantity of energy in high frequencies. Indeed, $F_e$, $\left(\frac{E_b}{E_t} \right)_{(1)}$ and $\left(\frac{E_b}{E_t} \right)_{(2)}$ increase.

The decrease of $\Delta$ (Fig. 4) indicates that the two CSPD draw nearer. There is a tendency for the second antiresonance frequency (frequency limit for the computing of $\Delta$) to become the spectral balance frequency ($\Delta$ minimal). This result may confirm our choice of the sharing frequency. An energetical balance arise. It may be due either to a decreasing of low frequency levels, an increasing of high frequency ones or the two cases. Results for $F_e$ and the two ratios confirm the first solution. Without considering the origin of energy movings, $\Delta$ is an indicator of spectral balance. Because its small values are unexpected for vowels, they may be induced by vocal stress. Although the two ratios and $F_e$ increase, their range of variation is smaller than for $\Delta$ (Table 1). Finally, $\Delta$ seems to be the best indicator: it has a unique sense and a great percentage of variation (Table 1).
**Fig. 7: Comparison of two envelope spectra.**

<table>
<thead>
<tr>
<th>Spectral characteristic</th>
<th>Range</th>
<th>Sense of variation with stress</th>
<th>Percentage of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_s ) (Hz)</td>
<td>500 - 5000</td>
<td>Undetermined</td>
<td>900%</td>
</tr>
<tr>
<td>( F_e ) (Hz)</td>
<td>2200 - 2650</td>
<td>Increasing</td>
<td>20%</td>
</tr>
<tr>
<td>( \frac{E_n}{E_1} )</td>
<td>0.5 - 0.65</td>
<td>Increasing</td>
<td>30%</td>
</tr>
<tr>
<td>( \frac{E_s}{E_1} )</td>
<td>0.55 - 0.65</td>
<td>Increasing</td>
<td>18%</td>
</tr>
<tr>
<td>( \Delta )</td>
<td>10 - 2</td>
<td>Decreasing</td>
<td>400%</td>
</tr>
</tbody>
</table>

**Table 1: Spectral characteristics modifications of the vowels studied in C.V.R analysis.**

**CONCLUSION:**

Although the number of vowels analysed is small, the experiment indicates that the spectral characteristics studied seem to vary with vocal stress. \( \Delta \) gives promising results. And even if the measurement of \( F_s \) has failed, the study must be also confirmed by a greater statistical experimentation. The indicators show an increasing amount of high frequency energy with stress and confirm the possibility of stress vocal detection by energetic shifting studies in the spectrum of vowels.

**ACKNOWLEDGEMENTS:** The Services Techniques de la Navigation Aérienne provided the magnetic record of the Cockpit Voice Recorder.

**REFERENCES:**


MODELLING THE NOISE SOURCE IN VOICED FRICATIVES

Christine H. Shadle
Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U. K.

SUMMARY
The classic model of voiced fricatives includes two sound sources, a periodic glottal source and a random noise source at the constriction. The amplitude of the noise source has long been assumed to be modulated by the voicing (Faant, 1960; Stevens, 1971; Flanagan, 1972), but the amount of this change has not been studied systematically, and the possibility of a change in spectral shape has been ignored. Spectral analysis of a voiced fricative that is F0-synchronized by use of a simultaneous laryngograph channel shows that modulation of the noise source by voicing causes not only amplitude changes but also changes in the spectral shape. These changes, consisting of a 5-dB variation in amplitude above 10 kHz in [v], are consistent with changes observed in time-averaged spectra of sustained fricatives at different effort levels, and with spectra of mechanical models having different constriction shapes. The amount of spectral change due to such modulation is likely to vary with the strength and place of the noise source, and thus will depend on the particular glottis-constriction coordination employed by the speaker. Extension of this work to more fricatives and more subjects is planned.

INTRODUCTION
The noise source in voiced fricatives has not received as much attention as that of unvoiced fricatives, in part because the voiced case, with two sound sources, is undoubtedly more complicated, and the unvoiced case cannot be considered to be solved. In this paper results from previous studies are considered together with data from three subjects to correct this imbalance.

The classic model of voiced fricatives includes two sources of sound: a periodic volume-velocity source located at the glottal end of the tract, and a noise source located in the vicinity of the primary tract constriction. The amplitude of the noise source has long been assumed to be modulated by the voicing (Faant, 1960; Stevens, 1971; Flanagan, 1972), although this effect is sometimes neglected (e.g. Stevens et al., 1992). The noise source before modulation is presumed to be similar to that used in models of unvoiced fricatives: it consists of white or broadband noise, and its strength depends primarily on the pressure drop across the constriction. This latter characteristic means that in general it is weaker than the noise source in unvoiced fricatives, that is, it produces less noise because the pressure drop across it is lower. This has been attributed to the need to maintain a significant transglottal pressure drop in order to maintain voicing, which therefore reduces the pressure differential that can be maintained across the constriction (Stevens, 1992). It has been noted, however, that different speakers use different strategies with regard to glottis-constriction coordination, and so the picture is somewhat more complex (Stromberg et al., 1994).

Apart from coordination issues, characterization of the noise source is more complex in certain other respects. First, the geometry of the vocal tract downstream of the constriction has a significant effect on the noise source spectrum, in particular by offering an obstacle to the emerging jet at which noise is generated (Shadle, 1990). Within a voiced-voiceless pair, one could assume that the
geometry and therefore the parameters controlling the noise source spectrum are the same. Some work has been done on characterizing the dependence of spectral amplitude and spectral tilt on pressure drop and constriction area for [s,t,f] (Badin et al, 1994).

Second, while variation in the flowrate through a constant-area constriction can be predicted to change spectral amplitude and tilt of the noise source, it is not clear how such modulation would be timed with respect to glottal vibration. Acoustic variations generated at the glottis will travel at the speed of sound to the constriction; hydrodynamic variations, which may be of similar strength, will convect at a slower rate that depends on vocal tract area and is therefore much more difficult to predict (Davies et al., 1993).

There are therefore two distinct problems in characterizing the noise source in voiced fricatives: understanding the nature of the glottis-constriction coordination, and describing the effect of the modulation imposed by voicing. We focus on the latter in this paper by describing the results of an \( F_0 \)-synchronous analysis of a voiced fricative, and comparing to results of mechanical model studies.

METHOD

The corpuses 2 and 3 used in this paper are part of a larger set of corpuses developed jointly at the University of Leeds, University of Southampton, and the Institut Communication Parlée, I.N.C.P., Grenoble. Corpus 2 consists of the fricatives /f,v,θ,ð,ʂ,z,ʃ,ʒ,ɣ,j,ɣ,ɦ/ sustained for 3 s. Six tokens of each fricative at three effort levels were recorded. In Corpus 3, the nonsense word /pV1,FV2/ was repeated 10 to 13 times during a single breath. For each such item, \( V_1 \) and \( V_2 \) were chosen from /a,i,u,: F \( F \) was one of the set given above. Three speakers, a woman speaker of General American English (CS), a man speaker of French (PB), and a man speaker of German (CD) were recorded speaking this corpus while a variety of acoustic, articulatory and aerodynamic measurements were made.

High-fidelity acoustic recordings were made in an anechoic chamber using a Bruel & Kjaer 4165 1/2” microphone located 1 m in front of the subject’s mouth. Recordings were made with a Sony PCM system at 16 bits with a sampling frequency of 44.1 kHz. A calibration signal was recorded to allow absolute pressure level to be retained. A laryngograph signal was recorded on the second channel. These Hi-Fi recordings were also compared with separately obtained recordings using a Rothenberg mask and oral pressure measurement, and electropalatography (details are given in Stromberg et al., 1994).

The analysis method differs for the two corpuses and is critical to the information obtained. For Corpus 2 an averaged power spectrum was computed for each 3 s token by time-averaging 8 consecutive 20 ms Hanning windows centered in the fricative. For Corpus 3, the eight middle tokens within each /pV1,FV2/ item were analyzed in two ways: by ensemble-averaging across the tokens at specific events throughout the vowel-fricative-vowel sequence (and thus producing one ensemble-averaged spectrum per event of an item), and by \( F_0 \)-synchronous analysis within one token. Both types of averaging depend on accurate labelling of events in the time waveform. The labelling criteria and the ensemble-averaging method have been described elsewhere (Shadle et al., 1992). The \( F_0 \)-synchronous analysis used the laryngograph signal (Lx) to isolate a portion of each cycle. A 2.9 ms Blackman window was then centered on the desired portion, typically corresponding to a peak or a trough of the laryngograph waveform. A time-averaged spectrum was then computed by averaging the eight DFT’s of the windowed portions during the steady-state of the fricative. This allowed computation of a ‘peak’ or ‘trough’ spectrum while retaining the benefits of spectral averaging.

RESULTS

Fig 1 shows spectra for one token each of three effort levels of sustained [f] and [v]. The primary difference is, of course, the increase in low-frequency energy in [avv] due to the first few voicing harmonics. Above 2 kHz the voiced fricatives are lower in amplitude than unvoiced; this effect is uneven across the frequency range for these examples, but is a consistent difference across the entire
corpus. Increasing effort increases the amplitude by 5 dB or more for both voiced and voiceless, particularly at higher frequencies. The peak frequencies are relatively constant, indicating that place of constriction and location of noise source are constant; the uneven increase of amplitude across the frequency range may indicate that cross-sectional shape of the constriction changes significantly with effort level (contrast the [ss] spectra for the same subject shown in Badin et al., 1994).

Ensemble-averaged spectra of Corpus 3 (not shown) show similar changes throughout unvoiced fricatives, but less evidence of change throughout voiced fricatives. This is probably an artifact of the labelling criteria; transition regions of voiced fricatives are significantly longer, and steady-state regions significantly shorter, than of unvoiced fricatives.

Fig. 2 shows three spectra generated by $F_0$-synchronous analysis of one token of PB saying [puvu]. There are virtually no differences between the 'Peaks' spectrum, corresponding to the peaks in the laryngograph spectrum, and the two 'Troughs' spectra, until about 8 kHz; above 10 kHz, there is a consistent 5 dB difference. This difference was maintained throughout the steady-state portion of the fricative.

DISCUSSION
The aerodynamic measurements, and previous analysis of them (Stromberg et al., 1994), were used to calculate the average fluctuation in volume and particle velocity mid-voiced-fricative in order to obtain a rough comparison to mechanical model data (Shadle, 1985) in which parameters can be measured more precisely. Typical variation in volume velocity of 30 cm$^3$/s corresponds to a variation in particle velocity in the constriction of 500 cm/s. This amount of variation at the same flowrate in mechanical models with three different constrictions produced a range of spectral shapes. The change in amplitude at 10 kHz was computed in each case; the 5 dB difference noted in Fig. 2 falls within the 3.7 dB range of variation noted in the three mechanical models. This helps to validate the short-time-window $F_0$-synchronous analysis, but also indicates that the amount of amplitude variation resulting from $F_0$ modulation may well depend both on the fricative and the particular 'operating point' chosen for a given token.

CONCLUSIONS
Spectral analysis of a voiced fricative that is $F_0$-synchronized by use of a simultaneous laryngograph channel shows that modulation of the noise source by voicing causes not only amplitude changes but also changes in the spectral shape. These changes, consisting of increased amplitude

- 147 -
Figure 2: Spectra generated by $F_0$-synchronous analysis of one token of PB saying [puvu]. 'Peaks' curve is average of 8 consecutive Lx peaks, centered on the steady-state portion of [v]; 'Troughs 1' is centered on the preceding troughs, 'Troughs 2' on the following troughs. (See text.)

above 10 kHz, are consistent with changes observed in time-averaged spectra of sustained fricatives at different effort levels, and with spectra of mechanical models having different constriction shapes. The amount of spectral change due to such modulation is likely to vary with the strength and place of the noise source, and thus will depend on the particular glottis-constriction coordination employed by the speaker. Extension of this work to more fricatives and more subjects is planned.

ACKNOWLEDGEMENTS
Thanks to Dr. Celia Scully for helpful discussions. This work was supported in part by a European CEC-ESPRIT project SPEECH MAPS (CEC-SCI* 0147C(EDB)).

REFERENCES


VOWEL-CONTINGENT EFFECTS IN THE PERCEPTION OF STOP CONSONANTS UNDER AUDIO-VISUAL PRESENTATION

Sumi Shigeno
College of Liberal Arts and Sciences, Kitasato University, 1-15-1, Kitasato, Sagamihara-shi, Kanagawa, 228 Japan

SUMMARY

This study investigated the vowel-contingent feature integration process of the perception of stop consonants under audio-visual conditions. The experiment explored whether the identification of a CV syllable, consisting of a stop consonant and a vowel, was vowel contingent or not under congruent and incongruent vowel contexts through auditory, visual, and audio-visual exposures. Results show that the integration process between audition and vision greatly depends on the vowel context, although the interaction between audition and vision takes place to a lesser extent even if the vowel context is discrepant between audition and vision. The audio-visual integration process was discussed mainly from the perspective of perceptual set and feature interaction of audition and vision.

INTRODUCTION

Cooper (1974), by carrying out selective adaptation experiments, indicated that the effect of voicing adaptation was contingent on vowel quality and called it the "contingent adaptation effect."[1] The results of vowel-contingent selective adaptation raised the question of whether adaptation occurs entirely at an auditory level of processing or at both an auditory and more abstract phonetic levels. Although vowel-contingent effects have not been examined in audio-visual presentation conditions, it was reported that the extent of the McGurk effect [2] was different among different vowel contexts. Green et al. (1988) found that the McGurk effect was largest for the /i/ vowel, moderate for /a/, and almost nonexistent for /u/ [3].

In the present study, the vowel contingent effects of audition and vision were examined with the expectation that results might give us an important clue in understanding the process of how visual information influences auditory information in speech perception.

METHOD

Subjects

The subjects were seven native speakers of Japanese with no known history of either hearing disorders or visual handicaps.

Stimuli

A face of a Japanese female speaker was videotaped using an
8 mm videotape recorder (SONY, ACP-80). She clearly pronounced each of nine consonant-vowel (CV) syllables. They consisted of three consonants (/b/, /d/, and /g/) and three vowels (/i/, /a/, and /u/). For selecting the best audio-visual stimuli, a pretest was conducted. Two listeners, including the experimenter, separately selected the best (or the most natural) sound and articulatory movement. After this selection, the best auditory stimuli were dubbed to the best visual stimuli. Thus, 81 kinds of audio-visual stimuli were prepared. The timing of dubbing was accommodated to give the most natural impression. Adjusting the dubbing timing by a 33-ms frame unit, the audio-visual stimuli were made so that the sound coincides with the articulatory movement of the release of the consonant in the CV syllable of each utterance. The duration of the sound was between 370-400 ms. Subjects were exposed to a 200 ms pure tone one second prior to each presentation of the face of the speaker. The duration of visual stimuli was 3 s and the voice was heard 1.5 s after the onset of the visual stimulus.

Procedure

Each syllable was played for the subjects under three presentation conditions.

Auditory-alone condition: The subjects were presented only with auditory stimuli. They were instructed to write down what they heard.

Visual-alone condition: The subjects were presented only with visual stimuli. They were required to keep watching the speaker's articulatory movements and asked to write down what they saw. Five of the subjects participated in this condition.

Audio-visual condition: The subjects were presented with audio-visual stimuli. They were required to keep watching the speaker's face focusing on her lips and her articulatory movements. They were asked to write down what they heard, not what they saw.

The speaker's face was presented at a 1 m viewing distance. Auditory stimuli were presented through two loudspeakers attached to both sides of the monitor. In the case of the visual-alone and audio-visual conditions, the experimenter always checked whether subjects kept watching the face on the monitor during the presentation of the visual and audio-visual stimuli.

The stimulus series consisted of 81 CV syllables that were presented in a randomized sequence. Each auditory or visual CV syllable was judged 18 times in the auditory-alone or the visual-alone condition and eight times in the audio-visual condition. In every condition the inter-stimulus interval (ISI) for judgment was 3 s. The auditory stimuli were presented to the subjects at about 75 dB(A) in a quiet room.

RESULTS AND DISCUSSION

Auditory-alone condition: The percentages of correct identification of the nine syllables nearly reached 100. There was no confusion among the different vowel environments at all.

Visual-alone condition: Generally, the averaged correct identification percentages were lower than those in the auditory-alone condition.

Audio-visual condition (Table 1): When the auditory and visual stimuli shared the same vowel (ex., /ba/-voice /ga/-lips), the visual information greatly influenced the auditory judgments.
except in the /u/ environment. The vowels /a/ and /i/ caused a number of fusion responses. For example, in a /ba/-voice and /ga/-lips presentation, subjects arrived at the unifying percept [da] by integrating auditory and visual information (12%). The results were confirmed by a chi-square test that compared the frequencies of /ba/ and /da/ responses in the /ba/-lips condition with those in the /da/-lips condition ($X^2=12.32, p<.005, df=1, N=112$). However, when the vowel environment was /u/, the fusion responses were very few. The present results clearly indicate that the McGurk effect was defined by the vowel context.

On the other hand, when the auditory and visual stimuli did not share the same vowel information, the visual presentation influenced the auditory judgments less than in cases in which

Table 1. The average percentage of correct CV-syllable recognition in the identification test at each of the three vowel contexts in audio-visual condition.

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>ba</th>
<th>da</th>
<th>ga others</th>
<th>Response</th>
<th>bi</th>
<th>di</th>
<th>gi others</th>
<th>bu</th>
<th>du</th>
<th>gu others</th>
</tr>
</thead>
<tbody>
<tr>
<td>V=ba</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>98</td>
<td>0</td>
<td>0 2</td>
<td>bi</td>
<td>98</td>
<td>0</td>
<td>0 0</td>
<td>bu</td>
<td>98</td>
<td>0 0 2</td>
</tr>
<tr>
<td>da</td>
<td>5</td>
<td>77</td>
<td>0 18</td>
<td>dl</td>
<td>0 100</td>
<td>0</td>
<td>0</td>
<td>du</td>
<td>4 96</td>
<td>0 0</td>
</tr>
<tr>
<td>ga</td>
<td>0</td>
<td>0 98</td>
<td>2  gl</td>
<td>0 0 88 12</td>
<td>0 0</td>
<td>100</td>
<td>0</td>
<td>gu</td>
<td>0 0</td>
<td>100 0</td>
</tr>
<tr>
<td>V=da</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>73</td>
<td>23</td>
<td>0 4</td>
<td>bi</td>
<td>86</td>
<td>14</td>
<td>0 0</td>
<td>bu</td>
<td>100</td>
<td>0 0 0</td>
</tr>
<tr>
<td>da</td>
<td>2</td>
<td>91</td>
<td>0 7</td>
<td>dl</td>
<td>0 98</td>
<td>0</td>
<td>2 0</td>
<td>du</td>
<td>0 100</td>
<td>0 0</td>
</tr>
<tr>
<td>ga</td>
<td>0</td>
<td>2 98</td>
<td>0  gl</td>
<td>0 0 98 0</td>
<td>0 0</td>
<td>2 98</td>
<td>0</td>
<td>gu</td>
<td>0 2</td>
<td>98 0</td>
</tr>
<tr>
<td>V=ga</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>88</td>
<td>12</td>
<td>0 0</td>
<td>bi</td>
<td>91</td>
<td>9</td>
<td>0 0</td>
<td>bu</td>
<td>100</td>
<td>0 0 3</td>
</tr>
<tr>
<td>da</td>
<td>2</td>
<td>89</td>
<td>0 9</td>
<td>dl</td>
<td>0 98</td>
<td>2</td>
<td>0</td>
<td>du</td>
<td>2 98</td>
<td>0 0</td>
</tr>
<tr>
<td>ga</td>
<td>0</td>
<td>0 98</td>
<td>2  gl</td>
<td>0 0 89 11</td>
<td>0 0</td>
<td>100</td>
<td>0</td>
<td>gu</td>
<td>0 0</td>
<td>100 0</td>
</tr>
<tr>
<td>V=bi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>100</td>
<td>0 0</td>
<td>0 0</td>
<td>bi</td>
<td>100</td>
<td>0 0</td>
<td>0 0</td>
<td>bu</td>
<td>100</td>
<td>0 0 0</td>
</tr>
<tr>
<td>da</td>
<td>0</td>
<td>84</td>
<td>0 16</td>
<td>dl</td>
<td>0 98</td>
<td>0</td>
<td>2 0</td>
<td>du</td>
<td>0 91</td>
<td>0 0</td>
</tr>
<tr>
<td>ga</td>
<td>0</td>
<td>0 96</td>
<td>4  gl</td>
<td>0 0 86 14</td>
<td>0 0</td>
<td>100</td>
<td>0</td>
<td>gu</td>
<td>0 0</td>
<td>100 0</td>
</tr>
<tr>
<td>V=di</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>86</td>
<td>12</td>
<td>0 2</td>
<td>bi</td>
<td>62</td>
<td>36</td>
<td>0 2</td>
<td>bu</td>
<td>100</td>
<td>0 0 0</td>
</tr>
<tr>
<td>da</td>
<td>0</td>
<td>89</td>
<td>0 11</td>
<td>dl</td>
<td>0 100</td>
<td>0</td>
<td>0</td>
<td>du</td>
<td>0 100</td>
<td>0 0</td>
</tr>
<tr>
<td>ga</td>
<td>0</td>
<td>0 96</td>
<td>4  gl</td>
<td>0 0 95 5</td>
<td>0 0</td>
<td>2 96</td>
<td>2</td>
<td>gu</td>
<td>0 2</td>
<td>96 2</td>
</tr>
<tr>
<td>V=gi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>84</td>
<td>16</td>
<td>0 0</td>
<td>bi</td>
<td>64</td>
<td>32</td>
<td>2 2</td>
<td>bu</td>
<td>100</td>
<td>0 0 3</td>
</tr>
<tr>
<td>da</td>
<td>0</td>
<td>93</td>
<td>0 7</td>
<td>dl</td>
<td>0 100</td>
<td>0</td>
<td>0</td>
<td>du</td>
<td>0 100</td>
<td>0 0</td>
</tr>
<tr>
<td>ga</td>
<td>0</td>
<td>0 100</td>
<td>0  gl</td>
<td>0 0 100</td>
<td>0 0</td>
<td>2 98</td>
<td>0</td>
<td>gu</td>
<td>0 0</td>
<td>2 98 0</td>
</tr>
<tr>
<td>V=bu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>95</td>
<td>0 0</td>
<td>5 0</td>
<td>bi</td>
<td>100</td>
<td>0 0</td>
<td>0 0</td>
<td>bu</td>
<td>100</td>
<td>0 0 0</td>
</tr>
<tr>
<td>da</td>
<td>0</td>
<td>89</td>
<td>0 11</td>
<td>dl</td>
<td>0 100</td>
<td>0</td>
<td>0</td>
<td>du</td>
<td>0 93</td>
<td>0 0</td>
</tr>
<tr>
<td>ga</td>
<td>0</td>
<td>0 93</td>
<td>7  gl</td>
<td>0 0 93 7</td>
<td>0 0</td>
<td>100</td>
<td>0</td>
<td>gu</td>
<td>0 0</td>
<td>100 0</td>
</tr>
<tr>
<td>V=du</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>89</td>
<td>7 0</td>
<td>4 0</td>
<td>bi</td>
<td>88</td>
<td>12</td>
<td>0 0</td>
<td>bu</td>
<td>88</td>
<td>5 7 2</td>
</tr>
<tr>
<td>da</td>
<td>0</td>
<td>88</td>
<td>0 12</td>
<td>dl</td>
<td>0 98</td>
<td>0</td>
<td>2</td>
<td>du</td>
<td>0 100</td>
<td>0 0</td>
</tr>
<tr>
<td>ga</td>
<td>0</td>
<td>0 96</td>
<td>2  gl</td>
<td>0 0 96 4</td>
<td>0 0</td>
<td>100</td>
<td>0</td>
<td>gu</td>
<td>0 0</td>
<td>100 0</td>
</tr>
<tr>
<td>V=gu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ba</td>
<td>91</td>
<td>9 0</td>
<td>0 0</td>
<td>bi</td>
<td>82</td>
<td>18</td>
<td>0 0</td>
<td>bu</td>
<td>87</td>
<td>5 4 4</td>
</tr>
<tr>
<td>da</td>
<td>0</td>
<td>89</td>
<td>2 9</td>
<td>dl</td>
<td>0 100</td>
<td>0</td>
<td>0</td>
<td>du</td>
<td>0 100</td>
<td>0 0</td>
</tr>
<tr>
<td>ga</td>
<td>0</td>
<td>0 100</td>
<td>0  gl</td>
<td>0 0 100</td>
<td>0 0</td>
<td>0 100</td>
<td>0</td>
<td>gu</td>
<td>0 0</td>
<td>0 100 0</td>
</tr>
</tbody>
</table>
they shared the same vowel information. The results from trials involving the /u/ environment show, however, that visual influence on audition is almost nonexistent. It was found that the /u/-voice itself was little influenced by any visual environment. The reason why visual presentation of /u/-vowel hardly influenced the auditory judgments might be due to the difficulty for the subjects to detect the articulatory movements of the speaker; the articulatory movements in cases of lip bursting are small and difficult to detect in the /u/ environment.

It goes without saying that there are no actual spectral overlaps between the two modalities—audition and vision. Present results indicate that auditory and visual information are integrated not at a unitary level but at, at least, two levels. One possibility is that the auditory and visual information is evaluated independently and then integrated to achieve perceptual recognition [4]. This process might explain the finding that vowel contingent anchoring is much greater when the same vowel is shared between vision and audition than when it is not. The other possibility is that the preparatory and actual movements of articulation of a speaker play an important role in helping subjects construct a perceptual set before the actual articulatory movements, and assist them in perceiving the audio-visual stimulus as they do. In the case of the /b/-vowel syllable we can easily recognize that a speaker is ready to utter a syllable starting with /b/ before seeing the actual utterance. Motor commands of articulation for speech might be involved in constructing the perceptual set.

CONCLUSIONS

The results clearly indicate that the visual influence on audition is greater when the vowel context of vision and audition is congruent rather than incongruent. and that /a/ and /i/ vowel environment produce more fused responses than the /u/ environment. It is quite natural to believe that several processing mechanisms are involved in the integration of visual and auditory information, such as vowel-dependent and vowel-independent channels.

ACKNOWLEDGMENTS

This research was supported by Grant-in-Aid for Scientific Research on Priority Areas (No. 06205211) from the Japanese Ministry of Education, Science, and Culture.

REFERENCES

PRONUNCIATION LEXICONS FOR SPEECH SYNTHESIS AND SPEECH RECOGNITION

Sverre Stensby, Berit Horvei, Georg E. Ottesen, Arne Kjell Foldvik*  
Acoustics Research Center, SINTEF DELAB, N-7034 Trondheim  
* Department of Linguistics, University of Trondheim, N-7055 Dragvoll

SUMMARY
This paper describes background and method for establishing two pronunciation lexicons for Norwegian. Parts of the work has previously been presented in Stensby (1991, 1992, 1993a,b).

INTRODUCTION
Large-scale, high-quality, machine-readable dictionaries are a necessity for progress in speech and language technology. Linguistic processing modules which perform syntactic parsing are included in systems for speech synthesis and recognition. Based on dictionaries and grammars such modules supply information for e.g. prosodic processing, disambiguation and pronunciation. This is further discussed in Hedelin (1992).

In reading aloud and listening, most of the words are known to us. We thus know their pronunciation, meaning and function in the sentence. The inclusion of lexicons in speech systems is a natural way of giving such systems similar information about the words used.

LEXICON CONTENT
By lexicon we mean a dictionary in machine-readable form. Each entry comprises the word in normal orthography, the phonemic transcription, information on the word class and if necessary a paradigm code. The paradigm code is used for rule-based generation of inflected forms.

The phonemic transcription is based on a dialect widely used in and around the capital of Oslo. In addition to the phonemes the transcription includes marks for main and secondary stress placed at the syllable boundary. The main stress mark also includes distinction between two types of word tone. In Norwegian there is a complex and irregular relation between orthography and pronunciation, and particularly the word tone is unpredictable from orthography alone. This entailed the inclusion of phonemic transcription.

Word tone
Norwegian has a limited use of word tone, and in most Norwegian dialects there is a distinction between two tones named tone 1 and tone 2. The difference is mainly manifested in the syllable carrying the main stress. In the Oslo dialect that we have worked on the fundamental frequency for tone 2 typically starts on a relatively high frequency and has a distinct fall in the course of this syllable, while tone 1 starts on a lower frequency and thus lacks this initial fall.
There are more than two thousand pairs of words with identical sound segments distinguished by word tone alone. Many of these are homographs. E.g. the Norwegian word *tømmer* has tone 1 or tone 2 depending on the meaning: [ˈtømər] = noun *timber* has tone 1, and [ˈtɔmɔr] = verb *empty* and noun *reins* has tone 2. The marks `1` and `2` indicate the position of the syllable which carries the main stress and marks tone 1 and tone 2 respectively.

**LEXICON DEVELOPMENT**

Two lexicons have been established; one for proper names and the other one for ordinary words. The latter is based on full word forms inflected by rule, i.e. most of the inflected forms of the headwords (the lexemes) are included in the lexicon. Compared to a lexicon of headwords only, this gives an easy lexicon entry for most words at the cost of increased data quantity. The inflected form entries also have a reference to the headword.

The basis for the lexicon is data-lists of headwords in normal orthography. The main outline of our work is:

- Suggestion for pronunciation
- Correction by students
- Formal phonotactical control
- Final control by senior phonetician

The lists may include rudimental pronunciation and grammatical information. In cases of no pronunciation, a suggestion for pronunciation is supplied by a rule-based program for text-to-phoneme conversion. The lexicon is divided into handy sized datafiles and these are examined by phonetics students. Then the corrected files are checked and adjusted for formal errors in the phonotactical control before the final preparation discussed below.

**LEXICON OF ORDINARY WORDS**

This lexicon has been established on behalf of Telenor Research. The basis is a limited Norwegian dictionary and a list of the most common words in newspaper text. Jointly this gives a basic word list of 14000 headwords. The pronunciation and grammatical information given in the original dictionaries have been improved and extended. The grammatical information for the words includes the word class and inflection pattern.

There is a distinction between open and closed word classes. The set of words belonging to the closed classes is mainly stable with time and is on the whole covered by a finite number of words. The closed word classes are practically completely covered by the lexicon.

**Inflection by rule**

Norwegian is partly an agglutinating language with many inflected forms created by adding suffixes to the words. The inflected full word forms are generated by rule from the words in the basic word list. Typically 4, 6 and 7 full word forms for nouns, verbs and adjectives respectively are generated, though there may be additional forms due to alternatives. The resulting lexicon of inflected full word forms has 56000 entries.

We found it handy to include most words in the common framework for inflection by establishing pattern paradigms also in less frequent cases, though this imposed an increase in the number of paradigms. The inclusion of inflection patterns for the phonetic transcription caused an even further enlargement in the number of paradigms increasing the number of patterns to 140 mainly due to unpredictable changes in word tone.
The passive form of verbs and the genitive forms of nouns and proper names, however, are not included in the lexicon. The reason is that in Norwegian orthography these forms are made by adding an s to the word. Words ending in an s and not found in the lexicon may thus be found removing the s. When a genitive or passive is found the appropriate phonemic transcription may be constructed by simple rules, in most cases just by adding an s to the word found.

Final check
The phonemic transcription and the inflection code were corrected by students. The final check was done by a program with combined inflection and synthesis. In this system all the inflected forms were generated and the results presented on screen and in the form of synthetic speech.

The procedure for the final check was: A word was selected from the list and all the derived inflected forms were displayed in normal orthography accompanied by the phonemic transcription. The paradigms were proof-read and synthesised in a speech synthesiser and these two presentations complemented each other. Wrong vowel quality and word tone were effectively detected through listening. An example of screen layout is shown in Table 1.

<table>
<thead>
<tr>
<th>Orthography</th>
<th>Transcription</th>
<th>Code</th>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>bil (=car)</td>
<td>b:il</td>
<td>s.m1m2</td>
<td>Indefinite singular</td>
</tr>
<tr>
<td>leve (=live)</td>
<td>&quot;leva&quot;</td>
<td>v.v3t1</td>
<td>Infinitive</td>
</tr>
<tr>
<td>kaste (=throw)</td>
<td>&quot;kasta&quot;</td>
<td>v.v1</td>
<td>Infinitive</td>
</tr>
<tr>
<td>leve</td>
<td>&quot;leva&quot;</td>
<td>v.v3t1.inf</td>
<td>Infinitive</td>
</tr>
<tr>
<td>lever</td>
<td>&quot;levor&quot;</td>
<td>v.v3t1.pres</td>
<td>Present tense</td>
</tr>
<tr>
<td>levde</td>
<td>&quot;levda&quot;</td>
<td>v.v3t1.pret</td>
<td>Past tense</td>
</tr>
<tr>
<td>levd</td>
<td>&quot; lev&quot;</td>
<td>v.v3t1.perfp</td>
<td>Past perfect</td>
</tr>
<tr>
<td>levende</td>
<td>&quot;levana&quot;</td>
<td>v.v3t1.presp</td>
<td>Present participle</td>
</tr>
<tr>
<td>lev</td>
<td>&quot;lev&quot;</td>
<td>v.v3t1.imp</td>
<td>Imperative</td>
</tr>
</tbody>
</table>

Table 1. Example page of the control and inflection program. A section of the input file with the selected verb "leve" is shown on top, while the paradigm of the selection is included below.

THE LEXICON OF NAMES
This lexicon is part of a European project named Onomastica, Schmidt (1993), whose aim is to establish an European multi-language pronunciation dictionary of proper names for persons, firms, places, and addresses. The project is managed by CCIR in Edinburgh. The Norwegian part comprises approximately 75000 Norwegian names and Norwegian pronunciation of 10000 names from other European countries.

In this project the suggested phonemic transcription is likewise corrected by students. All syllable boundaries are indicated in the transcription and this opens for an effective phonotactic control to improve and secure the overall quality. The system for the phonotactic control is DIPASYS (1993). The most important rules are:
- All symbols used in the transcription must be selected from the established set.
- Every syllable should contain one, and only one, nucleus.
- Each word should contain one, and only one, primary stress mark.
- Syllable initial and syllable final consonants or consonant clusters must be selected from the established set.
Based on the students' transcriptions the names were synthesised by a speech synthesiser. The final check and corrections were done by an associate professor in phonetics. For this task the synthesised speech and the error report from the phonotactic control were utilised in addition to the transcription itself.

**FORMAT CONVERSION**

The transcription is based on a subset of the symbols defined by the International Phonetic Association. In the final transcription the SAMPA (Speech Assessment Methods, Phonetic Alphabet) is used.

The persons involved in the project were more familiar with the IPA symbols so we decided to have a system displaying these symbols and finally converting them to SAMPA. The task of conversion was further complicated by the fact that more computer systems were involved. The linguists worked with MS-Word on Apple-machines while the engineers used DOS and WINDOWS on IBM-compatible PCs, and all symbols had to be displayed and printed correctly at every system. Flexible conversion between all these systems was established.

**CONCLUSION**

One Norwegian lexicon of ordinary inflected words and another one for proper names have been established. The entries comprise the orthographic word, a phonemic code, and grammatical information. The lexicons may be used in linguistic research, e.g. for text tagging, and in speech synthesis and speech recognition. High quality is ensured by a procedure for formal control and by thorough checking.

**ACKNOWLEDGEMENT**

The work reported here was funded by Telenor Research and the Royal Norwegian Council for Scientific and Industrial Research. The work has been carried out in co-operation with the Department of Linguistics at the University of Trondheim, Norway.

**REFERENCES**


ELECTRIC FIELD EFFECT ON THE ABSORPTION OF ULTRASONIC WAVES IN CONDENSER OIL

Andrzej Skumiel, Mikolaj Labowski
Institut of Acoustics, Adam Mickiewicz University, 60-769 Poznań, ul. Matejki 48/49, Poland

SUMMARY

The effect of an external DC electric field on the amplitude coefficient of ultrasonic wave absorption in condenser oils was studied. The theoretical electric field strength distribution in the neighbourhood and inside gas bubbles in the liquid was determined. Experimentally, the presence of the gas bubbles was shown to be decisive for the value of ultrasonic wave absorption. Our work extended over a wide range of field strengths and temperatures. The experimental procedure applied can serve to determine the effective size of the bubbles.

THE ELECTRIC FIELD SURROUNDING A GAS BUBBLE

It will be remembered [1] that a dielectric liquid sample, even if purified repeatedly by physico-chemical methods, still contains an amount from small gas bubbles. The gaseous cavity (the bubble) is surrounded by a region of strong inhomogeneity of the electric field [2] causing the free electric charges present in a real medium to be sucked away towards the surface of the sphere. If the resultant charge thus attached to the bubble is non-zero, the bubble will drift in the electric field in the direction of the electrode of opposite sign.

With regard to the initial component of the field, $E_o$, we finally obtain the following expressions for the real electric fields existing inside and outside the bubble:

$$E_{int(xe)} = \frac{3\varepsilon_r}{1 + 2\varepsilon_r} E_o$$  $$E_{ext(xe)} = \frac{3}{1 + 2\varepsilon_r} E_o$$  \hspace{0.5cm} (1)

The respective field gradient attaches the free charge to the bubble and causes the latter to wander towards the neutralizing electrode.

In practice, the formation and growth or vanishing of bubbles in oil can be due to a number of factors, such as temperature, moisture, the chemical composition of the oil, and the electric field strength. An intense electric field can moreover destroy the molecular structure of the paraffine hydrocarbons by disrupting the bonds between the carbons and hydrogens, leading to the emergence of hydrogen and methane bubbles. On the other hand, molecules of aromatic hydrocarbons adsorb atoms of free hydrogen thus impeding the formation of hydrogen bubbles.

In the range of lower temperatures, electrophoresis is the process responsible for this behaviour of the amplitude absorption coefficient of ultrasonic waves. In the present case, it consists in a drift of the air bubbles into the region of the oil where the electric field is weaker.

The expression for the respective electrostatic force runs as follows [3]:

$$F = R^2 \varepsilon_{ol} \frac{\varepsilon - \varepsilon_{ol}}{\varepsilon + 2\varepsilon_{ol}} E \nabla E$$  \hspace{0.5cm} (2)

where $R$ is the radius of the bubble, $\varepsilon_{ol}$ - the electric permittivity of the oil, $\varepsilon$ - the electric
permittivity of the gas in the bubbles, and $\mathbf{E}$ - the electric field strength vector in a given point of the field. The rate of removal of the gas bubbles from the region of greater electric field strength is dependent on the difference between the permittivity inside the bubble and that of the oil ($\varepsilon - \varepsilon_0$) as well as on the size $R$ of the bubble.

This contributes to modify the acoustic properties of the medium under the action of an external electric field, both by affecting the propagation velocity of ultrasonic waves and their absorption coefficient in the medium. These factors act in a way to remove the gas bubbles from the liquid [3].

THE METHOD OF MEASUREMENT

In our study of the changes in absorption coefficient of ultrasonic waves propagating in oils acted on by an external electric field we applied the pulse method. We made use of a laboratory setup from "MATEC" additionally supplied with a measuring block involving analog memory [4] thus achieving a high degree of accuracy in our measurements.

In the course of our studies we measured the amplitude of the ultrasonic pulse of frequency 6 MHz before and after applying the DC electric field at right angles to the oil sample. The DC electric field ranged from 0 to 300 kV/m and the temperature from 0 to 40° C.

The example of the time-variations of the ultrasonic pulse amplitude recorded for the mineral oil investigated is given in Fig 1.

We derive the following expression for $\Delta \langle \alpha_E \rangle$, the variation of the mean amplitude absorption coefficient of the wave due to the external field, as measured directly in our experiment:

$$\Delta \langle \alpha_E \rangle = \langle \alpha_E \rangle - \langle \alpha_o \rangle = \frac{1}{l} \cdot \ln \left( \frac{\langle u_o \rangle}{\langle u_E \rangle} \right).$$

With the values $l = 24.8 \text{ cm}$, $\Delta l = 0.01 \text{ cm}$, $\Delta u = 1 \text{ mV}$, $\langle u_o \rangle = 1 \text{ V}$ and $\langle u_E \rangle = 2 \text{ V}$, we arrive at a relative error of $\Delta (\alpha_E)/(\Delta \alpha_o) = 0.2\%$.

RESULTS

The time-variations of the ultrasonic pulse on traversing the oil sample in the external field were plotted in the form of $\Delta \alpha_E$ versus $E$ (Fig.2) as well as $\Delta \alpha_E$ versus the temperature $T$ (Fig.3). Moreover, we recorded the amplitude of the ultrasonic pulse on transition through the oil sample, pre-saturated with air bubbles from a compressor. In this case the growing pulse amplitude (Fig.4) achieved a steady-state level after a well defined lapse of time enabling us to determine the mean velocity of the bubbles drifting in the field of gravitation.

DISCUSSION OF THE RESULTS

The presence of great numbers of bubbles in the oil leads to very strong damping of the ultrasonic wave (chiefly due to scattering). In practice, the signal vanished.

Only on disconnecting the compressor did the signal amplitude begin to rise slowly as the oil underwent degassing. Degassing of an oil of density $\rho_o$ proceeds through the action of the Archimedes force on the air bubble. With regard to the Stokes force which, at uniform motion of the bubble, is equal to the Archimedes force, the effective radius $R_{ef}$ of the bubble is found to be

$$R_{ef} = \sqrt{ \frac{9 \mu \eta_o}{2 g \rho_o} } .$$
Fig. 1. The changes in ultrasonic pulse amplitude (6 MHz) previous and subsequent to the application of an external electric field \((E = 154 \text{ kV/m})\) in mineral condenser oil.

\[
\text{amplitude} \atop \text{time} \atop \text{E = 154 kV/m}
\]

\[
\begin{align*}
\text{mineral condenser oil}, & \quad f = 6 \text{ MHz} \\
\text{t = 40 °C} & \quad \text{I = 24.8 cm} \\
\sigma_{o,\text{exp}} = 0.1935 \text{ cm}^{-1} & \quad \Delta \alpha_g = -0.015 \text{ cm}^{-1}
\end{align*}
\]

Fig. 2. Variations in amplitude absorption coefficient \(\Delta \alpha_g\) of the ultrasonic wave in the oils \textit{versus} the external electric field strength \(E\).

\[
\text{[cm]} \atop \text{electric field strength} \atop \text{E = 154 kV/m}
\]

\[
\Delta \alpha_g
\]

\[
\begin{align*}
\text{AKB 10°C} & \quad \text{O MCO 10°C} \\
\text{AKB 40°C} & \quad \text{* MCO 40°C}
\end{align*}
\]

Fig. 3. The same as in Fig. 2. \textit{versus} the temperature of the oils.

\[
\text{temperature °C} \atop \text{electric field strength} \atop \text{E = 154 kV/m}
\]

\[
\Delta \alpha_g
\]

\[
\begin{align*}
\text{AKB 77 kV/m} & \quad \text{O MCO 77 kV/m} \\
\text{AKB 231 kV/m} & \quad \text{O MCO 231 kV/m}
\end{align*}
\]

Fig. 4. Time-variations of the ultrasonic pulse amplitude (6 MHz, 30°C), on traversal of a given path in the mineral oil after the production therein of a cloud of air bubbles.
with \( v \) - the mean velocity of the drifting bubbles determined from the recorded variation in ultrasonic pulse amplitude in the course of degassing, and \( g \) is gravitational acceleration.

With the values occurring in the preceding example, \( \rho_0 = 854.23 \text{ kg/m}^3 \), \( v = 294 \text{ \mu m/s} \), \( \eta = 13.33 \times 10^{-3} \text{ Ns/m}^2 \), we obtained for the effective radius of the bubble \( R_f = 45.8 \text{ \mu m} \).

If incomplete discharge takes place in the bubble oxygen \( O_2 \) goes over into ozone \( O_3 \).

Since ozone combines with the surrounding oil (especially at higher temperatures) the bubbles decrease in size [3]. Consequently, the ultrasonic amplitude absorption coefficient decreases. This process may well be responsible for the characteristic shape of the \( \Delta \alpha_{(E, T)} \) graphs of Figs 3 and 4 obtained experimentally. In both cases the ultrasonic absorption specifically decreases at higher temperatures, at which diffusion of the gas out of the bubbles proceeds more intensely. The electric field distribution in the vessel exhibits some divergences from homogeneity, especially at the edges of the vessel, where the density of the electric field lines is much greater than in the central region.

When the field is switched on, the bubbles are pushed out of the region where the electric field strength is greater toward the centre of the vessel leading to a higher concentration of the bubbles and thus to higher ultrasonic wave absorption.

In addition to the electrostatic interaction described above, we have to deal with a thermal interaction involving diffusion. The higher the temperature, the greater is the rate of diffusion. The velocity of the bubbles drifting toward the region where the electric field is weaker is limited by the viscosity of the oil.

At higher temperatures, diffusion considerably counteracts the ordering influence of the electric field on the bubbles. At a given temperature, the two factors cancel out mutually: the external electric field ceases to affect the amplitude coefficient of ultrasonic wave absorption. In the present case, this temperature lies in the range of 20 to 30°C.

Thus ultrasonic investigation permits the measurement of the effective size of gas bubbles and the gas content in oils. The method should be of considerable interest with regard to applications in the Oil Industry.

CONCLUSIONS

1. The theoretical evaluation of the electric distribution about a gas bubble immersed in a liquid dielectric points to the existence of considerable inhomogeneities of the field thus making degassing easier.
2. Ultrasonic measurements in liquids gas-containing medium permit the determination of the effective radius of the gas bubbles.
3. The decrease in amplitude absorption coefficient of the ultrasonic wave observed at higher temperatures points to a decrease in concentration of the bubbles or a decrease in their effective size.
4. Ultrasonic measurements are found to permit the determination of the effective size of bubbles and the degree of their concentration when testing various oils, thus providing a method meeting a vast range of practical aims.

This work was performed within the framework of KBN.

REFERENCES

EXTRACTION OF PRECISE FUNDAMENTAL FREQUENCY
BASED ON HARMONIC STRUCTURE OF SPEECH

Jouji Suzuki, Miyuki Setoh, Tetsuya Shimamura
Dept. Information and Computer Sciences, Saitama University, Shimo-okubo
225, Urawa, Saitama 338, Japan

SUMMARY

Correct and precise measurement of fundamental frequency $F_0$ is requested in the detailed analysis of speech sound and analysis synthesis telephony employing $F_0$. This article proposes a new method to estimate $F_0$ utilizing the harmonic number and its frequency. If the n-th harmonic frequency is identified from the spectrum, the fundamental frequency is computed as the weighted mean of several harmonic frequencies. Experimental results indicate that the proposed method gives more precise and reasonable pitch information.

1. Introduction

Measurement of $F_0$, fundamental frequency of voiced speech, is essential at phonetics, linguistics, analysis of pathological voice, synthesis of speech, analysis synthesis telephony and so on. Many methods have been proposed and employed to measure $F_0$ or fundamental period $T_0$. We have not obtained, however, the distinguished method on robustness or accuracy of extraction.

When the voiced sound is analyzed by BPFs, harmonic structure of voice source is observed as peaks on the spectrum. Fig.1 shows the spectrum analyzed by FFT. Harmonic number is indicated in circles.

Many pitch extraction methods have been proposed [1-3] by utilizing this feature. Authors propose a new method obtaining more precise $F_0$ information by employing harmonics.

2. Principle and procedures

If the harmonic order of peak $n$ is identified and the peak frequency $P_n$ is obtained, $F_0$ is easily computed by $P_n/n$. However, it is impossible to measure $P_n$ directly.

When the voiced sound is analyzed with narrow band BPFs, harmonic structure is observed as shown in Fig.1. The center frequency of BPF indicating the n-th peak is assigned as $f_n$. $F_0$ is estimated by $f_n/n$. Fig.2 shows the linear relation between $f_n$ and $n$. $F_0^*$, a speculated $F_0$, is computed by $f_n/n$. The expected error of $F_0^*$ is less than $B/2$. $B$ is the
Fig. 3 Procedure for determination of fundamental frequency by use of harmonic structure.

Fig. 2 Relation between harmonic number and channel number of FFT (1024 pts.)

bandwidth of the BPF. The larger the n and the narrower the B, the more precise Fo is obtained. Utilizing these features, a new algorithm to speculate Fo is proposed. It is shown in Fig. 3.

Speech signal of 256 points are segmented and weighted with Hanning Window after V/CV classification. Frequency spectrum is calculated with FFT of 1024 points after adding "0" to segmented data. When sampling frequency of 10 kHz, B becomes 9.8 Hz. Determination of n-th peak is carried out on the log spectrum. It is very difficult to identify peaks corresponding to harmonics on such detailed spectrum because of existence of many local peaks. Moving average with 5 points is carried out for smoothing the spectrum. This procedures brings the easier detection of peaks without increasing the analyzing bandwidth and changing fn. Fig. 4 shows the original spectrum and the spectrum after smoothing.

fn and n are determined below 2 kHz by referring to interval of peaks. fn is influenced with zero of glottal waveform, transfer function of vocal tract, and noise. In order to get reliable Fo, it is effective to employ many information on harmonics. Now, Fo' is calculated as follows:

\[ F_{o'} = \frac{\sum_{n=1}^{N} f_n}{\sum_{n=1}^{N} n} \]

N is selected in the interesting frequency range. This equation utilizes that (1) Error in Fo' is proportional to 1/n, or (2) Fo' is obtained by the approximation with a linear function of n and fn. This calculation is equivalent to get the coefficient of an autoregressive line passing on the origin. The expected error is estimated less than B/N.

Fig. 4 FFT spectrum of 1024 points and its smoothed spectrum by five points of moving average.
Fig. 5 Results of measurement of fundamental frequency.
Female voice; "byobue to nyoraizoo ni taisuru."
Upper; wave form, Middle; proposed method, Lower; cepstrum

3. Results

Fig. 5 is an experimental result on female voice, and a portion of this figure is enlarged and shown in Fig. 6. In these figures, the fundamental frequency extracted by the proposed is shifted by 20 Hz. Furthermore, Fo extracted by cepstrum and the interpolated are drawn as references. Interpolation only gives more smooth variation than the original.

Fig. 7 shows Fo' by the proposed and fundamental frequency extracted by cepstrum method at a sustained vowel. It is clearly understood that the proposed method gives more precise and detailed information of Fo.

Table shows extraction error of the proposed,
Table Comparison of fundamental frequency extracted by three method. (errors/total in %)

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Proposed</th>
<th>AutoCor</th>
<th>Cepstrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male 1</td>
<td>1/24(16.6)</td>
<td>8/(5.5)</td>
<td>29/(20.0)</td>
</tr>
<tr>
<td>Male 2</td>
<td>8/282(2.8)</td>
<td>6/(2.1)</td>
<td>7/(2.5)</td>
</tr>
<tr>
<td>Male 3</td>
<td>10/136(7.4)</td>
<td>9/(6.6)</td>
<td>10/(7.4)</td>
</tr>
<tr>
<td>Mean</td>
<td>42/563(7.5)</td>
<td>23/(4.1)</td>
<td>46/(8.2)</td>
</tr>
<tr>
<td>Female 1</td>
<td>16/197(8.1)</td>
<td>2/(1.0)</td>
<td>3/(1.5)</td>
</tr>
<tr>
<td>Female 2</td>
<td>7/96(7.3)</td>
<td>2/(2.1)</td>
<td>6/(6.3)</td>
</tr>
<tr>
<td>Female 3</td>
<td>4/109(3.7)</td>
<td>3/(2.8)</td>
<td>2/(1.8)</td>
</tr>
<tr>
<td>Mean</td>
<td>27/402(6.7)</td>
<td>7/(1.7)</td>
<td>11/(2.7)</td>
</tr>
</tbody>
</table>

confirm its effectiveness.

ACKNOWLEDGEMENT

This work is owed to Ms. Eri Sano for her original experiment and to Prof. Hiroyuki Yashima for his fruitful discussions.

REFERENCES


4. Conclusive remarks

A new method for measurement of fundamental frequency of voiced speech is proposed by use of harmonic structure. Experimental results indicate that the proposed will be powerful tool for obtaining precise and reasonable fundamental frequency even if it is high. It is considered that this method contribute to analysis of singing voice, pathological voice, naturalness or individuality of speech, and establishment of analysis synthesis system and speech enhancement system.

Further experiments, however, are needed on more various speech including noisy speech to...
APPROPRIATE GROUPING OF PHONEME BASED ON
DISTINCTIVE FEATURES FOR EVALUATING THE
PERFORMANCE OF HEARING AIDS

Yōiti Suzuki, Futoshi Asano, Shoei Sato*, Toshio Sone
Research Institute of Electrical Comm., Tohoku University, Sendai, Japan
* Japan Broadcasting Corporation, Sendai Station, Sendai

SUMMARY In order to group phonemes to obtain stable statistics from results of a reasonable number of hearing tests, use of distinctive features with appropriate modification is proposed in this paper. The modification is done so that confusion between phonemes can be regarded as being equivalent to the confusion between two groups of phonemes divided according to modified distinctive features.

INTRODUCTION

Careful examination of confusion matrices of phonemes yields information on detailed aspects of hearing impairment. With this method, however, extensive testing is needed to obtain statistically reliable results. If an evaluation is based not on individual phonemes but rather on grouped phonemes, the number of samples from a single test becomes larger, and thus, we can obtain more reliable statistics.

It is important that confusion between phonemes should be regarded as being equivalent to the confusion between two groups of phonemes discriminated from each other based on, for example, their distinctive features. This is the basic requirement for any method of grouping phonemes used for evaluating speech perception.

STATISTICAL TEST WITH 2×2 CONFUSION MATRIX

Method Whether or not the requirement stated in the previous section is satisfied for the distinctive features shown in Table 1, which is the Japanese version of Chomsky-Halle’s system [1], is examined. To do this, two hearing tests were first conducted with 12 subjects with normal hearing acuity. The experiment consisted of a mono-syllabic and a nonsense tri-syllabic speech test. Speech was fed to the left ear through an earphone (YAMAHA, YHD-3) in an anechoic chamber.

Then, 2×2 confusion matrices, an example of which is shown in Table 2, were calculated from the experimental results to conduct a statistical test as follows: The upper line in Table 2 (A) shows an example of the frequency distribution of /b/ being perceived as a +Voiced phoneme and as a −Voiced phoneme. The lower line (B) shows the frequency distribution of +Voiced phonemes other than /b/ being perceived as +Voiced and −Voiced phonemes, respectively. A null hypothesis, $H_0$, that the two distributions are samples from
the same population was examined. Since the frequencies are often below 5, Fischer’s exact method was used.

**Results** From the results of the statistical tests, the rejection ratio, \( R_{rej} \), was calculated. This ratio is defined as the ratio of the number of subjects by whom \( H_0 \) was rejected beyond a significance level of 0.01 for each phoneme and a specific distinctive feature to the total number of subjects.

The central column of Table 4 shows the \( R_{rej} \) averaged over all the phonemes for each feature. Figure 1 illustrates \( R_{rej} \) for each phoneme for Obstruent for example. The ordinate exhibits \( R_{rej} \) and the abscissa shows phonemes and the values, + or −, of the distinctive features of the phonemes. As shown in Table 4 and Fig. 1, \( R_{rej} \) is not always very small. When a phoneme exhibits large \( R_{rej} \) for a distinctive feature, the confusion of the phoneme cannot be regarded as having the same tendency as those of the other phonemes for this feature.

**MODIFICATION OF DISTINCTIVE FEATURES FOR THE EVALUATION OF PHONEME CONFUSION**

The results of the statistical test mean that the distinctive features shown in Table 1 cannot be used as is for the purpose of grouping the phonemes to evaluate confusion of phoneme perception. Therefore, modifications of the distinctive features were performed to obtain a set of features suitable for the grouping of phonemes [2].

An example of the results of these modifications is shown in Figure 2. This figure illustrates \( R_{rej} \) for each phoneme in Consonantal, distinctive features of which were modified as follows: /h/ is disregarded since acoustic characteristics of /h/ are strongly dependent on the following vowel in the Japanese language. Comparing Figure 2 with Figure 1, it can be seen that \( R_{rej} \) is dramatically reduced by the modification.

This kind of modification was examined for all the features. Continuant and Coronal, however, could not be treated in a unified way. Strident, one of the distinctive features proposed by Jakobson [3], was introduced for the grouping of /s/ and /z/ instead of Continuant, the modification of which did not yield a good grouping. Moreover, various modifications were attempted for Coronal, but none of them yielded better results. Thus, the feature was used as it appears in Table 1, though \( R_{rej} \) is not sufficiently small.

As a result of the modification, the new set of features shown in Table 3 was determined. In the table, # denotes that the phonemes are disregarded for the grouping of the features. The right column of Table 4 shows the \( R_{rej} \) averaged over all the phonemes for each modified feature. Comparing the right column with the central column, it can be seen that the modification of the features decreased the value of \( R_{rej} \) as a whole.

**APPLICATION TO HEARING IMPAIRED LISTENERS**

To check the validity of using the modified distinctive features for the grouping of phonemes, results of a hearing test with 42 sensorineural hearing-impaired subjects were statistically tested. Each subject listened to 50 monosyllables in CD:TY-89[4] for each test condition.
Table 1: Distinctive features of Japanese phonemes. 1(−) indicates presence (−absence) of the feature.

| Features | p | b | t | c | d | k | g | s | z | m | n | r | w | y | h |
|----------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Consonantal | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| Obstruent | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| Voiced | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| Continuant | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − |− |
| Nasal | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − |− |
| Coronal | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − |− |
| Anterior | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − |− |
| High | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − |− |
| Back | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − |− |

Table 4: Total rejection ratios with significant level of 1%.

<table>
<thead>
<tr>
<th>Feature</th>
<th>original</th>
<th>proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consonantal</td>
<td>0.173</td>
<td>0.066</td>
</tr>
<tr>
<td>Obstruent</td>
<td>0.185</td>
<td>0.009</td>
</tr>
<tr>
<td>Voiced</td>
<td>0.028</td>
<td>Same as left.</td>
</tr>
<tr>
<td>Strident</td>
<td>0.288</td>
<td>0.018</td>
</tr>
<tr>
<td>Nasal</td>
<td>0.028</td>
<td>Same as left.</td>
</tr>
<tr>
<td>Coronal</td>
<td>0.111</td>
<td>Same as left.</td>
</tr>
<tr>
<td>Anterior</td>
<td>0.119</td>
<td>0.018</td>
</tr>
<tr>
<td>Back</td>
<td>0.065</td>
<td>0.018</td>
</tr>
</tbody>
</table>

Table 2: Example of 2x2 confusion matrix used in the likelihood ratio test. (A: [Voiced] confusion frequency distribution of stimulus /b/. B: [Voiced] confusion frequency distribution of [VVoiced] stimuli other than /b/.)

<table>
<thead>
<tr>
<th></th>
<th>+Voiced</th>
<th>−Voiced</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:</td>
<td>161</td>
<td>8</td>
<td>169</td>
</tr>
<tr>
<td>B:</td>
<td>1249</td>
<td>13</td>
<td>1262</td>
</tr>
<tr>
<td>Total</td>
<td>1410</td>
<td>21</td>
<td>1431</td>
</tr>
</tbody>
</table>

Figure 1: Results of likelihood ratio tests. (Rejection rate: $R_{rej}$)

Table 3: Proposed phonemic grouping system. +(/−) indicates the presence (−/absence) of the feature, while # indicates phonemes excluded in this phoneme grouping system.

| Features | p | b | t | c | d | k | g | s | z | m | n | r | w | y | h |
|----------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Consonantal | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| Obstruent | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| Voiced | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| Strident | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| Nasal | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| Coronal | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| Anterior | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| Back | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |

Figure 2: Results of likelihood ratio tests using proposed phoneme grouping system. (Rejection rate: $R_{rej}$)

-167-
The following three conditions were applied: 1) no amplification, 2) linear amplification (half-gain rule), and 3) amplification with our Digital Hearing Aid (CLAIDHA) [5].

The statistical test was the same as that described in the previous section. The statistical test showed that rejection occurred only at a rate of 0.30% on the average and at a rate of 5.14% at the most (/r/ in Nasal). These results mean that the proposed method of grouping phonemes based on the new set of distinctive features shown in Table 3 can be used for hearing impaired subjects to analyze the results of hearing tests.

The grouping method proposed here is being used to evaluate the performance of our digital hearing aid (CLAIDHA IV) [6]. From $2 \times 2$ confusion matrices for all the features, information transmitted [7] and entropies are calculated and analyzed along with simple percent-correct values and precise confusion matrices of all the phonemes [2].

CONCLUSION

In order to group phonemes to obtain stable statistics from results of a reasonable number of hearing tests, use of distinctive features with appropriate modification has been proposed in this paper. After grouping the results of hearing tests based on the modified distinctive features, inclinations of phoneme confusion were represented by confusion between the groups of the phonemes discriminated from each other by one of the features.

Acknowledgments

The authors wish to thank former Prof. Nagatomo and Ms. Ogawa for their help in using the listening test materials and Prof. Makino for intensive discussion. The authors also thank Prof. Takasaka, Prof. Kobayashi, Dr. Ohyama, Dr. Kakehata, and Mr. Komine for their help in conducting the hearing test. This study was supported by a Grant-in-Aid for Developmental Scientific Research from the Ministry of Education, Science and Culture, Japan (B-03557072)

References

SPEECH FEATURE EXTRACTION USING A DSP SYSTEM FOR TACTILE DISPLAY

Yuichi Ueda and Akira Watanabe
Faculty of Engineering, Kumamoto University, 2-39-1 Kurokami, Kumamoto 860, Japan

ABSTRACT: Tactile representations have been used in transmitting speech information to the hearing impaired. In devising such systems, it may be desired that the speech information has to be compressed possibly in the analytic stage and then converted into the tactile patterns fitted to human tactile sense in the synthetic one. In this paper, we described the speech feature extractions suitable for the DSP operations and a DSP system developed to implement those in real-time.

INTRODUCTION

In transmitting speech information to the profoundly hearing impaired, it is desired to make the most use of his/her residual senses, such as the senses of vision or touch. Especially, the tactile representations of prosodic features have been regarded to be effective as an aid of lipreading[1]. On the other hand, although those of segmental features have been tried in many studies[2], the effective methods have not been obtained yet, because of the mismatch between the amount of speech information and the tactile processing ability. Therefore, in devising a tactile system to transmit such segmental features, it should be taken account of not only the transformation into the tactile patterns fitted to the sense of touch in the synthetic stage but the efficient extractions and compressions of speech features in the analytic one.

In this paper, we describe a spectral normalization for a tactile vowel representation and some segmental feature extractions using spectral shapes, from the analytic standpoint. And a DSP system to implement those algorithms in real-time is introduced also.

DSP SYSTEM

Fig.1 shows a block diagram of the DSP system. DSP_1 implements a spectral extraction and transfers the spectra to a Dual-Port RAM. DSP_2 calculates some speech features using those in the DPRAM. The frame period of spectral analysis is 5.33ms, which corresponds to 64 data at a A/D sampling rate of 12kHz. Each frame is shifted every 2.66ms which equals to the D/A conversion period. DSP programs are downloaded from a PC via each serial interface, SI0.

The processing flow realized in the system is illustrated in Fig.2.

VOWEL FEATURE

Generally, in order to transmit the phonemic information of speech tactually, the spatial shapes or the temporal variations of spectra are transformed into the tactile patterns. In such tactile representations, there are cases where vowel phonemes of speech uttered by various speakers cannot be received correctly owing to the differences in their vocal tract lengths[3]. Since it is not expected that such spectral distortions can be decoded by human tactile sense, some spectral normalizations should be considered in the analytic stage.
A. SPECTRAL NORMALIZATION

The shapes of the 32ch logarithmic spectra based on the FFT operation in DSP_1 were normalized on both the frequency axis and the level one as follows.

\[ f' = f/k = f/(0.001211F_0 + 0.5154) \quad (1) \]

where \( f \) and \( f' \) are the original frequency scale and the converted one and \( F_0 \) indicates its fundamental frequency.

The differences in vocal tract lengths are observed mainly as the formant shifts on its frequency axis. On the other hand, those lengths are correlated relatively with \( F_0 \). Therefore, we converted the frequency scale so that those correlations might become to be lower. A coefficient \( k \) in eqs. (1) is such scaling factor which has been obtained empirically by use of \( F_0 \) and the three formant frequencies. The speech data were the 1080 Japanese five vowels uttered by 208 speakers who were males, females and infants of ages 6 to 22.

After interpolating the original spectra (ORG) according to eqs. (1), each level of \( S'(n) \) was normalized as follows.

\[ S^*(n) = (S'(n) - Min)/(Max - Min) \quad (2) \]

where \( Max \) and \( Min \) are the maximum and the minimum level of the \( S'(n) \). \( S^*(n) \) will be called the normalized spectra (NFS).

In Fig. 3, both spectral examples of vowel /a/ uttered by a male and a child are illustrated.

![Fig. 3 Examples of spectral analysis and those normalization. (W: Vowel Vectors)](image)

- 170 -
B. VOWEL VECTORS

In order to extract the vowel features from the NFS, the template spectra of Japanese five vowels were generated by averaging over the 170 NFS every vowel which had been uttered by 34 speakers of ages 6 to 22. And then, the spectral distances with each template were calculated as follows.

\[ d_i = \frac{1}{L} \sum_{n=1}^{L} |S^*_n(n) - S^*(n)| \quad i = 1, 2, ..., 5 \quad (3) \]

where \( S^*_n(n) \) and \( S^*(n) \) are the \( n_{th} \) component of NFS and one of the \( i_{th} \) vowel template respectively. \( L \) equals to 23, which was determined from the results of the preliminary discrimination test using the distances.

Moreover, in order to transform those into vibrotactile stimuli of 5ch, the following standardization was carried out.

\[ W_1 : W_2 : ... : W_5 = \frac{1}{d_1} : \frac{1}{d_2} : ... : \frac{1}{d_5} \quad (4) \]

\[ \sum_{i=1}^{5} W_i = 1 \quad (5) \]

where \( W_i \) corresponds to the \( i_{th} \) vowel feature and indicates the similarity with its vowel template. We call the \( W = (W_1, W_2, ..., W_5) \) a vowel vector. In Fig.3, the vowel vectors of each spectrum are shown also.

C. VOWEL IDENTIFICATION TEST

Fig.4 shows the results of vowel identification test using three kinds of vibrotactile patterns with AM signals of 200Hz. As illustrated in the figure, the original spectrum (ORG), the normalized one (NFS) and the vowel vector components (VEC) were transformed into 16ch- and 5ch- vibrator arrays respectively. The computer-extracted frame spectra of the above 170 isolated vowels were used in the experiments.

After training the template patterns in each mode, those tactile patterns were perceived by placing his finger tip on the vibrator array. The results indicated that the spectral normalization and its vowel vector representation might be effective in identifying Japanese five vowels tactually.

SEGMENTAL FEATURES

A. SPECTRAL APPROXIMATION

The vowel vector components should be modulated by some segmental features because those stimulate his/her skin at all times. Since the original 32ch spectra might include some features for the purpose roughly, those were approximated using the GRAM's orthogonal polynomials[4] as follows.

\[ S(n) = a_0 + a_1 \xi_1(n) + ... + a_k \xi_k(n) + \epsilon \quad (6) \]

\[ a_j = \sum_{n=1}^{32} S(n)\xi_j(n) \quad j = 1, 2, ..., k \quad (7) \]

where the \( \xi_j(n) \) is a GRAM's polynomial of the \( j_{th} \) degree, \( a_j \) is the \( j_{th} \) expansion coefficient, and \( \epsilon \) indicates an error. Since the eqs. (7) can be implemented simply by the multiply/accumulate operations, it may be favorable for DSP realization.

Fig.5 shows an example of the original 32ch spectrum and its approximate synthesized by \( a_j \) and \( \xi_j(n) \), \( j = 0, 1, 2, ..., 5 \).
B. SPECTRAL FEATURES

In Fig. 6, it notes that the coefficients of the lower degrees may represent the features of spectral shapes. So the discriminate-analysis was carried out using six coefficients, \( a_0, a_1, \ldots, a_6 \), concerning some groups of features, such as voiced, unvoiced, fricative and so on. CV syllables uttered by a male were used in the analysis, where each frame had been labelled one of the above groups by visual segmentation of waveforms.

The results are as follows.

- Both \( a_0 \) and \( a_1 \) were effective in V-U-S classification.
- The differences between vowel and buzz (or nasal murmur) frames were appeared on the space of \( a_2, a_4 \), and \( a_6 \).
- Some features could be distinguished from others on the 6-dimensional vector space.

C. OUTPUTS of SPEECH FEATURES

According to the above results, the vowel vectors were modulated by a signal (\( M_v \)) which was a voiced-unvoiced gating function weighted by \( a_0 \) so as to make the stimulus strength correspond with the speech energy. Moreover, the fricative (\( F_r \)) and the buzz (\( B_z \)) signals were calculated as the 6-dimensional inner-products of the coefficient vectors and the gravity vectors of each group respectively.

Examples of the real-time outputs of those are shown in Fig. 6. The ratios of five vowel vector components, \( W_1 \), are preserved although the absolute levels of those vary with the speech energy variations. \( F_r \) and \( B_z \) vary between 0 and 1. These should be used in driving the external tactile devices, such as the above 5ch vibrator array or the other vibro/electrotactile one. The DSPs have spent the calculation times of about 55% in DSP.1 and 20% in DSP.2 of the duration under real-time operation.

CONCLUSIONS

Some speech feature extraction methods for tactile display and the DSP realization of those have been introduced. It will be need to confirm those performances by an interface this system with tactile device. In future, some extractions of pitch, other features and so on, will be put in the DSPs also using the extra calculation time.

Fig. 6 Examples of real-time outputs of the vowel vectors and some features. (utterance by a male speaker)

REFERENCES

CHARACTERISTICS OF MISPRONUNCIATION AND HESITATION IN JAPANESE TONGUE TWISTER

Hidenori Usuki, Jouji Suzuki, Tetsuya Shimamura
Dept. of Information and Computer Sciences, Saitama University,
Shimo-okubo 255, Urawa 338, Saitama, Japan

SUMMARY

Frequently we produce mispronunciation and hesitation in phonation due to the difficulty in pronunciation. Characteristics of such troubles in phonation, however, have not been clearly understood. This paper analyzes characteristics in Japanese tongue twister statistically and the characteristics of mispronunciation and hesitation in phonation by employing questionnaire. Employed speech samples are Japanese tongue twister, because we consider that tongue twister has many factors of mispronunciation and hesitation. The results of questionnaire show that hesitation frequently occurs at vowels, particularly at vowel /a/ and /i/. Palatalized sounds are found many times in the part of mispronunciation and hesitation. These results will contribute to improve the quality of speech synthesized by rule.

INTRODUCTION

At present, synthesized speech is used in various fields. The naturalness of synthesized speech by rule, however, is inferior to that of synthesized speech by compilation of recorded speech sound. The largest problem of synthesis by compilation is the restriction of sentences that can be created. The synthesized speech by rule hasn't such restriction. The naturalness of synthesized speech by rule must be improved in order to use in more fields.

When we listen utterance, mispronunciation and hesitation are generally found many times. Mispronunciation and hesitation have not been carried out in the utterance of speech synthesis by rule. It is considered that the difference makes the naturalness of synthesized speech by rule worse. Synthetic rule changed with concatenation between phonemes contributes to improve the naturalness.

Characteristics of mispronunciation and hesitation, however, have not been clearly understood. At first, we had analyzed the sentences of Japanese tongue twister by employing phonemes, morae, manner and place of articulation, and distinctive features. The analyzed results show that palatalized sounds consonant /j/, and concatenation between /a/ and nasals or voiceless plosives are found many times in Japanese tongue twister. In this case, the movement of place of articulation is greater than that in usual speech [1].

Global characteristics of sentences of tongue twister have been clearly understood, however, where mispronunciation and hesitation occur and how to mispronounce in phonation has not been yet understood. This paper analyzes where mispronunciation and hesitation occur on sentences.

II. QUESTIONNAIRE

This paper analyzes the characteristics of mispronunciation and hesitation employed Japanese tongue twister, because it is considered that tongue twister have many factors of mispronunciation and hesitation. The questionnaire contains 254 sentences of Japanese tongue twister. It is presented to nine male and a female who speak Tokyo dialect. They are asked to indicate the part being occurred mispronunciation and hesitation, in the case of mispronunciation how mispronounce in each sentences. The procedure of questionnaire is as follows.

First, they remember each sentences written in a mixture of Chinese characters and Japanese syllabic characters. Next, they utter each sentences without reading the sentence. Because it is easy for us to pronounce by reading the sentences [2]. It is considered that reading makes difficult to understand the characteristics of mispronunciation and hesi-
tation. This paper analyzes the indicated part and whole sentences for comparing by employing phoneme, place and manner of articulation.

These sentences are sampled from booklet[3] and dictionary[4]. Definition of place and manner of articulation of phonemes is assigned after modification of [5], because choked sounds /Q/ and palatalized sounds are considered as unique classification. Description and definition of mora and phoneme refer to [6]. They are shown in table 1 and 2.

Table 1: Definition of place of articulation

<table>
<thead>
<tr>
<th>Place of articulation</th>
<th>Phoneme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lips</td>
<td>/p/, /b/, /m/, /N/, /f/</td>
</tr>
<tr>
<td>Teeth</td>
<td>/t/, /s/, /c/</td>
</tr>
<tr>
<td>Front vowels</td>
<td>/i/, /e/</td>
</tr>
<tr>
<td>Tip of tongue</td>
<td>/d/, /n/, /s/, /r/</td>
</tr>
<tr>
<td>Soft palate</td>
<td>/k/</td>
</tr>
<tr>
<td>Back of tongue</td>
<td>/g/</td>
</tr>
<tr>
<td>Back vowels</td>
<td>/a/, /o/, /u/</td>
</tr>
<tr>
<td>Glottis</td>
<td>/b/</td>
</tr>
<tr>
<td>Choked Sounds</td>
<td>/Q/</td>
</tr>
<tr>
<td>Palatalised sounds</td>
<td>/j/, /w/, /x/</td>
</tr>
</tbody>
</table>

X means /b/, /g/, /s/, /h/, /m/, /n/, /s/, /j/, /p/, /c/, /k/.

Table 2: Definition of manner of articulation

<table>
<thead>
<tr>
<th>Manner of articulation</th>
<th>Phoneme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced Plosives</td>
<td>/b/, /d/, /g/</td>
</tr>
<tr>
<td>Voiced Fricatives</td>
<td>/s/, /z/</td>
</tr>
<tr>
<td>Nasals</td>
<td>/m/, /n/, /N/</td>
</tr>
<tr>
<td>Liquid</td>
<td>/t/</td>
</tr>
<tr>
<td>Voiceless Plosives</td>
<td>/p/, /t/, /k/</td>
</tr>
<tr>
<td>Voiceless Fricatives</td>
<td>/s/, /h/, /l/</td>
</tr>
<tr>
<td>Voiceless Affricate</td>
<td>/c/</td>
</tr>
<tr>
<td>Vowels</td>
<td>/i/, /e/, /a/, /o/, /u/</td>
</tr>
<tr>
<td>Semi Vowels</td>
<td>/a/, /e/, /i/, /y/</td>
</tr>
<tr>
<td>Choked Sound</td>
<td>/Q/</td>
</tr>
</tbody>
</table>

III. RESULTS AND DISCUSSIONS

Table 3 shows that phoneme distribution of the part where hesitation and mispronunciation occur, mispronounced phonemes and whole sentences. Distribution of vowels in the part where hesitation occurs is different from that in the part where mispronunciation occurs, and from the distribution of whole sentences. The frequency of vowel /a/ in hesitation is remarkably greater than that in the part where mispronunciation occurs, and that of whole sentences. On the other hand, vowel /a/ is appeared few times in the part where mispronunciation occurs. Vowel /a/ is uttered by opening the lips wide. It is considered that vowel /a/ requires us larger momentum than other phonemes in pronunciation. If vowel /a/ appears again and again in tongue twister, we need more efforts for utterance. It is considered that vowel /a/ is a factor of hesitation rather than a factor of mispronunciation.

Table 3: Distribution of mispronunciation, hesitation and whole sentences (%)

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Hesitation</th>
<th>Mispronunciation</th>
<th>Whole Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>8,0</td>
<td>3,2</td>
<td>2,7</td>
</tr>
<tr>
<td>e</td>
<td>4,1</td>
<td>1,2</td>
<td>1,5</td>
</tr>
<tr>
<td>a</td>
<td>20,3</td>
<td>2,2</td>
<td>3,5</td>
</tr>
<tr>
<td>o</td>
<td>13,6</td>
<td>2,9</td>
<td>2,0</td>
</tr>
<tr>
<td>u</td>
<td>6,8</td>
<td>2,8</td>
<td>3,1</td>
</tr>
<tr>
<td>i</td>
<td>0,5</td>
<td>0,8</td>
<td>1,1</td>
</tr>
<tr>
<td>u</td>
<td>0,1</td>
<td>0,8</td>
<td>1,1</td>
</tr>
<tr>
<td>b</td>
<td>1,3</td>
<td>4,5</td>
<td>4,9</td>
</tr>
<tr>
<td>d</td>
<td>0,6</td>
<td>0,9</td>
<td>1,0</td>
</tr>
<tr>
<td>g</td>
<td>4,3</td>
<td>12,6</td>
<td>9,2</td>
</tr>
<tr>
<td>s</td>
<td>2,4</td>
<td>7,3</td>
<td>7,1</td>
</tr>
<tr>
<td>h</td>
<td>0,7</td>
<td>1,4</td>
<td>1,3</td>
</tr>
<tr>
<td>j</td>
<td>0,7</td>
<td>0,5</td>
<td>0,1</td>
</tr>
<tr>
<td>m</td>
<td>8,3</td>
<td>9,9</td>
<td>14,3</td>
</tr>
<tr>
<td>n</td>
<td>4,2</td>
<td>5,1</td>
<td>4,4</td>
</tr>
<tr>
<td>N</td>
<td>0,3</td>
<td>0,0</td>
<td>0,2</td>
</tr>
<tr>
<td>z</td>
<td>0,4</td>
<td>0,8</td>
<td>1,5</td>
</tr>
<tr>
<td>r</td>
<td>1,6</td>
<td>2,3</td>
<td>2,5</td>
</tr>
<tr>
<td>p</td>
<td>1,9</td>
<td>2,3</td>
<td>4,8</td>
</tr>
<tr>
<td>t</td>
<td>4,4</td>
<td>6,2</td>
<td>5,1</td>
</tr>
<tr>
<td>c</td>
<td>1,2</td>
<td>1,1</td>
<td>0,8</td>
</tr>
<tr>
<td>k</td>
<td>9,9</td>
<td>12,1</td>
<td>10,3</td>
</tr>
<tr>
<td>Q</td>
<td>0,8</td>
<td>0,2</td>
<td>0,3</td>
</tr>
<tr>
<td>bj</td>
<td>0,1</td>
<td>0,7</td>
<td>0,4</td>
</tr>
<tr>
<td>gj</td>
<td>0,1</td>
<td>0,2</td>
<td>0,3</td>
</tr>
<tr>
<td>f</td>
<td>1,6</td>
<td>5,5</td>
<td>7,0</td>
</tr>
<tr>
<td>hj</td>
<td>0,6</td>
<td>0,6</td>
<td>0,1</td>
</tr>
<tr>
<td>mj</td>
<td>0,5</td>
<td>0,5</td>
<td>0,1</td>
</tr>
<tr>
<td>nj</td>
<td>0,3</td>
<td>1,8</td>
<td>1,5</td>
</tr>
<tr>
<td>rz</td>
<td>0,4</td>
<td>1,4</td>
<td>1,1</td>
</tr>
<tr>
<td>rz</td>
<td>0,4</td>
<td>1,3</td>
<td>0,1</td>
</tr>
<tr>
<td>cj</td>
<td>0,6</td>
<td>2,2</td>
<td>1,7</td>
</tr>
<tr>
<td>kj</td>
<td>1,4</td>
<td>3,1</td>
<td>4,4</td>
</tr>
</tbody>
</table>

The appearance rates of palatalized sounds in hesitation, the intended phonemes and whole sentences are 5.4%, 18.4% and 3.5% respectively. Palatalized sounds rarely appear in usual conversational speech but frequently appear in
Japanese tongue twister[1]. The appearance rates of palatalized sounds in hesitation and mispronunciation are larger than them in whole sentences. As to palatalized sounds in the intended phonemes in mispronunciation are three times as many as hesitated phonemes. It is considered that palatalized sounds is a factor of mispronunciation rather than a factor of hesitation.

Plosives are frequently found in distribution of the intended phonemes. In distribution of hesitated phonemes, plosives appear many times comparing to their in whole sentences. Plosives need the motion to stop air stream in vocal tract. It is considered that plosives are factor of hesitation and mispronunciation owing to such motion.

Table 4: Distribution of phoneme change in mispronunciation (%)

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td>m</td>
<td>7.6</td>
</tr>
<tr>
<td>s</td>
<td>f</td>
<td>5.9</td>
</tr>
<tr>
<td>m</td>
<td>g</td>
<td>4.7</td>
</tr>
<tr>
<td>/</td>
<td>s</td>
<td>4.2</td>
</tr>
<tr>
<td>k</td>
<td>kj</td>
<td>3.9</td>
</tr>
<tr>
<td>n</td>
<td>m</td>
<td>3.9</td>
</tr>
<tr>
<td>t</td>
<td>k</td>
<td>3.5</td>
</tr>
<tr>
<td>k</td>
<td>t</td>
<td>3.0</td>
</tr>
<tr>
<td>g</td>
<td>b</td>
<td>2.8</td>
</tr>
<tr>
<td>kj</td>
<td>k</td>
<td>2.5</td>
</tr>
<tr>
<td>b</td>
<td>g</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Distribution how we mispronounce phonemes is shown in Table 4. "From" is the intended phoneme to the mispronounced phoneme, "To" represents mispronounced phonemes. This table shows that few mispronounced phonemes are palatalized sounds, except phonemes // and /kj/. Palatalized sounds distribution in mispronounced phoneme is fewer than that in the intended phonemes, except phonemes // and /kj/, which is also shown in Table 3. As aforementioned, palatalized sounds are difficult to pronounce at tongue twister. It is considered that even if we mispronounce, we don't mispronounce from original phonemes to palatalized sounds, because we usually feel some difficulty to pronounce them. Palatalized sounds /kj/ and // appear frequently in usual speech, so we can easily pronounce them comparing to other palatalized sounds.

This table also shows that we mispronounce phonemes having similar features, such as voiced sounds to voiced sounds and plosives to plosives. It is considered that we mispronounce from original phonemes to phonemes that have the common feature with the intended and mispronounced phoneme. It is difficult to pronounce a chain of phonemes whose characteristics are slightly different each other rather than a chain of phonemes with different characteristics.

Table 5: Distribution of place of articulation change in mispronunciation (%)

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back Tongue</td>
<td>Lips</td>
<td>10.7</td>
</tr>
<tr>
<td>Teeth</td>
<td>Palatalized Sounds</td>
<td>7.3</td>
</tr>
<tr>
<td>Lips</td>
<td>Back Tongue</td>
<td>7.2</td>
</tr>
<tr>
<td>Palatalized Sound</td>
<td>Teeth</td>
<td>6.1</td>
</tr>
<tr>
<td>Tip of Tongue</td>
<td>Lips</td>
<td>4.4</td>
</tr>
<tr>
<td>Soft Palate</td>
<td>Palatalized Sounds</td>
<td>4.0</td>
</tr>
<tr>
<td>Teeth</td>
<td>Soft Palate</td>
<td>3.9</td>
</tr>
<tr>
<td>Palatalized Sounds</td>
<td>Tip of Tongue</td>
<td>3.8</td>
</tr>
<tr>
<td>Back Vowels</td>
<td>Back Vowels</td>
<td>3.8</td>
</tr>
<tr>
<td>Soft Palate</td>
<td>Teeth</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Table 5 shows that change in place of articulation in mispronunciation. Mispronunciation rather occurs between phonemes whose place of articulation are comparatively apart, as shown in this table. Japanese tongue twister have many concatenations that the movement of place of articulation is large[1]. Such concatenation cause mispronounce in Japanese tongue twister. Table 6 shows that change in manner of articulation in mispronunciation. Distribution of vowels and consonants are fifty-fifty in Japanese sentences including tongue twisters. Total percentage of mispronounced consonants is about 90%. Consonants have various manner of articulation comparing to vowels. It is considered that consonants are
easy to mispronounce comparing to vowels.

This analyzed result shows that we mispronounce hardly from vowels to consonants or from consonants to vowels. Change from voiced sounds to voiceless sounds or from voiceless sounds to voiced sounds hardly occur in Japanese tongue twister. It tends to continue the vibration of vocal cords when the tongue twister is pronounced.

We also analyze the relation of phoneme distance between original phoneme and mispronounced phoneme. Fig 1 shows that example of definition of phoneme distance. This analyzed results are shown in Table 7. Mispronounced phonemes frequently appear in the range from 2 to 4 phonemes. As aforementioned, we mispronounce at consonants. It is considered that we are easy to mispronounce when mispronounced phonemes appear in the range from 2 to 4 phonemes away from the intended phoneme.

**Table 7: Distribution of phoneme distance between original phoneme and mispronounced phoneme**

<table>
<thead>
<tr>
<th>Phoneme distance</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 phoneme</td>
<td>0.7</td>
</tr>
<tr>
<td>2 phonemes</td>
<td>45.0</td>
</tr>
<tr>
<td>3 phonemes</td>
<td>11.7</td>
</tr>
<tr>
<td>4 phonemes</td>
<td>13.4</td>
</tr>
<tr>
<td>5 phonemes</td>
<td>5.0</td>
</tr>
<tr>
<td>6 phonemes</td>
<td>4.5</td>
</tr>
<tr>
<td>over 6 phonemes</td>
<td>7.8</td>
</tr>
<tr>
<td>not appear</td>
<td>11.9</td>
</tr>
</tbody>
</table>

**Figure 1: Example of definition of phoneme distance**

**IV. CONCLUSION**

The characteristics in the part where mispronunciation and hesitation occur and how we mispronunciation are obtained through analysis of a questionnaire. Hesitation occurs at vowels, particularly at vowel /a/. This is considered as a remarkable feature of hesitation.

On the other hand, most of mispronunciation occurs at consonants. Majority of mispronounced phonemes become the preceding phonemes in each sentences. To mispronounce from original phoneme to phoneme whose place of articulation are comparatively apart is frequently found in Japanese tongue twister. In manner of articulation, we mispronounce from original phonemes to phonemes that have common feature with the intended phonemes. These are considered as features of mispronunciation. Mispronounced phonemes appear frequently in the range from 2 to 4 phonemes from the intended phoneme.

Physical analysis is carrying out when mispronunciation and hesitation occur. Furthermore, we will synthesize the sentences of tongue twister referring to these results, and estimate their naturalness.

**REFERENCES**


NONLINEAR FILTERING OF SPECTRAL FEATURES FOR ROBUST SPEECH RECOGNITION

Dirk Van Compernolle
Lernout and Hauspie Speech Products, NV, Koning Albert I - Laan 64, 1780 Wemmel, Belgium.

SUMMARY

In this paper we present some new ideas on how, by the application of nonlinear filtering to filterbank log energies, one can maximally reduce the disturbing influence of background noise and channel characteristics on the commonly used cepstral features for speech recognition. The non-linearities add following properties: variable time constant and transformation to fixed dynamic range.

INTRODUCTION

In most of today's speech recognition systems the feature representation is based on cepstral coefficients in one way or another. The most commonly used signal processing is Fourier spectral estimation, followed by filterbank simulation with filters equally spaced along some auditory scale (back, mel, ..) which is then followed by a cepstral transformation. These cepstral parameters combined with their derivatives over time are then presented as basic features to the statistical recognition modules. Features derived in such a way do not only model the speech signal itself but capture at the same time background noise and recording channel characteristics. The background noise is additive in the power spectral domain while the transfer function of the recording channel manifests itself as an additive component in the log spectrum or cepstral domain. The effect of these disturbances is very considerable, especially when there is a significant mismatch between training and testing environments.

Overcoming the mismatch problem can be done in a few different ways:

- by obtaining training material from different sources such that the expected variability is reflected in the statistical models
- by filtering out the disturbances from the features
- by adjusting the statistical parameters on line such that they represent the new acoustic environment.

The first solution is to be rejected because it reduces the discriminating power of the resulting statistical models. The second and third try to do the same thing but in a significantly different
way. Proponents of model adaptation will argue that changing the statistics has a greater potential in dealing with all kinds of nonlinear feature transformations. Proponents of feature filtering and transformation will argue that this additional signal processing step is significantly simpler. This paper deals with some new ideas on how such an optimal feature filtering might look like.

ENVIRONMENT NORMALIZATION

A Signal Capture Model The recorded signal contains apart from speech influence from the background noise and the recording channel. A reasonable model, covering a very wide range of situations, for speech recordings in adverse conditions is:

\[ y(t) = h(t) * s(t) + n(t) \]  

(1)

Thus the power spectral estimate for frame \( i \) is:

\[ Y_i(f) = H_i(f) \cdot S_i(f) + N_i(f) \]  

(2)

The above equation is valid for FFT power spectrum coefficients and approximately for filterbank values in general. From these cepstral coefficients \( c_y(k) \) are computed:

\[
\begin{align*}
  c_y(k) &= \text{FFT}^{-1}\log\{Y(f)\} \\
  c_y(k) &\approx \begin{cases} 
  c_h(k) + c_s(k) & H(f)S(f) \gg N(f) \\
  c_n(k) & S(f) = 0
  \end{cases}
\end{align*}
\]  

(3.a, b)

Thus during silence the cepstral coefficients represent the background noise and during sections with larger SNR the cepstral coefficients are a sum of contributions of the speech and the channel, while the background noise is masked by the speech.

The Ideal Compensator Under the assumption that both background noise and channel are stationary disturbances it is easy to write down the ideal compensator:

\[ \hat{S}(f) = (H(f))^{-1}(Y(f) - N(f)) \]  

(4)

Of course both \( H(f) \) and \( N(f) \) must be estimated on line and the estimate will never be perfect, nor will \( H \) and \( N \) ever be perfectly stationary.

The functionality of the ideal compensator is very reminiscent of the long term adaptation in the human auditory system. By intricate adaptation mechanisms the human ear is virtually insensitive to channel characteristics and the observation of stationary background noise gradually fades.

Histogram Based Estimation Methods Attempts to suppress background noise in speech enhancement and speech recognition applications have been abundant since the introduction of spectral subtraction in the late seventies [1]. It took much longer till a serious attempt was made to suppress the channel effects as well. The long-term adaptation in the early IBM Tangora [2] used histograms to measure long term maxima and minima after which the dynamic range was linearly rescaled. While the histograms reasonably characterize the environment, the linear rescaling is by no means a good fit to the ideal compensator presented above. Continuing on this work, a successful method that jointly suppressed background and channel effects using the signal capturing perspective was presented in [3]. The peaks in the histograms were used for the channel estimate and the noise floor was estimated as mean over the low energy frames. The
ideal compensator was further modified to cope better with estimation errors and to mask the residual noise typical in spectral subtraction like algorithms:

\[ \hat{s}(f) = (\hat{H}(f))^{-1}(Y(f) - \hat{N}(f)) + M \]  

(5)

in which \( M \) is a masking constant set to a reasonably small level such that the effective dynamic range in each channel is on the order of 25dB. The histogram estimation technique proved very reliable, but cumbersome. The greatest weakness of the method, however, is that creating reliable histograms implies time constants of several seconds. Thus the method, developed for dictation applications, could not easily be ported to the large class of telephone applications in which a whole session might just contain a few words and for which recognition performance is as important for the first word as for the latter ones.

**Cepstral Compensation Techniques** Commercialization of speech recognition technology, especially for telephone applications, made us realize that the channel distortion was often much more important than the background noise. For reasonably high SNRs it was shown that the channel characteristic is almost purely additive to the speech cepstral parameters. Adding to this the (verified) assumption that the long term cepstral mean of speech is quite constant leads to very simple algorithms. One can either subtract the long-term cepstral mean (CMS) or apply high pass filtering to the cepstral or log spectral features (RASTA). As might be expected, more complex variants, such as SNR dependent mean subtraction, are required to extend robustness to higher noise situations. Similarly the modification of plain RASTA filtering by J-RASTA has proven much more robust against changes in background noise.

**RECURSIVE NONLINEAR SPECTRAL FILTERING**

Cepstral mean subtraction and RASTA/J-RASTA filtering are implementations of a similar concept: a highpass linear filtering of the log spectral or cepstral features in which a number of parameters may be adapted in a feedforward fashion. By this we mean that the parameters are set on the basis of measurements on the input feature vectors: e.g. in adaptive J-RASTA the \( J \) may be set inversely proportional to the noise floor. The time constants used in CMS are typically quite a bit larger than in RASTA; which makes that the latter is quicker to adapt but also prone to accentuate coarticulation in the feature space, making phoneme recognition virtually impossible and alternatively implying some lack of swift adaptation for CMS.

While going in the right direction, there is still something missing in these algorithms. First of all there is some danger implied by the open loop steering, more important, however, these algorithms do not map both noise and loud speech to consistent values. The solution we propose to use the concept of the histogram method, i.e. trying to preserve a predictable fixed dynamic range after environment normalization and doing it in a simple adaptive filter fashion, but with variable learning rate and several hard nonlinear constraints. This can be achieved in the following way:

\[ y = \log(Y + A) \]  

(6.a)

\[ \hat{s} = y + b \]  

(6.b)

in which both \( A \) and \( b \) are values steered by output feedback from \( \hat{s} \). The general goal is to obtain a spectral estimate of \( s \) that is mapped to a fairly constant - and predictable - long-term dynamic range. The \( A \)-parameter takes a role as masker and must reduce the variability due to the noise. (It is somewhat reminiscent of the \( J \) in J-RASTA but it's steering criterion is actually the inverse.) The \( b \)-value takes the role of cepstral mean in cepstral mean subtraction.
The main distinction with the existing methods is the way in which all parameters are adapted. The final signal estimate is tracked constantly for minimum and maximum values. The adaptation is immediate if a new value outside the current estimates is observed and resettling is obtained by natural decay from both ends. In this way the "attack" time constant is identical to the frame shift and the "decay" time constants can be set independently for the noise and the speech. Furthermore the enforcement of a number of hard constraints is reasonably straightforward: minimal SNR must be guaranteed, and limitations on the range of A and b. The good insight in these hard constraints is a major advantage. It avoids such problems as CMS starting to model the average noise cepstrum if the speech detector has gone of track, or the problems posed by RASTA emphasizing the transitions between the unpredictable noise floor and the speech, making them also unpredictable transitions.

Experiments At this point evaluation is ongoing of the parameters steering the nonlinear adaptation constants and at variants that can be derived from the general principle. Preliminary experiments have been run using the described algorithm. These indicate a reduction in error rate of as much as 25% compared to more conventional environment normalization algorithms in the case of high mismatch between training and testing environment.

CONCLUSION

In this paper we presented a new algorithm for environment adaptation for robust speech recognition. The algorithm is based on nonlinear filtering of spectral features, using output feedback from the final spectral estimate to steer the whole adaptation process. Initial experiments show significant improvement in results for the high mismatch case.

ACKNOWLEDGEMENTS

Several of the ideas in this paper are results of discussion and work performed by Hugo van Hamme (L&H) and Tom Claes (KULeuven).

The author is also part time at the KULeuven and an honorary Senior Research Associate of the National Fund for Scientific Research of Belgium (N.F.W.O.).

References


PITCH SYNCHRONOUS LINEAR PREDICTIVE CZECH
AND SLOVAK TEXT-TO-SPEECH SYNTHESIS

Robert Vich
Institute of Radio Engineering and Electronics, Academy of Sciences of the
Czech Republic, Prague

SUMMARY

The contribution describes a new system for the conversion of written text into speech. The
system is based on the concatenation of suitably chosen elementary speech units representing
diphones and sound segments of natural speech. For the description of the speech units the linear
prediction analysis combined with vector quantization has been chosen. The system has low
memory requirements owing to the chosen inventory of speech units (only 441) as well as to
their parametric description enabling also an easy implementation of prosodic rules. The system
runs in real time with a TMS320C25 signal processor board.

INTRODUCTION

In cooperation with the Institute of Phonetics of the Philosophical Faculty, Charles University,
a text-to-speech (TTS) system has been proposed [1,2,3]. The system is based on the
concatenation of elementary speech units representing diphones and semidiphones of natural
male/female speech. For the description of the speech units linear predictive coding (LPC)
combined with vector quantization was used. The speech units are represented by a sequence of
LP coded models, the corresponding rms-values of the residual signal and the pitch frequency
changes within the unit. The application of linear prediction improves significantly the quality
of synthetic speech in comparison with the former Czech synthesizers that used formants or the
Baumwolspiner method.

The TTS system involves:
- linguistic preprocessing
- phonetical transcription of the text
- transformation of transcribed text into a sequence of speech units
- application of prosodic rules to the speech units
- synthesis of the speech signal.

The block scheme of the TTS system is shown in Fig. 1. The system runs in real time with the
sampling frequency 8 kHz on a PC XT/AT equipped with a signal processor board with
TMS320C25 [4].

SPEECH UNITS INVENTORY

The transcription of Czech texts comprises the formal transcription of some characters into
double characters and vice versa, the coarticulation at boundaries between words (with possible
glottal plosives for words with initial vowels), the softening of some consonants preceding i, i
(long i) or e (ye) inside a word, the change of voicing of paired consonants in final positions
(specially before a pause), and the regressive (in a few cases progressive) assimilation of
voicedness. The transcription of foreign or loan words is solved by tables of exceptions or by
transcription based on Czech pronunciation. The inventory of Czech transcribed sounds consists
of 6 vowels (further V) (a, e, i, ã, o, u) 2 diphthongs (au, ou) and of 29 consonants (further C)
and consonantal units (b, c, č, dz, dě, dż, d, f, g, h, ch, j, k, l, n, ň, p, r, ř, s, š, t, ť, v, z, ž, glottal plosive, pause).

Fig. 1 Text-to-speech system

The inventory of the TTS synthesis system consists of 441 speech units by which all sounds in Czech and Slovak can be synthesized. There are three types of units:
- diphones representing connection CV and VC
- central parts of vowels and diphthongs
- each consonant represented by two parts (initial/final) for the synthesis of CC connection.

For the construction of the inventory of the speech units about 300 Czech and Slovak words pronounced by a male/female speaker were used. The words were of approximately the same length and they contained all necessary sounds. The speaker pronounced them orthoepically with constant speed and loudness and with neutral accent and with as few further prosodic features as possible. From this material a sequence of models for each speech unit was found and included into the inventory by means of a programme with analog input/output, signal segmentation, analysis and synthesis and a number of auxiliary functions, e.g., manual correction of the fundamental frequency, a graphical display of smoothed spectra for a chosen interval and of spectral transitivity function based on cepstral distances. A further function - displaying the rate of the change of spectral envelopes - is of special importance for finding the stationary part in vowels. The sampling frequency is 8 kHz, the frame length is 24 ms with overlapping in analysis and 12 ms in synthesis. The speech signal is pre-emphasized with the coefficient 0.9, and Hamming-windowed. The prediction order is 8. The reflection coefficients gained by the analysis are transformed into the cepstral domain and then used for the calculation of spectral transitivity functions in the inventory construction and for the codebook generation and vector quantization of the speech units [5]. The information for each speech unit consists of a sequence of models (within the range of 0 to 24 models for one unit), each of them described by the vector of reflection coefficients, the root-means-square value of the residual signal and the increment of the fundamental frequency (the sum of increments within one speech unit equals zero).

From the LPC models a cepstral codebook of 256 models was constructed. Then the inventory was vector-quantized by this codebook. The intensity information given by the rms-value of the residual signal is logarithmically quantized to 64 values. The fundamental frequency increments are not affected by vector quantization.
PITCH SYNCHRONOUS LP SYNTHESIS

The synthesis part of the TTS system is shown in Fig. 2.

![Diagram of pitch synchronous LP synthesis]

The synthesis is carried out by a modified vocal tract model. For voiced excitation a periodical multipulse unit sample response of a broad-band Hilbert transformer is used, unvoiced segments are excited by coloured noise [6].

The naturalness of the synthetic speech is to great deal dependent on the implementation of the prosodic features of the text. For this reason, the TTS system controls the word and sentence prosody not only by changes of the fundamental frequency in voiced units, but also by changes of duration of diphones and by intensity variation. That means that also the segment length in the synthesis is time-varying [7]. The fundamental frequency changes requested by the prosody are linearly interpolated within the speech units and the microintonation is added. The TTS system enables the choice of the speech rate by changing the number of samples generated in the synthesis for the models from the inventory and also the fundamental frequency can be chosen for the male/female voice in a reasonable interval. In the application of the TTS system for the blind, an increase of the speech rate up to 300% of the basic rate (given by the speech rate in diphone labelling of natural speech) was required.

The changes of the fundamental frequency and of the duration of the diphones caused by the prosodic rules, together with the possible choice of the fundamental frequency level and of the speech rate, lead to consideration on synchronization of the speech synthesis with the actual fundamental frequency, in order to maintain the quality of the synthetic speech for different combination of these parameters. Pitch synchronous linear predictive (PSLP) synthesis has been proposed to minimize the occurrence of disturbing transients and other effects in the synthetic speech due to uncoordinated parameter updating which can be perceived as clicks or buzz. The algorithm of the PSLP synthesis generally sets the segment length in the synthesis according to the duration change given by prosody combined with required speech rate; the result is the number of samples that should be calculated with the LPC parameters for all models of the actual diphone. However, this length is not used directly. In voiced models the actual fundamental frequency is calculated from the required level of the fundamental frequency, the increments of the models for the actual diphone and from the change of the fundamental frequency required by the prosody. The segment length is then set to the nearest multiple of the combined fundamental frequency period, or the segment is dropped in case that a former
segment has already generated too many samples. In an unvoiced model following a voiced one, the number of samples is balanced, in the other unvoiced models the segment length is given by the speech rate and by the duration requirements given by the prosody. In real-time implementation of the PS LP synthesis carried out in the signal processor, some tricks must be applied, e.g. in situations for high speech rate where many successive segments must be dropped or when the "balancing" segment is very short, as there must be left some time for the communication of the diphone parameters to the signal processor board [4].

CONCLUSION

The intelligibility and naturalness produced in the PS LP TTS-system is good. Further improvement of the quality could be reached by extending the speech unit inventory, by using pole/zero modelling of the speech production and by defining better prosodies rules. Research is also performed to improve the quality of the female voice by more sophisticated excitation [8].

The TTS-system is in production in the Cooperative of the Blind (SPEKTRA) in Prague, mainly as a reading machine for the blind, but other application in information and telecommunication services are under development.

ACKNOWLEDGEMENT

The TTS-project in JREE was supported in 1991-1992 by the Grant of the Czechoslovak Academy of Sciences 26701. At present, the research is being supported by the Commission of European Communities in the frame of the PECO Programme - Action COST 233 "Prosodics of Synthetic Speech" under the Ref. No. ERBCIPCT 926101-12534.

REFERENCES

SPEECH VISUALIZATION BY EXTRACTING FEATURES WITH NEURAL NETWORKS

Akira Watanabe, Shigeki Tokunaga*, Shingo Tomishige
Faculty of Engineering, Kumamoto University
* Kyusyu Matsushita Electric Industrial Co., Ltd.

SUMMARY
A study of speech visualization has been continued to verify the existence of new visible speech by which we can communicate each other.

In this study, consonantal features have been extracted with four three-layered neural networks and overlapped as peculiar patterns to the features on the old representation of words.

Output magnitudes of the neural networks control intensity of the peculiar patterns to the respective feature and all of the patterns are overlapped as they are. This means that one of those patterns is not assigned to a consonant as a result of automatic recognition but appears naturally on the display. So that, we can usually see the clearest pattern of them. Sometimes, though we might notice two or more overlapped patterns, it can be expected that human decision using contextual information brings about good results.

In the experiments, after a subject learned only the Japanese meaningless words and monosyllables, he was tested reading meaning words. In spite that nobody taught the answers to the subject until the end of all tests, he could read correctly 96.3% of 75 words (300 words in total) pronounced by 4 speakers. The average response time was 1.33 s/word. These results give us a great hope to realize the excellent visible speech.

INTRODUCTION
Speech visualization aims at complete representation by which we can read any word or short sentence pronounced by anyone. Therefore, the basic idea is to naturally integrate several patterns corresponding to speech parameters, and never to recognize speech automatically. One of the key technologies in this area is formant extraction technique. We developed inverse filter control method to extract 4 - 5 formant frequencies and succeeded in representing voiced sounds by vivid colors decided from the formant frequencies. This technique has been applied on visualization of telephonic speech too.[1]

However, many consonantal features are too weak in the visual impression of the colors to understand for subjects. In this study, we tried to utilize neural networks to extract consonantal features. Neural networks are useful to extract several features by one processing network in the case where we cannot indicate clearly the processing algorithm. Signals concerning manner and place of articulation have been extracted from acoustical features with four three-layered neural networks and overlapped as visual patterns on the others. Finally, the new representation has been tested in reading patterns. The results made us have a hope that we might be able to realize true representations for visible speech near future.
NEW VERSION OF THE SPEECH VISUALIZATION SYSTEM

Old system

Previous system represents formant frequencies, pitch and normalized spectrum. The lowest three formant frequencies decide the color of voiced sounds and pitch frequency restricts the horizontal length of the color pattern. Continuous speech patterns flow from the bottom to the top on the screen. 2 seconds pattern always appears on the screen. The normalized spectrum is overlapped as a black-and-white pattern on the whole.

Tests to read the visual patterns showed that the rate of correct answers was 83% in 90 words pronounced by two adult males. [2]

Neural networks for new system

The new system is shown in Fig.1. In this study, consonantal information has been appended with neural networks. There are four three-layered networks which are called a manner of articulation net, a nasal (m/n/N) net, a plosive (b/d/g or p/t/k) net and a fricative (s/z/h) net. 42 acoustical parameters like 32 mel band filter outputs, four formant frequencies, pitch, zero crossing frequency etc., are used for input of the manner of articulation net. The output is 8 features that is fricative, burst, buzz-bar, nasal, vowel, voiced, voiced and silence. The other 3 networks have 50 inputs constructed by 42 acoustical parameters plus 8 outputs of manner of articulation net. The outputs are m/n/N, labial/dental/alveolar (p/t/k or b/d/g), s/z/h in nasal, plosive, fricative networks respectively.

Nasal and fricative networks have a structure of feed forward type because of stationary property in the acoustical features. On the other hand, recurrent net is used in the plosive networks because plosive sounds are specified by transitional feature. Since the acoustical features of plosives are influenced by the following vowel, five sub-networks which learned using transitional parts different in the following vowel are integrated by a main network.

Each network learned consonantal features from mono- syllables and /vcv/ uttered by 5 adult males. In order to investigate the performance, open tests have been conducted using mono- syllables and /vcv/ pronounced by the other five adult males than the speakers in learning. Average probabilities that the largest output is correct are shown in Table.1.

Since this system requires real-time processing, each frame of parameter changes is independently treated. So, the above results are not necessarily good comparing to
terns to the respective features and all of the patterns are overlapped as they are.

Fig.2 shows the system to synthesize patterns of the manner of articulation as an example. Each pattern which is assigned to each consonantal feature represents a visual image intuitively corresponding to auditory one. In Fig.2, weights \((W_1, W_8, W_n, W_v, W_u)\) mean output magnitudes of the neural networks and \(Q_1 - Q_5\) indicate binary patterns of consonantal image. \(R, G,\) and \(B\) are three primary colors which are decided by the lowest three formant frequencies.

Thus, as all of the signals are added each other, the strongest pattern will appear. Sometimes, subject might notice two or more overlapped patterns. It can be expected in such case that human decision using contextual information brings about correct answers.

Information of place of articulation for nasal and plosives also has been visualized in the same idea as the above. Manner of articulation is represented by a geometric pattern and place of it horizontal position of the pattern. According to the visual image of the new patterns, we feel that the new representations in which consonantal information is appended gives us strong impression and makes the whole pattern clear.

TEST FOR READING WORDS

In auditory perception, if basic words have been learned in the past, we can recognize any new word by applying the judgment ability. Therefore, to investigate the true effect of learning, the basic words to be used for training should differ from a group of words for test.

In the experiments to read the visual representation of words, we have used one to four syllables meaningless words for the training and two to four syllables meaning words for tests.

Training session

Meaningless words pronounced by two adult males were presented as visual patterns
with sounds. Three groups of 25 meaning words by the same speakers as in the meaningless words were prepared for preliminary tests. The meaningless words with sounds and first group of the meaning words without sound were presented alternatively and repeatedly to a subject (a male student) observing the rate of correct answers in the meaning words.

After the correct rate saturated, the second group of 25 words were appended and 50 meaning words were tested likewise. Finally, 75 words were tested. It took 11 days in this training session. The subject was not taught the answer of the meaning words at all during this training.

Test sessions
A new group of 75 meaning words were pronounced by four adult males including 2 speakers who were adopted in the training. This word group which consists of 300 words in total were consistently used for three test sessions. The results of the tests are shown in Fig.3. A period between first and second session is 5 months, and second and third is 7 months. During those periods, the subject was inactive in reading pattern. In the first and the second sessions, three types of neural networks that is, those for manner of articulation, plosives and nasals are used for the visualization. The fricative network was appended for the third session.

According to Fig.3, learning effect is remarkable. In spite of the long rests between the sessions, the correct rate and the average response times in the beginning of each session little degrade comparing the end of previous session. As a final result in this experiment, the subject can read correctly 96.9% - 96.5% of 300 words in a speed of 1.33 sec/word of the average response time. (The standard deviation of the response time is 0.8 s.) Since this response includes reaction time for pushing button after understanding, it is presumed that the judgment is instantaneous conducted. It should be asserted again that the correct answers were not taught at all to the subject until the end of all tests.

CONCLUSIONS
Speech processing by neural networks used in this study is not necessarily complete because of restriction to get information in real time. However, the obtained information is very useful for human judgment to the visual patterns.

We have been learning understanding speech through daily conversation since we were born. While, in this study, the learning period was only 11 days. Improving the neural networks and thorough training might prove the existence of complete visible speech near future.

References
ACOUSTIC NORMALIZATION AND ADAPTATION FOR MICROPHONE-CHANNEL CHARACTERISTICS

Yunxin Zhao*, Ted H. Applebaum, Brian A. Hanson

* Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL, USA
Speech Technology Laboratory, Panasonic Technologies Inc. Santa Barbara, CA, USA

SUMMARY

An unsupervised on-line adaptation technique for microphone/channel and speaker normalization is presented. Before decoding a sentence, a phone-independent spectral bias is estimated and removed from the distorted speech spectra. The spectral bias is initially estimated assuming the distorted spectra as uniformly distributed and then re-estimated assuming the normalized spectra as Gaussian distributed within each phone. After decoding the sentence with Gaussian-mixture phone models, the phone model parameters are adapted via Bayesian estimation, for use in recognizing the next sentence. An iterative batch adaptation procedure is further performed when appropriate. The experiments were on speaker-independent continuous speech recognition with a vocabulary size 853 and a grammar perplexity 105. The training data was from the TIMIT database, and the test data was acquired via three types of microphones. Recognition accuracy has been significantly improved using the proposed technique; e.g., on a test set containing 198 sentences from a male speaker where a mismatched microphone and microphone-mouth distance degraded the baseline accuracy down to 0.3%, with adaptation the average word accuracy increased to 85.2%.

1. INTRODUCTION

Of the various factors affecting automatic speech recognition, the types of microphones and their positions relative to speakers are important sources of variation in the speech signal, where mismatches between training and test in microphones and positions are known to severely degrade recognition accuracy. Recent attention has been given to three classes of techniques for dealing with microphone variations: cepstral filtering techniques, mapping of cepstral input features, and modifications to recognition model parameters. Comparisons of different techniques have been studied [Chang, 1994; Liu, 1994; Acero, 1993].

Cepstral filtering techniques, represented by cepstral mean normalization [Acero, 1993] and RASTA [Hermansky, 1991], yield significant recognition gains. As will be shown in this paper, the conventional cepstral mean normalization is equivalent to removing a bias obtained from maximum likelihood estimation assuming uniform distribution of the spectral features. Cepstral filtering techniques are often combined as a first step [Liu, 1994; Anastasakos, 1994; Neumeyer, 1994] with other, more computationally intensive, cepstral mapping techniques. Cepstral mapping techniques have been based on direct cepstral comparisons of stereo training data [Neumeyer, 1994; Liu, 1994; Anastasakos, 1994] or on a structural model of the acoustical degradation [Acero, 1993; Rahim, 1994]. An hierarchical approach that first removes a global cepstral bias and then modifies phone model parameters, based on short calibration speech from new speakers [Zhao, 1993b], has been shown effective for improving recognition accuracy under both matched and mismatched conditions.

There are disadvantages to approaches that require stereo training data, e.g., the need for a representative set of microphones and suitable interpolation algorithm [Liu, 1994] or microphone selection algorithms [Anastasakos, 1994; Neumeyer, 1994]. A spectral bias estimate can be obtained via an unsupervised maximum likelihood estimation [Cox, 1989]. The method described in this paper does not require stereo or adaptation data, since it is a self-learning procedure that combines the three classes of techniques described above, and the procedure can be performed on-line as new input speech is acquired. Self-learning is accomplished by two steps of spectral bias estimation, followed by unsupervised sequential phone model adaptation and iterative batch phone model adaptation [Zhao, 1993b; Zhao, 1993c]. The baseline speaker-independent continuous speech recognition system is based on phone-unit HMM: each phone unit has three states, and each state is modeled by a Gaussian mixture density [Zhao, 1993b]. Two other recent efforts were reported by [Gauvain, 1992] and [Paul, 1993] on unsupervised phone model adaptation for speaker adaptation.

2. ADAPTATION SYSTEM

The self-learning, adaptive speech recognition system is illustrated in Fig. 1. Assuming the speaker
Table 1. Recognition word accuracy using different adaptation methods

<table>
<thead>
<tr>
<th></th>
<th>baseline</th>
<th>acu0</th>
<th>PAU-1 (acu0)</th>
<th>PAS-1 (acu0)</th>
<th>acu1</th>
<th>acu2</th>
<th>PAU-I(acu2)</th>
<th>PAU-II(acu2)</th>
<th>PAS-I(acu2)</th>
<th>PAS-II(acu2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>86.9</td>
<td>88.1</td>
<td>88.3</td>
<td>88.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-CT</td>
<td>74.8</td>
<td>83.0</td>
<td>86.1</td>
<td>86.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-CT</td>
<td>56.5</td>
<td>77.6</td>
<td>85.2</td>
<td>85.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-1D</td>
<td>44.2</td>
<td>64.4</td>
<td>72.8</td>
<td>76.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-1D</td>
<td>0.3</td>
<td>54.5</td>
<td>71.3</td>
<td>78.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-OC</td>
<td>35.6</td>
<td>59.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-OC</td>
<td>17.6</td>
<td>62.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where \( \lambda^{(u_n)} \) is the interpolation parameter, \( \mu_k^{(p_u)} \) and \( \sigma_k^{(p_u)} \) are the sample mean and the sample covariance of the adaptation data. Define

\[
\lambda^{(u_n)} = \frac{1}{N} \sum_{i=1}^{N} \lambda_i^{(u_n)} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{2N} \lambda_i^{(u_n)} + \nu_i^{(u_n)} \right),
\]

\[
\mu_k^{(p_u)} = \frac{1}{N} \sum_{i=1}^{N} \mu_i^{(p_u)} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{2N} \mu_i^{(p_u)} \right),
\]

\[
\sigma_k^{(p_u)} = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^{(p_u)} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{2N} \sigma_i^{(p_u)} \right).
\]

2.3. Iterative Phone Model Adaptation

Denote the sequentially recorded sentences as the completion of one session as \( \{S_k, k = 1, 2, \ldots, K\} \) and the decoded word strings as \( \{ \mathcal{A}^{(u_n)}(t), k = 1, 2, \ldots, K\} \). In the sequential adaptation, the phone models that are used to recognize the \( j \)th sentence have been adapted from the previous sentences \( \{S_1, S_2, \ldots, S_{j-1}\} \) and the Viterbi segmentations have been guided by the word strings \( \{ \mathcal{A}^{(u_n)}(1), \mathcal{A}^{(u_n)}(2), \ldots, \mathcal{A}^{(u_n)}(j-1)\} \). In applications such as dictation, the recorded sentences could very well be reused to carry out iterations of phone model adaptation and recognition in order to further reduce the recognition error. Denoting the decoded word strings that are generated in the first iteration as \( A^{(1)}(k) \), and those in the \( n \)th iteration as \( A^{(n+1)}(k) \), for recognizing the \( j \)th sentence in the \( n \)th iteration, the phone models are adapted from all the available sentences excluding the \( j \)th sentence: \( \{S_1, \ldots, S_{j-1}, S_{j+1}, \ldots, S_K\} \) and the Viterbi segmentations are guided by the word strings \( \{ \mathcal{A}^{(n+1)}(1), \mathcal{A}^{(n+1)}(2), \ldots, \mathcal{A}^{(n+1)}(j-1)\} \), \( j = 1, 2, \ldots, K \). For example, when \( n = 1 \), the sentences that were uttered before the \( j \)th sentence are segmented according to the newly decoded word strings, and those after the \( j \)th sentence are segmented according to the word strings generated from the sequential adaptation.

3. EXPERIMENTS

The baseline speaker-independent HMM phone models were trained from 717 sentences by 325 speakers in the TIMIT database. The cepstral coefficients of the PLP analysis (8th order), log energy, and their first-order 50-msec temporal coefficients were taken as features. The task vocabulary size was 853, and the grammar perplexity was 105. The test set has one female (F) and one male (M), each reading 198 sentences, and the speech were simultaneously recorded using three microphones: a Sennheiser close-talking microphone (CT), a unidirectional dynamic microphone (UD), and an omnidirectional condenser microphone (OC). Compared to the TIMIT data, the test data were collected under a higher level of ambient noise (approximately 15 dB higher). In order to calibrate the mismatched conditions and to evaluate the adaptation effect on the matched condition, a TIMIT test set is also included with two to three sentences from each speaker giving a total of 186 sentences. The following cases were evaluated:

1. speaker-independent continuous speech recognition (baseline)
2. acoustic normalization assuming \( \hat{h}_3^{(u_n)} = 0 \) and using only the second bias-estimation block in Fig. 1 (acu0)
3. unsupervised phone model adaptation after acu0 (PAU-I (acu0))
4. supervised phone model adaptation after acu0 (PAU-I (acu0))
5. acoustic normalization using only \( \hat{h}_3^{(u_n)} \) (acu1)
6. acoustic normalization using \( \hat{h}_3^{(u_n)} \) (acu2)
7. unsupervised phone model adaptation after acu2 (PAU-II(acu2))
8. supervised phone model adaptation after acu2 (PAS-I(acu2))
9. iterative (one iteration) unsupervised phone model adaptation after acu2 (PAU-II(acu2))

The average recognition word accuracies are summarized in Table 1 for each case. On the data from the omnidirectional condenser microphone, the results of phone model adaptation are not available at the time of writing this paper. The word accuracy as a function of the amount of on-line speech data is further illustrated in Figure 2 for the male speaker using the unidirectional dynamic microphone. In Figure 2(a), the word accuracies were averaged over the number of sentences indicated on the horizontal axis (cumulative-averaged results), i.e., if \( n = 60 \), the result was averaged over the last 60 sentences. In Figure 2(b), the word accuracies were averaged over 20-sentence blocks (interval-averaged results), i.e., for \( n = 60 \), the result was averaged over the sentences 41 through 60. In the plots, the squares \( \square \) mark the baseline results; the diamonds \( \Diamond \) mark the results from the acu2; the upward triangles \( \Delta \) mark the results from the PAU-II(acu2); the downward triangles \( \nabla \) mark the results from the PAU-II(acu2) (one iteration); the circles \( \bigcirc \) mark the results from the PAS-II(acu2).

Acoustic Normalization. The acoustic normalization improved the recognition accuracy significantly under the mismatched conditions. Under the matched condition TIMIT (86.9%) and the good baseline condition
2.1 Spectral Bias Estimation and Removal

A standard speaker is characterized by a set of unimodal Gaussian densities \( N(\mu_i, C_i) \), \( i = 1, 2, \cdots, I \), which are estimated from a speaker-independent training set. For a speaker \( q \), the spectral sequence of a sentence is denoted as \( s_1, s_2, \cdots, s_T \), with the accompanying phone label sequence \( i_1, i_2, \cdots, i_T \). The spectral bias \( \hat{\kappa}(i) \) is estimated via the EM algorithm as

\[
\hat{\kappa}^{(1)}(i) = \arg\max_{\kappa(i)} \sum_{t=1}^{T} \log f(s_t | i = i_t, \kappa(i)) P(i_t = i) 
\]

with

\[
f(s_t | i = i_t, \kappa(i)) \sim N(\mu_i + \kappa(i) \epsilon_i, C_i), P(i_t = i) \sim \frac{N_i}{N} (1)
\]

\[
P(i_t = i | s_t, \kappa(i)) = \frac{f(s_t | i = i_t, \kappa(i)) P(i_t = i)}{\sum_{i'} f(s_t | i = i', \kappa(i')) P(i_t = i')}
\]

where \( N_i \) is the sample size of the \( i \)th phone unit, and \( N = \sum_i N_i \).

As the EM algorithm is only locally optimal, the choice of the initial bias \( \hat{\kappa}^{(0)}(i) \) has the direct bearing on the quality of the subsequent bias estimate. Under good conditions, the spectral bias is small and one could simply choose \( \hat{\kappa}^{(0)}(i) = 0 \). Under severely mismatched conditions, the spectral bias is likely to be large, so choosing \( \hat{\kappa}^{(0)}(i) = 0 \) could lead to bias estimate that is far from the true bias. In order to relieve the need for guessing \( \hat{\kappa}^{(0)}(i) \), the probability densities \( f(s_t | i = i_t, \hat{\kappa}(i)) \) are assumed uniform, leading to

\[
P(i_t = i | s_t, \hat{\kappa}(i)) = P(i_t = i) = \frac{N_i}{N}
\]

The estimate \( \hat{\kappa}^{(1)}(i) \) therefore becomes

\[
\hat{\kappa}^{(1)}(i) = \frac{1}{T} \sum_{t=1}^{T} s_t - \frac{1}{N} \sum_{i=1}^{N} N_i \mu_i = \bar{s} - \bar{\mu} (2)
\]

i.e., it is the difference between the mean spectra \( \bar{s} \) of the current sentence and the mean spectra \( \bar{\mu} \) of the entire training set. In Fig. 1, this bootstrap estimation of \( \hat{\kappa}^{(1)}(i) \) is shown as “bias estimation I.”

For \( n \geq 1 \), the density functions \( f(s_t | i = i_t, \hat{\kappa}^{(n)}(i)) \) are computed as the Gaussian densities in Eq.(1). Under the simplifying conditions that the posterior probabilities \( P(i_t = i | s_t, \hat{\kappa}^{(n)}(i)) \) are each approximated by the decision operation \( \hat{i}_t(i) = \arg\max_i P(i_t = i | s_t, \hat{\kappa}^{(n)}(i)) \) and the covariance matrices are the unit matrix, the spectral bias estimate becomes

\[
\hat{\kappa}^{(n+1)}(i) = \frac{1}{T} \sum_{t=1}^{T} (s_t - \mu_{\hat{i}_t(i)}), (3)
\]

where \( \mu_{\hat{i}_t(i)} \) is the mean of the phone unit labeled for \( s_t \) from computational considerations, only one iteration is carried out using Eq.(3), which yields the refined bias estimate \( \hat{\kappa}^{(n)}(i) \) shown as the box “bias estimation II” in Fig. 1. Acoustic normalization is defined as \( \bar{s}_t = s_t - \hat{\kappa}^{(n)}(i) \), \( \forall t \).

2.2 Phone Model Parameter Adaptation

A size-\( M \) Gaussian mixture density is defined as

\[
\sum_{k=1}^{M} \alpha_k s_t, \text{ with } f(s_t | \theta_k) \sim N(\mu_k, C_k); \text{ the parameters to be adapted are the means and the covariances } \theta_k = (\mu_k, C_k); \text{ the mixture weights satisfy the constraints } \alpha_k \geq 0 \text{ and } \sum_k \alpha_k = 1. \text{ The prior distributions of } \theta_k \text{'s are assumed to be i.i.d., and the mixture weights } \alpha_k \text{'s are taken as constant. The prior means and covariances, } \mu_0(0) \text{ and } C_0(0), \text{ are the speaker-independent estimates from a training set with sample size } n_k. \text{ The joint distribution of } (\mu_k, C_k^{-1}) \text{ is taken as a conjugate prior distribution. Based on the EM algorithm, the Bayesian estimation on the means and the covariances are}
\]

\[
\mu_k(n+1) = (1 - \lambda^{(n)}) \mu_k(n) + \lambda^{(n)} \mu_0(0(n))
\]

\[
\nu_k(n+1) = (1 - \lambda^{(n)}) \nu_0(0(n)) + \lambda^{(n)} (\nu_0(0(n)) - \mu_k(n)), (4)
\]
F-CT (74.6%), the set yield the best acoustic normalization result; for the rest of the cases, set1 yielded better or equivalent results compared to the set2, and the set2 improved set1 for all the cases. Especially noticeable is the case M-UD: the baseline accuracy was below 0% because the speaker held the microphone too close to his mouth which caused a significant boost to the low frequency components of the speech. In this case, the set2 is inferior to set1 because the posterior labeling probabilities based on the Gaussian models in Eq.(1) were very unreliable. On the other hand, it is observed that the set1 deteriorated the baseline result on the TIMIT set. Examining the results closely reveals that the method set1 increased errors on the sentences that have high correlations among spectral features; on these sentences the method set2 improved the decoding accuracy significantly.

Phone Model Adaptation The phone model adaptations further improved the recognition accuracy: the trend is that the longer the learning period, the larger the gain of accuracy. If the amount of adaptation data is very small, the unsupervised sequential phone model adaptation could degrade the decoding accuracy because the data is unreliable. The unsupervised adaptation yielded similar improvements as the supervised adaptation if the recognition accuracy after the acoustic normalization is high. The main effect of the iterative adaptation is on the sentences at the beginning of a session where the amount of adaptation data is too small for the sequential adaptation. The interval-averaged results of Fig. 2(b) indicate that the relative effects between the acoustic normalization and the phone model adaptation varied considerably among the subsets of sentences; the cumulative-averaged results of Fig. 2(a) indicate the trending effects for the individual methods.

4. CONCLUSIONS

The proposed adaptation technique has led to significant improvements for a speaker-independent continuous speech recognition system. The current work demonstrates that an accurate estimate of the distortion spectral bias is important for speech recognition under mismatched conditions. The proposed two-step method for spectral-bias estimation improves the conventional method of cepstral-mean subtraction and the previous unsupervised maximum likelihood estimation. The unsupervised sequential phone model adaptation improves the decoding accuracy with the increasing amount of speech data. Iterative phone model adaptation further improves recognition accuracy, which should be applied in applications where delaying the decoding output is permissible.

REFERENCES

PHYSIOLOGICAL AND PSYCHOLOGICAL ACOUSTICS
THE CANADIAN FORCES HEARING CONSERVATION PROGRAM: A THREE-YEAR PROSPECTIVE EVALUATION

S. M. Abel, E. O. Pelausa, J. Simard, I. Dempsey
Samuel Lunenfield research Inst., Mount Sinai Hospital, Toronto, Canada
and Div. of Otolaryngology, National Defence Medical Ctr., Ottawa, Canada.

SUMMARY

This study investigated the development of noise-induced-hearing loss in Canadian military recruits over the first three years of employment. The 134 subjects were employed in four trades, three associated with high noise levels. The data comprised hearing thresholds and responses to a questionnaire relating to noise exposure at work and during leisure. Group audiograms at the three-year recall were characterized by a 6 kHz notch, indicative of noise-induced hearing loss, although mean threshold values were within normal limits. Eleven percent of the infantry had a mild to moderate hearing loss in the left ear, exceeding 25 dB HL. Subjects appreciated the benefit of hearing protectors but instructions in their use and education on the hazards of noise exposure at work or leisure were poor.

INTRODUCTION

The course of noise-induced hearing loss has been well documented in studies of changes in hearing thresholds over the working lives of large numbers of individuals. (e.g., Bauer, Körpert et al., 1991; Pirilä, Sorri et al., 1991). A study of mining and steel-making industries showed that the effect of daily noise exposure was maximal in the region of 2-6 kHz (Abel and Haythornthwaite, 1984). The average rate of hearing loss at 4 kHz was 1.5 dB/yr, compared with 0.5 dB/yr in office workers. Individual differences were wide-ranging. Brühl, Ivarsson and Toremalm (1994) found the same rate of decline at 4 kHz in automobile sheet-metal workers. Absolute hearing thresholds increased with exposure level.

Several studies have focused on pre-employment hearing, in some cases to provide a baseline at the start of military service. Axelsson, Rosenhall and Zachau (1994) found a 14% prevalence of hearing loss greater than 20 dB HL in a sample of 500 18-yr old Swedish male conscripts. The loss was in the region of 4-8 kHz. Leisure noise was suggested as a possible factor. However, no correlation was found between hearing loss and listening to pop music, attendance at rock concerts, shooting or the use of firecrackers.

Impulsive sounds from weapons in military settings are particularly damaging to hearing. The effectiveness of hearing protectors for preventing hearing loss in infantry officers was studied by Christiansson and Wintzell (1993). The subjects ranged in age from 30 to 60 years. Peak exposure levels were between 136-185 dB SPL. All, including the youngest, had higher thresholds than norms published in ISO 1999 (1990). The maximum loss was at 6 kHz. By 45 years of age, the mean hearing threshold at 6 kHz was 40 dB HL. Hearing loss, in spite of hearing protector usage, suggests that hearing conservation programs developed for industrial settings may not be adequate for military trades.
RATIONALE

The present study of Canadian military recruits was a prospective within-subject investigation of the early development of hearing loss associated with four different trades, three characterized by high-level noise. The aim was to evaluate the prevailing hearing conservation program. Of particular interest were changes in the prevalence of hearing loss and problems associated with the prevention of loss.

METHODS AND MATERIALS

Subjects:

The sample comprised 134 subjects tested on admission to the Canadian Forces and after three years of employment: 38 infantry, 16 artillery, 37 armour and 43 in non-noisy trades. All had been screened for congenital or familial hearing loss, chronic ear disease or ear surgery, chronic complaints of tinnitus and/or dizziness, head trauma causing hearing loss, ototoxic drug exposure, employment in noisy trades for longer than three months, participation in noisy hobbies, and congenital or acquired external, middle or inner ear abnormalities or chronic eustachian tube dysfunction.

Procedure:

The testing was performed in quiet rooms with ambient noise levels below 30 dBA. The measurements were made using a portable audiometer (Soundlinked Data, Max 1000). Hearing thresholds were obtained for seven standard pure-tone test frequencies, ranging from 0.5 to 8 kHz. These stimuli were presented to the right and left ears separately over TDH headphones with MX 41AR cushions. A multiple-choice questionnaire completed at the three-year recall provided information about the utilization of hearing protective devices, instruction on the hazards of unprotected noise exposure and concerns that hearing protectors might interfere with job performance. Details are given in Abel and Pelausa (1995).

RESULTS

The percentage of subjects with hearing thresholds exceeding the 25 dB HL clinical fence for hearing loss (AAO-ACO, 1979) at 3-8 kHz was calculated at baseline and the three-year recall. Since the prevalence of hearing loss for the control group was 5% at baseline, only those percentages exceeding this value were considered significant. In the infantry, 11% of the subjects had a hearing loss exceeding 25 dB HL in the left ear at both 4 kHz and 6 kHz, after three years. Eight percent of the armour group showed a hearing loss in the right ear at 6000 Hz. A similar analysis was carried out using a 20 dB HL. fence for hearing loss (Axelsson et al., 1994). Application of the lower fence resulted in a relative increase in the prevalence of hearing loss. Baseline percentages ranged from 0% to 14% in the 3-8 kHz region. At the three-year recall, 21% of the infantry showed a hearing loss at 6 kHz in the left ear. The results for the other groups were within or close to the overall baseline range. The differences in threshold for right and left ears were then calculated within subject for each test frequency at baseline and at the three-year recall. These data showed the emergence of a between-ear difference in hearing in the infantry of about 7 dB on average at 4 kHz and 6 kHz.
Responses to the questionnaire indicated that in the noise-exposed groups, 31%-41% of individuals perceived that their hearing was slightly to moderately worse since joining the Canadian Forces, while only 2% of the control group felt that their hearing had changed. Forty-seven to 69% of the noise-exposed groups, compared to 12% in the control group, said that they were often or constantly exposed to loud noise. Eighty-three to 94% of the noise-exposed subjects compared with 38% in the control group judged the noise to be moderate or severe.

At least two-thirds of each of the four groups had not received any lectures or training films on the dangers of excessive noise. One-third or less rated their training as adequate or good. About half the noise-exposed groups received no instruction on the proper use of hearing protectors. About 80% said that they wore protectors in noise. However, 60% of these wore the devices less than half the time. Almost all subjects agreed that hearing protectors were definitely helpful. About one-third found them uncomfortable. Across groups, 70% believed that they could not hear as well when they wore protectors but only 25% felt that they posed a danger in their jobs. Forty-four percent to 86% of the noise-exposed groups experienced moderate to great difficulty understanding orders in a noisy room when protectors were worn, compared with 34% to 57% when protectors were not worn.

Half of each of the four groups indicated that they were exposed to loud sounds during their free time. The greater proportion, 66%-81%, associated these loud sounds with rock music and disco/dance bars. Across groups, on average 75% of individuals did not wear hearing protectors during these noisy activities. If protectors were worn during noisy leisure activities, plugs were generally selected in preference to muffls.

**DISCUSSION**

This study assessed the hearing status of military recruits after three years of employment in the infantry, artillery and armour trades. The results were compared with those of a control group working in non-noisy trades. These data showed that in all four groups, a small percentage of subjects, 2%-5%, had a high-frequency hearing loss exceeding 25 dB HL on admission to the Canadian Forces. Based on the information reported in ISO 7029 (1984), threshold shifts of at most 11-14 dB would be expected at the higher frequencies in 10% of highly screened, otologically normal 20-year-old males. The outcome observed in the current study is not surprising in view of published reports of hearing loss in teenagers from injurious levels of recreational noise.

By the three-year recall, the prevalence of high-frequency hearing loss exceeding 25 dB HL had increased to 11% at 4 kHz and 6 kHz in the left ear for the infantry. Using the 20 dB HL criterion, the prevalence of hearing loss in the infantry was 21% at 6 kHz in the left ear. According to ISO 7029 (1984), only 10% of a highly screened population of 30-year-old males would be expected to exceed a threshold of 16 dB HL at 6 kHz. By comparison, in an unscreened population with a high likelihood of both leisure and occupational noise exposure, 50% would be expected to exceed a threshold of 12 dB HL and 10%, 25 dB HL (ISO 1999, 1990). The direction of the interaural difference in hearing within subjects is consistent with the argument that the hearing loss was due to the use of small caliber weapons (Johnson and Riffle, 1992).
The questionnaire confirmed that the experimental groups were exposed to occupational noise which they judged to be moderate to severe. However, education regarding the hazards of noise exposure and instruction on the use of personal hearing protectors appeared to be poor. Almost all subjects believed that the devices were helpful, and only a small proportion felt that hearing protectors significantly impeded performance. A substantial number of the control group also felt that their place of work was noisy. This finding underscores the need to examine those trades which have traditionally been considered to be of low risk with respect to hearing. In addition, a large proportion of each of the four groups tested indicated that they were exposed to high-level leisure noise. Relatively few subjects wore hearing protectors during these activities, and it is likely this exposure contributed to the hearing loss observed.

ACKNOWLEDGEMENTS

The research described in this report was supported by a contract from the Department of National Defence, Canada. The authors are indebted to Sgt Pierre Lamontagne, Sgt Ken Murray, Sgt Fernand Niquette and Mrs Brenda Beauchamp for their help with the testing of subjects. Mr. Errol Davis and Mr. Soumen Kamarkar provided technical assistance.

REFERENCES


LOUDNESS AND ANNOYANCE OF INTERMITTENT AND CONTINUOUS ENVIRONMENTAL SOUNDS

Birgitta Berglund, Anna Preis*
Inst. Environmental Medicine, Karolinska Inst., and Dep. Psychology, Stockholm Univ., Sweden
* Dep. Psychology, Stockholm Univ. & Inst. Acoustics, A. Mickiewicz University, Poznan, Poland.

SUMMARY

Intermittent sounds have by some authors been claimed to be louder and more annoying and by others to be softer and less annoying than continuous sounds. By exchanging the A-weighted sound levels of the same two components of intermittent sound two hypotheses were empirically tested, namely if the very nature of intermittence makes sounds more annoying or if the stronger component of intermittent sounds predominates its annoyance. Depending on segment durations, some intermittent sounds were found to be louder and more annoying and some softer and less annoying than continuous sounds. The results show that the intermittence hypothesis for annoyance (and loudness) should be rejected. The annoyance of two-component intermittent sounds is rather determined by the most annoying but not necessarily the component highest in sound level. Thus the annoyance of sounds created from car plus pink noise is determined by the annoyance of pink noise whether or not the pink noise component is of higher or lower sound level than the car-noise component. It is probable that a masking effect between successive components in intermittent sound determines if the continuous or the intermittent sound becomes more annoying.

INTRODUCTION

Research on loudness and annoyance of intermittent sound has produced ambiguous results. For example, it has not been settled how these perceptual attributes of intermittent sound relate to those of continuous sound. Some pursues that intermittent sounds are louder and more annoying than their continuous counterparts [3], but the opposite have also been claimed [5]. It seems to us that this is an apparent contradiction that may be solved by adopting a more systematic research approach. The present findings emanate from two experiments in a planned set of converging experiments on comparative research on intermittent and continuous environmental sounds.

Each intermittent sound is composed of two components. If one component conveys information, the other is seen as distorting. In some contexts the distinction between the distorted and distorting part of an intermittent sound may seem artificial, for example, if both components are noises (continuous spectra). Which of two noises is perceived as distorted and distorting seems to be arbitrary. We [2] have earlier shown that the component that is distorting may be decided from the relationship between annoyance and the distortion ratio for intermittent sound. Here we ask if annoyance of intermittent sound is determined by its physical nature of intermittence or by the annoyance of its components (i.e., the loudest or most annoying component). These two hypotheses were empirically tested by scaling loudness and annoyance of continuous and intermittent sounds in which the sound pressure level (SPL) of the two components were reversed.

METHOD

Twenty subjects participated in Experiment I (7 men & 13 women; 19-37 years old) and 19 subjects in Experiment II (7 men & 12 women; 20-39 years old). All subjects were paid and none had earlier served in psychoacoustical experiments. All had normal hearing (less than 20 dB hearing loss in the frequency range 250-8,000 Hz; Bruel & Kjaer type 1800 audiometer).

Two continuous sounds (pink noise or car noise) each at two sound levels and different intermittent sounds composed thereof were used as stimuli. Pink noise was created from a random noise generator (Bruel & Kjaer type 1402) whereas car noise was a digitally recorded passage on an audio tape recorder (TCD-D10PROII; condenser microphone Bruel & Kjaer type 4155; power supply Bruel & Kjaer type 5935).

- 199 -
Fig. 1. Experiment I: Loudness or annoyance of continuous and intermittent sound. Segment duration refers to pink noise or gaps.

Fig. 2. Experiment II: Loudness or annoyance of continuous and intermittent sound. Segment duration refers to pink noise or gaps.
The intermittent sounds were composed from the continuous sounds using a digital sound editing program (Sound Designer II of the Sound Tools System & a Macintosh Quadra 950). In both experiments, two types of intermittent sounds were created as regular sequences of car noise and silent gap (C/G), or car noise and pink noise (C/PN). The time parameters were identical, the segment durations of the two-component intermittent sounds were 25, 50, 100, 250, and 500 ms. By combining the different segment durations for the two components, 50 unique intermittent sounds were created for each experiment: 25 C/G and 25 C/PN sounds. In Experiment I, the intermittent sounds were created from segments of varying durations of car noise at 63 dBA and pink noise at 66 dBA. These sound levels for the two components were reversed in Experiment II.

Each experiment was divided into two parts in which loudness or annoyance of the 52 stimuli of 4-s duration were judged separately. A balanced order between parts was used. Both loudness and annoyance was scaled with the method of free-number magnitude estimation. The instructions used were the original one developed by Stevens [4]. The annoyance instruction was supplemented as follows: Please imagine the following situation in which the particular sound will be heard: "You have just come home from a day's heavy work and have sat down in your favorite armchair to read a newspaper when you hear this sound" [1].

RESULTS AND DISCUSSION

For both experiments, the scales of loudness and annoyance of continuous and intermittent sounds were calculated and found to be of acceptable quality at individual as well as group level. The main results for the two groups are exhibited in detail in Figs. 1 and 2 for Experiment I and II, respectively. The segment duration on the abscissa refers to pink noise or silent gaps. In Fig. 3,

![Fig. 3. Loudness and annoyance of intermittent and continuous sounds as a function of distortion ratios.](image-url)
the same group scales of loudness and annoyance are shown as a function of the distortion ratio
defined as the duration sum of segments of one component to the total duration of the intermittent
sound [2].

In both experiments the continuous pink noise was perceived as louder and more annoying than
the continuous car noise notwithstanding the 3-dB difference in A-weighted SPL. The results con-
firm that A-weighted sound pressure level is neither an adequate correlate of loudness nor of anno-
ynce. In both Figs. 1 and 2 it is obvious that higher scale values were used for annoyance than for
loudness; the tendency is more explicit in Experiment II.

One general regularity in the results may be observed for car noise. Almost in all cases, contin-
uous car noise was perceived as the softest and the least annoying compared to the intermittent
sounds. The few exceptions refer to very short car-noise segment durations. For pink noise the re-
sults are different. In Experiment I (Fig. 1), continuous pink noise was louder and more annoying
for PNC segment ratios smaller than one but softer and less annoying for ratios larger than one.
On the other hand, in Experiment II (Fig. 2) continuous pink noise was the loudest and most an-
noying sound. The few exceptions were the very short car-noise segment durations.

A more comprehensive picture of the results is given in Fig. 3 by plotting loudness and anno-
ynce against the distortion ratio. The findings presented here may be viewed as a confirmation of
the results presented in Figs. 1 and 2.

CONCLUSIONS

(1) The hypothesis that annoyance of intermittent sounds depends on their intermittence is re-
jected. Rather the annoyance is determined by the most annoying component. Thus, the annoyance
of car plus pink noise sounds is determined by the annoyance of pink noise although its sound
pressure level may be higher or lower than that for car noise.

(2) The results show that some intermittent sounds are louder and more annoying and some
softer and less annoying than continuous sounds. It is probable that a partial masking effect be-
tween successive components in intermittent sound determines if the continuous or the intermittent
sound becomes more annoying. For example, in Experiment II continuous pink noise was more
annoying than intermittent sound because car noise segments masked pink noise segments (more
annoying component).

ACKNOWLEDGEMENTS

This research was supported by a grant from the Swedish Environmental Protection Agency.
We appreciate the technical and experimental assistance received from Mr. Tommy Bech-
Kristensen, B.A., Dr. Piotr Miecznik, Ph.D., and Mr. Mats E. Nilsson, B.A.

REFERENCES

[1] Berglund, B., Berglund, U., & Lindvall, T. Scaling loudness, noisiness and annoyance of aircraft
INRETS No 34 bis).
A Review. US Department of Commerce, National Bureau of Standards (National Engineering
Laboratory, Washington, DC 20234), Report No. NBSIR 81-2377, October 1981.
NEW RESULTS ON THE PITCH OF INTERAULLY DELAYED WHITE NOISE

Frans A. Bilsen, Han H. Dols
Perceptual Acoustics Lab., Faculty of Applied Physics, Delft University of Tech.,
The Netherlands.

SUMMARY

A white noise signal presented simultaneously to the left and right ear with an interaural time difference (ITD) between 0 and about 1 ms gives rise to the perception of a lateralized image. For ITD's larger than about 3 ms a faint pitch can be perceived corresponding to the reciprocal of the given ITD. The pitch image is positioned near the middle of the head (dichotic repetition pitch: DRP). Recently, it was discovered that the DRP-image can be lateralized by means of an interaural intensity difference (IID), similar to the classical "intensity image" though with a remarkable asymmetry.

Other dichotic pitch phenomena like Huggins pitch, Fourcin pitch, MPS-pitch, binaural edge pitch (BEP), most probably belong to a different class because they are more salient than DRP. More specifically: they don't show IID-lateralizability. Instead, they show ITD-lateralization common to the pure "time image". A tempting conclusion will be that time and intensity are processed separately at the earlier stages of binaural interaction.

INTRODUCTION

Pitch phenomena evoked by dichotic noise signals have influenced our thinking about binaural interaction significantly. For as far as dichotic pitch values were concerned, Bilsen and Goldstein (1974) showed that the similarity of dichotic and monotic repetition pitch with the low-pitch of normal periodic signals requires the existence of centrally generated spectral patterns with resolved (lower) harmonics for dichotic repetition pitch (DRP). From such spectral patterns pitch is extracted by pattern recognition (Bilsen, 1977).

Experiments on the lateralization of dichotic-pitch images of Huggins pitch (HP), Fourcin pitch (FP) and MPS-pitch (Raatgever and Bilsen, 1977; Raatgever, 1980) revealed that these pitch images behave like pure "time images" as postulated by Haferl and Jeffress (1968) for lateralized signals. This led to the conclusion that, for wide band signals, the time image is the result of spectral-pattern recognition in which frequency information is pooled across frequency for particular interaural delays. Based on the Jeffress (1948) scheme, a Central Spectrum (CS) theory of binaural processing was developed (Bilsen, 1977; Raatgever and Bilsen, 1986) that stresses both interaural cross correlation and central spectrum recognition. Recently, the related concept of "straightness" was introduced by Stern and Trahiotis (1991).

From the historical point of view, it is remarkable that dichotic repetition pitch (DRP) triggered our thinking about central spectra (Bilsen, 1972; Bilsen and Goldstein, 1974), whereas it (still) is the
dichotic pitch phenomenon that is not easily explained by the CS-theory in its present form. Lately, we discovered that DRP, contrary to the other dichotic pitch phenomena mentioned above, can be lateralized by means of an interaural intensity difference (IID), though with a remarkable asymmetry (Bilsen, 1994, see fig.1 and 2; B: delayed signal attenuated, A: undelayed signal attenuated). In the present paper these results will be pursued further.

**LATERALIZATION EXPERIMENT**

DRP-stimuli derived from lowpass-filtered gaussian white noise with a high-cutoff frequency of 2000 Hz, were generated in a digital signal processor (Loughborough DSP 96002; sample frequency 25 kHz; program code generated by a Comdisco SPW system). Thus a dichotic stimulus (see simplified block diagram in Fig. 1) with an ITD and an IID was obtained.

![Diagram](image)

**Fig.1. DRP-lateralization experiment**

In order to have a comfortable perception of DRP the ITD was programmed such that a 6%-interval in pitch was presented with a stimulus duration of 700 ms and a silent interval of 50 ms. The IID was programmed such that 19 different values were randomly presented ranging from -22.5 dB to +22.5 dB in steps of 2.5 dB, for each of seven values of the ITD in the range of 5.0 to 10.0 ms. The subject was seated in a sound proof booth and listened with headphones (Beyer DT 770) at a sensation level of about 40 dB at the unattenuated side. He had to indicate the in-head position of the DRP-image by adjusting the position of a dichotic white noise (pointer). He could switch between DRP-stimulus and pointer at will. For each of the seven ITD-values, a best linear fit was made (compare fig.2) providing the slope, viz. pointer-ITD-increment (in ms) divided by stimulus-IID-increment (in dB). This slope will be called: *laterability (ms/dB)*. Thus 14 laterability points were obtained, two for each ITD-value.

**RESULTS**

Three experienced subjects participated in this experiment. Their results are presented in Fig. 3. Open squares indicate delayed signal attenuated (positive attenuation; compare B in fig.1 and 2), while closed circles indicate undelayed signal attenuated (negative attenuation; compare A in fig.1 and 2).

![Graph](image)

**Fig.2. DRP-lateralization (Bilsen, 1994)**

It was investigated in a separate experiment that the pitch of the DRP-stimulus for a particular ITD did not change with varying IID. Thus it makes sense to present the results as a function of interaural delay (ITD).
CONCLUSIONS

- A DRP-image can be lateralized significantly by the introduction of an IID,
- The direction of movement is always towards the unattenuated ear,
- The lateralization seems proportional with the IID,
- The laterability is dependent on the ITD-value,
- The asymmetry reported by Bilsen (fig.2) is confirmed.

For ITD's smaller than 6 ms, case A (closed circles in fig.3) hardly shows any shift in DRP-position with changing IID, whereas case B shows a substantial shift with changing IID. In the latter case, the relation between perceived position and IID is about comparable to the function one observes for the lateralization of dichotic noise with changing IID (Blauert, 1983). A simple explanation for this asymmetry could not be found. (An intuitive argument could be that, in normal situations, the sidedness of dichotic white noise tends to the undelayed side if the ITD decreases beyond 9 ms (border of decorrelation). If the auditory system contains two separate mechanisms for ITD and IID, a conflicting situation might appear when the undelayed side is attenuated. Maybe, for an interaural delay less than 6 ms, the influence of the ITD on the lateral position of the DRP is stronger than the effect of the IID.)
For ITD-values larger than 6 ms, the results are rather subject-dependent. The asymmetry becomes less pronounced. Thus a general conclusion cannot be drawn.

In former papers (Raatgever and Bilsen, 1977, 1986), it has been shown that the dichotic-pitch images of HP, FP, and MPS can be lateralized by an (extra) ITD. The lateralization shows great similarity with the behaviour of the time image. IID's, on the contrary, are ineffective in lateralizing these dichotic pitches; they only affect their salience. An extra ITD, of course, cannot influence the lateralization of a DRP-image; it only determines its pitch. In the present paper it is shown, however, that for DRP IID's are effective in lateralizing the pitch image, be it in an asymmetric way.

Summarizing, Dichotic repetition pitch (DRP) is a faint binaural pitch phenomenon, evoked by the interaural delay of a single (white) noise signal. It appears not to be an artefact due to trivial interaural cross-talk (Bilsen, 1994). It has been shown in the past that its pitch behaviour provides evidence for a central-spectrum processing of pitch, rather than cross-correlation processing for which (according to Fourcin) two uncorrelated noise sources would be required. Nevertheless, DRP-extraction is still not fully understood. The present experiments were intended to get further understanding of the binaural processes involved.

The final tempting conclusion, could be that there are two separate mechanisms: a mechanism to evaluate ITD's and a mechanism to evaluate IID's. DRP could very well be the product of the IID-mechanism exclusively, contrary to HP, FP and MPS that are clear exponents of the ITD-mechanism.

REFERENCES


THE AUDITORY REPRESENTATION IN VIRTUAL REALITY

Jens Blauert, Hilmar Lehnert
Lehrstuhl fuer allg. Elektronotechnik und Akustik
Ruhr-Universitaet Bochum, D-44780 Bochum, Germany

SUMMARY

Virtual Reality (VR) refers to a technology where the sensory input into human beings is generated by computers. As this computer-generated input can be formed to be physiologically adequate, e.g. like sensory input from the real world, perception of immersion and presence can be created for the subjects who are exposed to it. To this end, adequate treatment of the prominent sensory modalities, including the auditory one, is mandatory.

Technically, auditory representation in VR systems is implemented by means of a sound system. However, in contrast to conventional sound systems, the auditory scenario is generated and controlled by computers. Furthermore, the auditory representation is non-stationary and interactive, i.e., among other things, dependent on listeners' actions. As a consequence, the auditory component of VR systems can no longer be dealt with simply in terms of linear time-invariant systems.

INTRODUCTION

A traditional task of electroacoustics is to transmit auditory scenes across time and space. However, authentic reproduction is rarely required, even when so-called high fidelity is at stake. At the play-back end of the electroacoustic chain, the aim usually is an "enhanced" representation of the original auditory scene - e.g. enhanced with respect to the optimum artistic effect.

Consequently, electroacousticians have developed a battery of methods and tools over the years which enable them to modify recorded sounds in many ways. For example, the sound of an orchestra on a CD disk has usually been mixed from a number of individual sound tracks representing different musical instruments or groups of them. During the mixing process the auditory perspective has been re-arranged, and acoustic room properties like early reflections and reverberation have been added.

On the basis of the knowledge that enables electroacousticians to modify and arrange sounds, numerous applications have been developed focusing on simulation, enhancement and/or creation of auditory scenes. Examples for such applications are: public address systems to amplify natural voices for better intelligibility, so-called electronic-architecture systems to modify acoustic properties of halls, multi-channel systems to produce a sense of auditory environment in the cinema or when watching TV at home (home-theatre), so-called spatialiser to create interesting sound effects in the context of electronic music and/or computer games.

Auditory representation in virtual-reality systems is based on the same methods and techniques which are used to produce auditory illusions in applications such as the ones mentioned above. Yet, a new dimensions has been added: VR systems aim at creating a state of mind in the users which is described as immersion and presence. In other words, VR systems are designed to prompt the illusion of actually being present in a world which is de facto virtual, i.e. created and
presented by a computer. The users, consequently, being in a state of mind where they take the virtual world as real (virtual reality), act and react very much in the same way - in many cases intuitively - as they would in a corresponding real world. Their action may, to be sure, include interactions with the virtual world.

VR systems, as a rule, are multimodal systems, i.e. they provide two or more of the following representations: visual, auditory, tactile (incl. thermal), olfactory, gustatory, and kinaesthetic.

ROOM-RELATED VS. HEAD-RELATED REPRESENTATION

Consider a modern flight simulator as a system that comes close to a VR system. Such a simulator has the purpose of providing pilots the illusion that they are flying a real plane in a real world. The cockpit in these systems is a mock-up from components of a real cockpit. The visual illusion of a world outside the plane is provided via video-projection on screens in front of the windshield, while the auditory illusion is generated with a set of loudspeakers. The idea with respect to the acoustics of the simulator is to produce a sound field in the cockpit mock-up which is perceptively equivalent to the sound field in cockpits in corresponding real flight situations. This approach is called the room-related approach to auditory representation.

If the sound field is generated in a realistic way, the pilots can move their heads in the cockpit without losing the impression of realism. In particular, sound sources at fixed positions in space will not give rise to moving percepts when the pilots move their heads. It is not a trivial task to achieve this perception of "spatial constancy". As a rule, more than 15 loudspeakers driven by 15 independent channels are necessary to achieve this end.

In the flight simulator considered, the cockpit mock-up is built from real elements. The perception of these is superimposed by visual, auditory and other cues generated by a computer. Such a combination of real and virtual representations is called augmented or joint reality. In an integral virtual reality, no real objects are used at all. In the case of a flight simulator, the cockpit would thus be computer-generated as well. This implies, for the auditory representation, that physiologically adequate sound signals have to be delivered directly to the auditory systems of the listeners, namely to their eardrums. The technical way to accomplish this is via transducers at the entrances to the earcanals (headphones).

Headphones are fixed to the head and thus move simultaneously with it. Consequently, head and body movements do not modify the coupling between transducers and earcanals (so-called head-related approach to auditory representation) - in contrast to the case where the transducers, e.g. loudspeakers, are positioned away from the head and were head and body can move relative to the sound sources. As a result of these movements, the transmission paths from the sources to the ear-drums vary - depending on the directional characteristics of sources and external ears (skull, pinna, torso) and with respect to reflections and reverberation.

Virtual-reality systems must take account of all the specific variations and modifications of the sound paths from the sound sources to the listener's ears which occur in reality when the listeners and the sound sources move. Only if this task is performed with sufficient sophistication will the listeners accept their auditory percepts as real - and develop the required sense of presence and immersion.

ARTIFICIAL-HEAD TECHNOLOGY

Early head-related sound systems are characterised by the use of artificial heads. These heads are dummies which represent natural heads with respect to their acoustic properties. Particularly, they represent a mechanical implementation of the transfer characteristics from each sound source to each of the eardrums, including reflective paths. In the cases of spatially fixed scenes, the sound paths are linear and time-invariant - and can thus be characterised by so-called head-related transfer functions (HRTFs). Advanced head-related sound systems substitute the artificial heads by using banks of electronic HRTF filters.
A problem with HRTFs is as follows. These transfer functions are different for individual ears. Thus, if a head-related system is not individually adjusted to a particular listener, he/she will "listen through somebody else's ears." As a consequence, undesired perceptual effects may occur, such as front-back inversion or colouration of timbre.

BINAURAL ROOM SIMULATION

To include the physics of sound fields in enclosed spaces into head-related systems, the reflective components of the sound have to be considered (binaural room simulation). Today's systems individually compute up to a few hundred early reflections. Later reflections are typically accounted for by an additional reverberation algorithm. Please note that each direct and reflective sound path must be weighted by a specific HRTF - selected with respect to the specific angle of sound incidence and the specific distance of the sound source.

HRTFs denote linear time-invariant systems. Binaural room simulation systems which rely on this approach mimic cases where the sound sources are fixed in space, and the listeners do not move in relation to them. Systems of this kind may be sufficient for the auditory exploration of the "acoustics" of architectural design or, e.g., spaces for musical performances. Yet, they cannot account for more vivid situations required by VR systems. For example, if listeners to such binaural room-simulation system move their heads, the perceived auditory scenes will also move - a sense of spatial constancy will thus not be achieved.

HEAD TRACKING

Advanced binaural sound systems perform all HRTF filtering and sound-field simulation electronically. Consequently, if the computer is informed about head movements, the input signals to the listeners' ears can be controlled accordingly. To this end, head positions are tracked by devices based on, e.g., mechanical, optical or magnetic principles.

For relatively slow head movements, the system may synchronise by appropriate switching or interpolating between different steady-state HRTFs. The listeners will then perceive the sound sources at fixed positions in space. It goes without saying that this simple approach will fail for faster head and/or source movements. It is worthwhile noting that perceptual front-back inversion no longer occurs when head-tracking is employed, even with HRTFs which have not been adjusted individually.

INTERACTIVITY

VR systems typically allow for interaction between the subjects and the virtual reality. For example, in a system currently under development in the authors' laboratory in co-operation with four European partners (ESPRIT project SCATIS), the following experimental scenario is about to be implemented.

A subject will be exposed to a virtual space with various (invisible) auditory/tactile objects distributed in it. He/she will localise and identify these virtual objects auditorily and be able to reach out to them and grasp them individually. Upon tactile contact, contour, texture and thermal attributes of the virtual objects will be perceived. The task of the subject will be to manually move the objects around, i.e., to re-arrange their spatial position and orientation according to an experimental plan. Auditory feedback will be given.

A system of this kind aims, among other things, to precisely generate and auralise a considerable number of sound sources and early reflections in prescribed directions of incidence - under circumstances where the head and body, as well as the sound sources, vary their positions in space at realistic speeds. It goes without saying that the computer has to know the positions and trajectories of the sound sources as well as those of head and hand (palm and fingers) of the subject at any given time. The architecture of the system includes "afferent" as well as "efferent" channels to provide the necessary links between the computer and subject.
PERCEPTUAL REQUIREMENTS

As to the quality of auditory representations, there are two important perceptual features which are unique to computer-generated auditory scenes: smoothness and responsiveness. Smoothness has to do with the refresh rate with which the display is updated. This rate should be high enough so as to not to generate auditory discontinuities ("flicker"). Responsiveness has to do with the time lag between the presentation of synchronous events in different sensory modalities and/or between an action of the subject, e.g. the turn of the head, and the reaction of the display. This time lag should ideally be short enough for the subject to not be aware of any unnatural assynchrony.

Limits for refresh rates and time lags cannot be given in the form of fixed numbers as they depend heavily on the specific scenario. As to the refresh rate, it is easily understood that the speed of movements incurred is of paramount importance - for a still scene no refresh would be necessary at all! The ESPRIT project SCATIS aims to provide more detailed information on requirements with regard to smoothness and responsiveness of an auditory-tactile virtual scenario.

ISSUES OF IMPLEMENTATION

A typical VR system may, following its afferent auditory pathway, be composed from the following chain of components: a computer model of the virtual world (containing the objects with their attributes and the laws which govern the relations between them), an event-handling layer (tracking events in the virtual world and actions of the subject, and initiating and controlling the responses of the system), an auditory renderer (performing the binaural sound-field modelling), an auralisation module (providing binaural signals based on the sound-field model and the source signals) and, finally, the auditory effectors (headphones).

Special hardware is particularly needed for the auralisation module. While in a binaural room-simulation system for a still scenario the renderer could be a two-channel convolution engine for impulse responses with a length of a few hundred milliseconds, VR systems for time-variant interactive scenarios need more complex hardware. As a matter of fact, in such systems the direct sounds from each source have to be processed individually as well as a sufficient number of early reflections. The need for individual treatment becomes evident when considering that in non-static scenarios the characteristics of different sound paths vary in different ways. For example, upon movement of a source relative to a listener, the direct sound path may become shorter while reflective paths become longer. Additional reflections may occur in some paths, while reflections disappear in others. Reflection coefficients of walls and radiation coefficients of sources may vary as a function of changing directions - differently for different sound paths. Effects of time-variance, e.g. Doppler shifts, may come into play.

With hardware which is commercially available at this time, up to about 36 different sound paths can be processed simultaneously in real time. This is already sufficient for applications where "naturalness" is not a key issue, e.g. video games. More complex effects of time variance have not yet been considered for auditory representation in commercial VR systems. As the number of available sound paths is limited in auralisation modules, decisions will have to be taken on how to use them most effectively. The more the simultaneously-active sound sources require direct sounds paths, the less the reflections can be mimicked. The decisions can be revised framewise. The number of available sound path may depend on the number of filter coefficients employed for each individual sound path at a given time (for HRTFs, reflection coefficients, and directional characteristics of the source). As a rule, direct sound paths and strong early reflections need preciser filtering than the rest of the reflective sound paths. Late reflection with similar angles of incidence may be bundled and then treated with the same HRTFs, as the spatial selectivity of the human hearing is limited.

There is some evidence that between 100 and 200 sound paths per sound source are sufficient to generate complex auditory scenarios, e.g. a musical performance in a concert hall, in a perceptually adequate way - provided that adequate reverberance is added by special algorithms.
EFFECT OF TONE FREQUENCY RANGE ON TEMPO SENSITIVITY

Marie-Claire Botte, Carolyn Drake, Nathalie Panissal-Vieu, Caroline Bey
Laboratoire de Psychologie, Université René Descartes, CNRS (URA 316) 28,
rue Serpente F-75006 Paris

SUMMARY

Tempo sensitivity is known to be influenced by the length, temporal structure and rate of sequences. In this paper we examine whether tempo sensitivity is affected by the range of tone frequencies in the sequence. Tempo IND for isotonic sequences were compared with those obtained for sequences in which the frequency of tones varied within either one or six critical bands. The results show that tempo sensitivity is not impaired when the frequency variations are limited within one critical band, whereas impairment is observed for larger frequency variations. These data suggest the existence of interactions which might be attributed to a relatively peripheral process (streaming mechanisms based on critical bands), and more central processes involving the complexity of the sequence representation.

Auditory sequences are composed of auditory events (sounds having certain physical characteristics: frequency, amplitude, spectral envelope etc.) which occur at certain points in time (temporal structure). Whereas there has been a tendency to study these two dimensions separately, two fields of research have concentrated on the relations between the two. The first concerns auditory stream segregation in which listeners may hear an auditory sequence as two streams depending on the physical characteristics of the events in a sequence: one stream is perceived if the frequency between the events is close, whereas two streams are perceived if the frequency range is wider. However, this tendency also depends on the temporal structure: the faster the sequence, the more likely a listener is to perceive two streams. The classical studies of van Noorden (1975) and Jones (1976) have described relatively precisely the limits under which one or other type of perception is observed. The second field of research concerns rhythm and time perception, where, in addition to studying the role of temporal factors in the structuring of sequences, some effort has gone into examining the effect of physical characteristics such as frequency on time perception. For instance, it is harder to detect a temporal lengthening between two sound events if they vary in frequency (Drake, 1993; Fitzgibbons, Pollack & Thomas, 1974). Research in these two fields has tended to remain separate, with studies in stream segregation concentrating primarily on physical characteristics in relatively fast sequences, and with studies in rhythm perception concentrating primarily on temporal aspects in relatively slower sequences. In this paper we hope to demonstrate how these two approaches should be considered as complementary.

Tempo sensitivity

In previous studies on tempo sensitivity (Drake & Botte, 1993; 1994), we have demonstrated that tempo sensitivity is maximum in a range of interonset intervals covering approximately 200 to 600 ms, a range larger than those found previously by Michon (1964)
and Fraisse (1967). We have also shown that tempo JND decrease as a function of the number of intervals, that is the number of tones in the sequence. All these studies examine tempo sensitivity in isotonic sequences, that is sequences in which all the events are of the same frequency. The present experiment aimed at evaluating the effect of a frequency variation from tone to tone on tempo JND. Two theoretical positions exist. First, temporal and physical characteristics of tones may be processed independently. Evidence in support of this position has been proposed by Peretz (1993) whose neurological observations indicate that patients can selectively process either the rhythmic or melodic pattern of musical sequences as a function of the location of their cerebral impairment. Could it be that tempo is processed independently of melodic contour by normal subjects? If this is true, tempo sensitivity should not vary as a function of the melodic contour of sequences. The second position considers that the two are closely inter-related with temporal aspects of a sequence influencing the processing of the physical characteristics of events and visa versa (Jones, 1987). If this is the case, frequency variations of tones should affect the tempo representation of sequences.

We therefore measured tempo JND in the same way as previously, but this time, instead of using only isotonic sequences, we varied the frequency of each tone in a sequence. The tone frequencies were either within a narrow frequency range (one critical band), or a wide frequency range (6 critical bands). A wide range of tempo rates were used in order to test whether different processes were involved at different rates.

How may frequency variations of tones affect tempo sensitivity? One explanation concerns the mental representation for the sequence: this representation should be more complex when it involves both temporal and melodic aspects than when it only involves temporal aspects (isotonic sequences). This may lead to greater difficulty in extracting information about tempo in sequences with both types of variations. Also, frequency range may influence the complexity of the representation, so the lowest sensitivity would be expected in sequences with the widest frequency range. Another phenomenon, that of auditory streaming may intervene. Streaming may occur for some of our faster sequences with a wide frequency range. For sequences which are perceived as more than one stream, overall tempo perception is difficult because adjacent tones are not perceived in the same perceptual unit. Tempo representation should therefore be impaired in conditions of auditory streaming. Different patterns of results, reflecting different processes, may therefore be observed for fast and slow sequences.

METHOD

Subjects. Four musicians (who had been receiving formal training for at least five years and played regularly), and four non-musicians (who had received not more than one year of formal training) participated in the experiment. They were paid for their services.

Materials. Sequences were constructed with 50-ms tones which were presented binaurally, via earphones, at 70 dB SPL. Each sequence was presented at 6 different standard tempi (with IOIs ranging from 100 to 750 ms). Sequences were composed of twelve tones at the fast tempi (100-, 150-, and 200-ms IOI), and six tones at the slow tempi (250-, 500-, and 750-ms IOI). This was done to ensure minimal tempo JND by using an optimal temporal window for tempo representation. Four types of sequences were constructed which varied according to the frequency range of the tones they contained: 1) isotonic: all tones were of the same frequency (440 Hz), 2) narrow frequency range: the frequency of each tone in a sequence was randomly chosen (without replacement) among 12 frequencies which were regularly spaced within one critical band (about 3 semi-tones) centered at 440 Hz, 3) wide frequency range: the frequency of each tone in a sequence was randomly chosen (without replacement) among 12 frequencies which were regularly spaced within six critical bands (about one and a half octaves) centered at 440 Hz, and 4) musical wide frequency range: the frequency of each
tone in a sequence was randomly chosen (without replacement) among 12 musical semi-tones (from C4 to G5), which also corresponded to about six critical bands. In the last three cases, the contours of the first and second sequences were either the same (the same random sequence) or different (two random sequences). This manipulation was of course impossible for the isotonic sequences. There were therefore seven frequency range conditions: isotonic, narrow identical, narrow different, wide identical, wide different, musical identical, musical different.

Procedure. Tempo JNDS were measured by the same method as previously, in a 2AFC paradigm, with an adaptive procedure converging on the difference in IOI which was correctly detected 84.1% of the time. Subjects heard one sequence followed by another which was slightly different in tempo and their task was to indicate which sequence was slower. The six standard tempi were tested successively in a counterbalanced order. For each standard tempo, the JND was measured twice in each of the seven frequency-range conditions. Therefore, eighty-four differential thresholds were measured for each subject, which took about six 2-hour sessions. Subjects sat in a sound-proofed cabin and listened over headphones.

RESULTS

JNDS for each frequency range condition will be examined as a function of tempo. Results are averaged over musicians and non-musicians because an ANOVA by group (musicians/nonmusicians), frequency-range condition (7), and tempo (6) revealed no significant difference between the two groups.

Figure 1: Relative tempo JND for four frequency ranges (identical, narrow, wide and musical wide).

Effect of frequency range. Figure 1 presents the mean results of the eight subjects for the isotonic, narrow, wide and musical wide conditions averaged over the 7 tempi and the identical or different contours. JND were low for the isochronous (2.6%) and narrow range (2.6%) conditions, and higher for the wide (3.4%) and musical wide (3.8%) conditions (main effect of condition: F(6,36) = 15.35, p<0.01). Planned comparisons revealed no significant differences between the isochronous and narrow range conditions. However, JND for the isochronous conditions were significantly lower than both the wide (F(1,6) = 12.95, p<0.01) and musical wide conditions (F(1,6) = 19.30, p<0.01). JND for the narrow range conditions were lower than both the wide (F(1,6) = 41.81, p<0.01) and the musical wide conditions F(1,6) = 39.96, p<0.01). Therefore, events may vary in frequency within one critical band without impairing tempo sensitivity (compared with isotonic conditions), whereas impairment is observed if the events vary within a wider frequency range (in the present case 6 critical bands). Interaction with tempo. However the story is more complicated than this. The expected main effect of tempo was observed (F(5,30) = 15.52, p<0.01) with a U-shaped curve. There was
also a significant interaction between tempo and condition ($F(30,180) = 2.52$, $p<0.01$). When each tempo was examined separately, the pattern described above was observed at all tempi, but the condition effect is only significant for the fastest tempo (100 msec) and the two slowest tempi (500 and 750 msec). The processing of tempo at intermediate rates is less affected by frequency range than at fast and slow rate. Such a pattern suggests that different types of processing may occur for sequences at different rates.

**Effect of sequence musicality.** Overall, JND were significant higher in the musical wide (3.8%) than in the wide (3.4%) conditions ($F(1,6) = 11.25$, $p<0.01$), but when each tempo was examined separately, this effect was only significant for the fastest tempo (100 msec). Thus, the coherence of events in a sequence leads to impairment in tempo processing, particularly in fast sequences.

**Effect of contour.** Tempo discrimination was harder when the contours of the two sequences to be compared were different (3.44%) than when they were identical (3.06%)($F(1,6) = 5.40$, $p<0.05$), and this pattern was seen for all the tempi. Thus, the processing of frequency contour interferes with tempo representation.

**DISCUSSION**

These results demonstrate that tempo sensitivity is not influenced by frequency variations of successive tones when they are limited within one critical band. However, tempo sensitivity is impaired when the frequency variations are wider (in this case within six critical bands). Whereas this pattern was observed for all tempi it was particularly marked at the very fast tempo of 100 ms. This is probably due to the fact that the creation of the mental representation for the sequence was hindered by the process of auditory streaming which prevents adjacent events being processed in the same unit.

Rather than an independence between the processing of temporal and melodic aspects of tone sequences as suggested by Peretz (1993), these data suggest the existence of interactions which might be attributed to a relatively peripheral process (streaming mechanisms based on critical bands), and more central processes involving the complexity of the sequence representation. This highlights the importance of incorporating findings from both streaming and rhythm perception.

**REFERENCES**


A MATHEMATICAL DESCRIPTION OF PSYCHOACOUSTIC SPACE

P. Camiz, M. Beato
Università di Roma "La Sapienza", Roma, Italy

Human beings as well as some animals are able to recognize many features of sounds, such as pitches, harmonic relationships, vowels, instrumental characters, which are related to different levels of analysis of the frequency spectrum, namely to the large, intermediate or small scale. This ability depends on the physiological structure of the auditory apparatus, but also on the training during the early period of life, when the capacity of developing neural connections is at its maximum level.

The auditory experience develops both in time as well in frequency domain, and its mathematical description is satisfactory accomplished by a time-dependent spectral distribution: our main interest is focused on the spectral aspects and on their influence on the learning process, so that we consider the auditory experience as a succession of independent events. In a communication presented at SMAC93 (1) one of us stressed the difficulty, for a simple system, to recognize at the same time pitch and vowel; in a recent work Beato (2) simulates the learning process for sung vowels by a self-organizing neural network of Kohonen type, and finds an interesting topological transition depending on the frequency resolving power of the sensorial units of the first layer.

In this communication we present a slightly different approach, more transparent from a mathematical point of view, based on the eigenvectors and eigenvalues of a correlation matrix (3), which has proved fruitful also in describing visual processes, such as the organization of perceptive fields in colour perception(4).

The model consists of two layers of n units, completely connected by $n^2$ links; the units of the first layer are resonators tuned to a given frequency, whose responses $r_i$ ($i=1,...,n$) to acoustic stimuli run continuously from 0 to 1. The set of auditory experiences consists of N normalized vectors $r(a)$ ($a=1,...,N$) in an n-dimensional space, with non negative components $r_i(a)$; the normalized correlation matrix $G$ is defined as the average over the whole set of auditory experiences of the outer products of the input vectors with themselves: $G_{ij} = (1/N) \sum_{a=1}^{N} r_i(a) r_j(a)$. 
$G$ is real, symmetric, its elements are non negative numbers, such that $(G_{ij})^2 \leq G_{ii} G_{jj}$, and $\text{Tr } G=1$; its eigenvalues $g_\lambda$ range from 0 to 1, and represent the weights of the corresponding eigenvectors $\psi_\lambda$ in the auditory experience; an "incomplete auditory experience" generates a matrix with rank smaller than $n$. The eigenvectors, in the $n$-dimensional space of the first layer, may have positive and negative components $\psi_\lambda(i)$, so that they cannot interpreted as possible auditory stimuli: a more appealing interpretation is that the $i$-th component of the $\lambda$-th eigenvector represents the strength of the connection between the $i$-th unit of the first layer and the $\lambda$-th unit of the second, positive connections being excitatory and negative ones inhibitory.

According to this interpretation the (auditory) experience selects the units of the second layer according to the corresponding eigenvalues: if a threshold has to be chosen a reasonable value seems to be $1/n$, which corresponds to the $n$-degenerate eigenvalue of the unit matrix divided by $n$; the learning process drives the eigenvalues of the more relevant eigenvectors over the threshold, pushing downward those belonging to the less stimulated eigenvectors.

Correlation matrices may exhibit a variety of structures, depending on the choice of input vectors among pure noises (continuous frequency spectrum), simply or multiply periodic sounds (discrete frequency spectrum), with or without vowel formants, or a suitable mixture of all of them, more adequate for the reproduction of a realistic acoustic environment.

In this communication I want to present some preliminary results of a computer simulation in which $n$ has been chosen equal to 60, and the resonance frequencies were one semitone apart from each other, so that our first layer covers 5 octaves; we have produced two correlation matrices: the first one is obtained from simulated musical sounds (a fundamental accompanied by 8 overtones of amplitude decreasing as $1/k$, $k=1..8$) whose pitch runs over 5 octaves; the second one uses the same sounds filtered by the first two formants of the Italian vowels A-E-I-O-U. The eigenvalue equations $G \psi_\lambda = g_\lambda \psi_\lambda$ ($\lambda=1..n$) have been solved by standard computational techniques. The first 12 eigenvectors of the two matrices are presented in fig.1 and 2: in both cases the largest eigenvalue corresponds to an eigenvector with components of the same sign, meaning that the most important unit of the second layer is sensitive to the overall intensity of the stimulus and only in the second case to the large scale structure of its frequency distribution. Roughly speaking half of the eigenvalues are larger than 1/60, which is the actual value of the threshold previously defined: this means that the relevant information contained in a stimulus already experienced is well represented by only 30 amplitudes out of the 60 availables.

The first correlation matrix is invariant under reflection with respect to the secondary diagonal: as a consequence of this invariance its eigenvectors are either symmetric or antisymmetric under reflection with respect to the middle point of the frequency axis.

A detailed analysis of all the eigenvectors is beyond the scope of this paper, but some interesting features can be outlined: almost all eigenvectors of the first matrix present large structures modulating rapidly varying oscillations. The large structure can be used for a rough localization of the dominant frequency of a sound, while the fine structure is able to analyze its
musical content: it is possible to recognize in many eigenvectors strong positive correlations for consonant intervals and strong negative ones for intervals like semitones.

The second matrix does not have the invariant property of the first one, so that its eigenvectors are much less regular, and their large and small scale structures less evident: in this case the large structure is relevant for the recognition of vowels, the fine structure being more sensitive to the musical details, as in the previous case. As an example of a rough vowel discrimination we can compare eigenvectors n.1 and 6: a vowel whose formants are far apart has strong negative components on both eigenvectors, whereas a vowel with close formants has negative component on n.1 and a positive one on n.6. The particular choice of the frequency scale with respects to the formant's width has produced a strong depression of matrix elements corresponding to very low and very high frequencies: as a consequence the eigenvectors belonging to large eigenvalues have components distributed over the most relevant frequency region, whereas the last eigenvectors have only one or two large components concentrated in the extreme frequency regions; it must be stressed that the corresponding units of the second layer are ready to label sounds which have not been previously experienced.

The second part of our work has been devoted to the analysis of the second layer response to acoustic stimuli and to their reconstruction: for this purpose we have chosen the vowels "I" and "O" whispered, and we selected 6 pitches in the low, middle and high frequency region, corresponding to primary units n.6,7,18,25,37,49, (their difference gives the interval in semitones): each one has been presented alone, with its 8 overtones, and filtered with the formants of vowels "I" or "O". For each stimulus the projections over the 60 eigenvectors of both the correlation matrices have been calculated: according to our interpretation of the eigenvectors the projections represent the level of excitation (positive or negative) of the second layer units, therefore the attention should be focused on those with high absolute value.

The reconstruction of stimuli has been done adopting different threshold criteria in order to compare the faithfulness of different partial reconstructions: therefore we selected either the eigenvectors belonging to eigenvalues larger than 1/n=1/60, or the eigenvectors whose weight (in each stimulus) was larger than 1/n, and this for the two sets of eigenvectors. Some results are shown in fig.3: for the two whispered vowels the best reconstruction is the one obtained with the most important eigenvectors of the second correlation matrix: for the vowel "O" it seems that the formants separation is even more pronounced than in the original sound; in both cases the most relevant amplitudes are n.1,6 and 10. For sung vowels the reconstructions via the large components criterion are almost equally good; their reconstructions via the importance criterion are less satisfactory, but deserve a comment due to the appearance of some "harmonic ghosts", which occurs also for sounds without overtones: the strong harmonic correlations in the G matrix are responsible for the presence in the partial reconstruction of some frequencies belonging to the harmonic or subharmonic spectrum of the fundamental, or for the reappearance of the fundamental itself even when it was filtered out by the formants.
These preliminary results show that different acoustic environments can generate different structures in the model's psychoacoustic space and correspondingly differences in musical and linguistic individual abilities; this kind of behaviour is not in contradiction with observations on living beings, and we believe that our model may represent an idealized description of real neurologic processes. A full account of this investigation will be published elsewhere.

Fig.1) The first 12 eigenvectors of the correlation matrix G1 in frequency space: the origin corresponds to 150 Hz.

Fig.2) The 12 first eigenvectors of the correlation matrix G2.
Fig. 3) In each of the 6 sectors; up-left: the response to the stimulus; up right: its projections on G1 eigenvectors (down) and on G2 eigenvectors (up); down left: partial reconstructions of the stimulus with G1 eigenvectors via the large components criterion (up) and with the large eigenvalue criterion (down); down right: the same with G2 eigenvectors. - The stimuli: 1)-vowel "I" whispered, 2)-vowel "O" whispered, 3)-pure fundamental in position 18, 4)-the same as 3 with its overtones, 5)-the same as 4, filtered through the "I" formants, 6)-the same as 4, filtered through the "O" formants.

REFERENCES

1) P. Camiz, "A physico-mathematical model for the spontaneous organization of the auditory cortex" in Proceedings of SMAC93.

2) M. Beato, "A topological transition in the map of acoustic space generated by a Kohonen neural network" in preparation.

3) P. Camiz, "The role of correlation matrices in the organization of perceptive spaces" in preparation.

MECHANISMS OF COCHLEAR FUNCTION:
OBSERVATIONS FROM MAMMALIAN HAIR CELLS

M.A. Cheatham
Auditory Physiology Lab., Hugh Knowles Center, Frances Searle
Building, Northwestern University, Evanston IL, 60208 USA.

SUMMARY

Although it is well known that the peripheral auditory system is nonlinear,
characterizations have been based largely on reports about single unit responses. Only
recently, in fact, have direct measures of cochlear nonlinearities been made in basilar
membrane mechanics (Rhode 1971) and in sensory hair cell responses (Russell and Sellick
1978). The latter provide advantages over neural measures because they are not complicated
by the nonlinearities and time delays inherent in synaptic transmission and spike generation.
In the present experiments, a very low frequency tone is used to bias the cochlear partition
(Deatherage et al. 1957; Nieder and Nieder 1968 a,b) in order to reveal the relationships
between cochlear mechanics and the excitatory processes associated with cochlear hair cells
(Nieder and Nieder 1971).

INTRODUCTION

Interactions between two simple, sinusoidal inputs are evaluated here in order to
understand how signals are coded in the peripheral auditory system. While it is known that a
two-tone input results in production of numerous combination tones (Nuttall and Dolan 1990,
1993), we focus on the degree to which one tone can suppress the response produced by
another. This phenomenon, known as two-tone suppression, has been observed in basilar
membrane (Rhode 1977), inner hair cell (IHC) (Sellick and Russell 1979) and single unit
responses (Rupert et al. 1963) for suppressors both above and below the characteristic
frequency (CF) of the place under study. In this communication, however, we employ a very
low frequency tone at 40 Hz to investigate changes in responsiveness as a function of time in
IHC and organ of Corti (OC) responses from second turn of the guinea pig cochlea where
characteristic frequency is approximately 3500 Hz. The low frequency tone is combined with
a variable frequency probe to determine the degree to which tones at and below CF are
modulated. Phase dependent changes in magnitude are then correlated with position of the
basilar membrane.

METHODS

In the paradigm used here, the starting times of the two sinusoids are staggered allowing
responses to each individual input to be recovered in addition to those in the region of overlap
where both tones are presented together. The position of the basilar membrane is determined
using the polarity of the cochlear microphonic (CM) response produced by the 40 Hz tone
alone (Konishi and Nielsen 1978; Patuzzi et al. 1984b). The positive phase of the CM is
associated with basilar membrane displacement to scala vestibuli (SV); the negative phase
with displacement to scala tympani (ST). Since the CM is measured from the OC fluid space in the same region where IHCs are recorded, it provides a better indication of basilar membrane position than does the CM recorded from the round window (Konishi and Nielsen 1978; Klis and Smoorenburg 1985; Cheatham and Dallos 1994). Because IHCs can respond to both inputs, the averaged response waveforms are high-pass filtered off line so that modulations of ac receptor potentials are clearly displayed. This investigation was approved by the National Institutes of Health and Northwestern University’s Animal Care Committee.

RESULTS

Figure 1A shows an IHC ac receptor potential for a near CF probe at 3500 Hz measured in the presence of a second tone at 40 Hz. The CM response to 40 Hz alone, plotted with dashed lines, is also appended. The waveforms reveal responses to 40 Hz alone on the left and to 3500 Hz alone on the right. In the region of overlap, the ac receptor potential is reduced but only during the negative phase of the CM response when the basilar membrane is displaced to ST. Similar modulations are observed for the OC response in Fig. 1B. These response patterns are similar to those from single units (Sellick et al. 1982; Patuzzi et al. 1984a) and to those from first turn IHC (Patuzzi and Sellick 1984) and basilar membrane (Patuzzi et al. 1984b) responses. They are also consistent with masking period patterns measured psychophysically by Zwicker (1977).

The neural literature suggests that modulation is observed only for tones in the tip regions of single unit tuning curves. In contrast to these reports, very limited results from the base of the cochlea by Patuzzi and colleagues (Patuzzi and Sellick 1984; Patuzzi et al. 1984b) indicate that modulation for low frequency inputs can occur but in the opposite direction to that observed at the CF of the cell. Results in Fig. 1C are consistent with this demonstration. Here the CM response to a probe at 280 Hz is measured in the presence of the 40 Hz tone. The waveform illustrates that magnitude reductions are associated with the positive phase of the CM during the time when the basilar membrane is displaced to SV.

In order to better understand this frequency dependence, additional information was collected from the OC where recordings are stable over relatively long periods of time. These measures demonstrate that the gross dc response, the summing potential (SP), is bipolar (Dallos et al. 1972; Cheatham and Dallos 1994) such that negative hyperpolarizing dc shifts are produced at low stimulus frequencies. Positive depolarizing responses occur in the CF region around 3500 Hz. By varying probe frequency, it was determined that probe tones associated with the depolarizing phase of the SP response are reduced during basilar membrane displacement to ST. Low frequency probes that produce hyperpolarizing dc shifts are decreased in phase opposition, during the time when the basilar membrane is displaced to SV. However, as level of the modulating tone increases, magnitude reductions are measured for displacements in both directions independent of probe frequency.

CONCLUSION

The fact that the direction of modulation for CM responses is associated with SP polarity suggests that the frequency dependence of the magnitude changes may be linked to the asymmetries responsible for generation of dc responses by cochlear hair cells. Further study may shed light on the mechanisms of hair cell transduction and on the ways in which receptor populations interact within the peripheral auditory system. Since modulations well below CF have not been adequately investigated, it may also be helpful to search for counterparts at both mechanical and neural recording locations.
Fig. 1A. A 3500 Hz probe tone is presented at 40 dB in combination with a 40 Hz tone at 75 dB. The IHC ac receptor potential has been high-pass filtered to remove the response to the low frequency tone in order to illustrate phase-dependent changes in magnitude. The CM recorded from the OC fluid space near the IHC is also included. Displacements to ST (SV) are associated with the negative (positive) phase of the CM response. The magnitude scale on the left applies to the probe; that on the right to the CM response at 40 Hz. 1B. Companion results for the CM at 3500 Hz. 1C. Results, similar to those in 1B, except probe frequency is 280 Hz.
ACKNOWLEDGEMENTS

Work supported by NIH-DIDCOD Grant DC00089.

REFERENCES


TEN YEAR TEST OF THE NORWEGIAN REGULATIONS RELATING TO NOISE AT THE WORKPLACE

Gordon Flottorp
Rikshospitalet, National Hospital, University of Oslo, Oslo, Norway

SUMMARY
A noise project was started in 1983 in a tobacco plant (JLT), in which 250-300 employees were exposed to factory noise about the highest limits allowed according to the Norwegian regulations relating to "Noise at the Workplace" (1982), in order to study whether these limits serve the goal: No noise induced hearing loss (NIHL) shall be acquired at the workplace. In the start and in the eleventh year an exhaustive examination of the noise level at all working places was carried out. Every year through 11 years extensive audiometric hearing control, supplied with Temporary Threshold Shift (TTS) study of 55 selected employees, was performed of all noise exposed employees.
Methods and results are reported. The conclusion is: When the equivalent level for 8 hours workday \( L_{eq} \) does not exceed 85 dB(A) and the max. sound level exposed to never exceeds 110 dB(A) FAST, the chance for acquiring NIHL is approximate zero.

INTRODUCTION
A thorough examination of the noise situation in a Norwegian tobacco plant (JLT) in 1982 indicated the necessity to carry out a hearing control of the employees, according to the Norwegian regulations of 1982: "Noise at the Workplace". A contract between the University of Oslo and JLT was agreed upon in 1983 with the aim to control whether "Noise at the Workplace" does secure that none of the employees acquire noise-induced hearing loss during work. The control of hearing of the employees necessary for the project secured that JLT fulfilled the requirements described in the regulations regarding control of hearing of the employees.

The program for the noise project was simple: First a new thorough analysis of the noise situation directed by the project leader (the author), measuring the equivalent and the maximum noise levels at all actual working places and secondly, starting the yearly audiometric control of hearing in the frequency range 125-8000 Hz (11 frequencies) of all the employees that JLT found necessary in order to comply with "Noise at the Workplace" Chapter VI, even with a better margin of safety than required in this act. In addition it was found desirable to carry out temporary threshold shift (TTS) measurements on about 55 selected employees. A room, satisfying international standards for audiology had to be provided, located in a central part of the plant.

The project started in the fall 1983 and was at first considered to last about three years. We learnt, however, from our own experience and from the literature that a duration of about 10 years might be necessary in order to obtain sufficiently reliable results.
METHOD

A. Noise measurements.

The work at different working places, together with the different machines, were studied in order to determine the actual measuring sites. It ended up with around 150 measuring places. The measurements were carried out using specially designed sound measuring equipment from Norwegian Electronics (Norsonic), the microphone mounted on a tripod and the amplifier/calculator registration outfit placed at secure places, not disturbing the work. In addition some separate measurements were made using the Bruel & Kjær precision sound level meter, type 2209. The weighting curve A and the time constant FAST always were used.

The Norsonic equipment determined automatically the equivalent sound levels for 15 min intervals, stored the results and printed them out.

Later these data were used in calculation of the 8 hours equivalent sound level (nominal equivalent level), using the information about the employees' eating and smoking pauses. Special attention was paid to the maximum sound levels, since the Norwegian regulations beside the limit for nominal equivalent level, contain the following sentence: "... no employee shall be exposed to sound level A exceeding 110 dB - FAST".

Such high maximal levels might be caused by:

1. Use of compressed air for cleaning part of the machinery
3. Technicians doing special adjustments or repair of the machinery.

In order to obtain precise values for such maximum levels, it was arranged occurrence of these noise types.

The first thorough analysis of the noise situation took several weeks. Later the noise was measured from time to time at working places having more "dangerous" noise levels, using hand-hold noise level meter.

In the eleventh year of the noise project (April 1994) a new thorough measurement and mapping of the results at the different working places was carried out, using hand-hold sound level meters, registering the equivalent level for 18 seconds and repeating the measurements when the working situation indicated this. Norsonic sound level meter, type SLM 116 and Bruel & Kjær, type 2230 were used.

B. Control of hearing

The Norwegian regulations, chapter VI states: "The employer shall carry out audiometric control of hearing of all employees who are exposed to nominal equivalent sound levels exceeding 80 dB or to a maximum sound level A exceeding 105 dB" and further: "Initial control shall determine the hearing threshold at the frequencies 500, 1000, 2000, 3000, 4000 and 6000 Hz. Regular control of hearing shall be carried out annually at least at the frequencies 3000, 4000 and 6000 Hz".

In our noise project we wished to be on the safe side of these requirements, and therefore perform audiometric control of hearing of all employees in all departments in which some employees were exposed to noise levels about the limits specified in the regulations.

We also decided to carry out the hearing control every year for the 11 standard frequencies from 125 Hz to 8000 Hz (750 and 1500 Hz in addition to the series in the regulations).

Threshold measurements were of two types, always using MAICO MA 20 audiometer, which every year was controlled and calibrated, using the IEC artificial ear (IEC 318):

I. Pure tone audiometry in the morning carried out before the start of work (one employee), continuing during the first two hours, the employees using hearing protectors until leaving the work place and going to the quiet audiometry room.

A new group of employees starts work in the afternoon. The hearing threshold of this group is measured in the same way as for the employees starting at 06.45. All employees in
this afternoon group have been at the working place using hearing protectors before coming to the audiometry room, being exposed to noise for 25 min to 2 2/3 hours.

After a short explanation of what should be done in order to control the hearing and how to respond etc., eye glasses and larger earclips were removed and the earphone was carefully placed by the operator. The thresholds were determined for the standard (ISO 11 frequencies mentioned (125 Hz - 8000 Hz), starting at 1000 Hz, proceeding with 500, 250, 125, 200 ... - 8000 Hz, concluding at 1500 and 750 Hz. The ascending method was used, in some cases supplemented with the bracketing method (ISO 8253). The operator and the employee were seated in the same room, on each side of the audiometer table, facing each other.

The subject had been instructed to raise the forefinger whenever hearing a tone, the right hand forefinger when hearing the tone in the right ear, and the left when hearing the tone in the left ear.

Cerumen was removed and otoscopy carried out, when audiometric findings indicated this to be warranted.

II. For temporary threshold shift (TTS) investigation the same procedure for threshold determination as described under I was used. The selected 55 employees (various age and hearing loss, although mainly normal hearing persons of both sexes) for this examination were instructed to leave their working place at a given time, and go fast to the audiometry room in order to start the threshold determination 2 min after leaving the noise area, - no hearing protection used.

The audiometric results were in the usual way reported in audiograms. These were carefully examined and stored.

Audiograms from the regular hearing control were stored separately for each of the 11 departments of the plant, and the TTS-audiograms as a separate group.

RESULTS

I. Noise measurements

The difference between the two main noise analyses is very small, and may be explained as a result of the different methods for placing the microphone of the sound level meter. In the first measurements the microphone usually was placed on a tripod, whereas the last analysis was carried out with hand-hold sound level meters, which permitted a closer measurement to some of the production machineries.

The median value was 81.1 dB(A) in 1983/84 and 81.2 dB(A) in 1994, - expressed as the nominal equivalent noise level $L_{Aeq}$ dB(A), and calculated based upon 147 measuring points in 1983/84 and 144 in 1994. In table I the results are presented in three different intervals of 5 dB size. It is concluded that the actual noise situation has been fairly stable through the 10-11 years of examination.

<table>
<thead>
<tr>
<th>Nominal equivalent sound level $L_{Aeq}$ dB(A)</th>
<th>I $L_{Aeq}$ ≥85</th>
<th>II 80≤$L_{Aeq}$&lt;85</th>
<th>III 75≤$L_{Aeq}$ &lt;80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of measuring places 1983/84: 147</td>
<td>4 (3%)</td>
<td>104 (71%)</td>
<td>39 (26%)</td>
</tr>
<tr>
<td>1994: 144</td>
<td>19 (13%)</td>
<td>78 (54%)</td>
<td>47 (33%)</td>
</tr>
</tbody>
</table>

Table 1: Noise data from 1983/84 and 1994 distributed in three 5 dB intervals noise levels.

II. Control of hearing

Each year the result of the audiometric hearing control was presented for each of the 11 departments in which the noise situation required hearing threshold control. Each audiogram
was carefully analysed and classified into one of the following five categories, after being
corrected for age dependent hearing loss (ISO 7094)
N: Normal hearing i.e. no threshold point below 10 dB hearing loss.
N+: Practically normal hearing i.e. max hearing loss 20 dB at two frequencies.
L1: Hearing loss from 25-40 dB for one or more of the frequencies 3000, 4000, 6000 Hz,
or 20 dB loss for all these frequencies.
L2: Hearing loss at one or more of the frequencies 3000, 4000, 6000 Hz greater than 40 dB
and the hearing loss at 2000 Hz not greater than 20 dB.
L3: As for L2, but hearing loss for 2000 Hz greater than 20 dB.
P: Hearing loss of other pathology than noise induced.

In this rather short publication it has been necessary to present results for all the
departments together. It is, however, very important in the practical control of noise exposed
people to follow the development of possible hearing loss separately for each department, and
of course for each employee.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>57</td>
<td>57</td>
<td>54</td>
<td>51</td>
<td>51</td>
<td>52</td>
<td>56</td>
<td>52</td>
<td>52</td>
<td>49</td>
<td>50</td>
</tr>
<tr>
<td>N+</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>23</td>
<td>24</td>
<td>21</td>
<td>25</td>
<td>25</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>L1,</td>
<td>15</td>
<td>14</td>
<td>16</td>
<td>16</td>
<td>17</td>
<td>16</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>L2,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>No</td>
<td>232</td>
<td>241</td>
<td>259</td>
<td>256</td>
<td>257</td>
<td>264</td>
<td>272</td>
<td>315</td>
<td>307</td>
<td>294</td>
<td>272</td>
</tr>
</tbody>
</table>

Table 2. The number of employees tested each year, and their percentage distribution in the
four categories hearing status.

The result for 1984 is presented in the table, although it does not represent exactly the
same departments as in the following years. The time 1983/84 was more or less a trial period.

As it appears from the table, the percentage distribution of employees upon the different
categories has been rather stable through the period of the project. The percentage of noise
induced hearing loss has varied between 14% and 17%. A variation of ±1% may be caused
by the changing of employees, either beginning or leaving the plant. The increase from 1993-94
is f.inst. due to a high percentage of noise induced hearing loss among the 10 new
employees in 1994.

The impression conveyed by this rather rough summarizing about a stabilized hearing
situation among the employees, is confirmed by the detailed study of the individual
audiograms. A further proof is the fact that for the 44 employees having their hearing
measured 11 years, the percentage NTHL (L) in 1994 was 16%, and for the group 11,10 and
9 years (128 employees) 13% (L).

CONCLUSION
The conclusion is therefore that the noise situation at JLT, which for about 230 working
places is fairly close to the Norwegian noise limits for working places, has not caused a
significant number of new or increased noise induced hearing loss during the last 10 (11)
years. It must be emphasized that the employees have obeyed the rules laid down in "Noise
at the Workplace", guided by an efficient main safety officer and his staff.
DO INSERT EARPHONES REALLY MAKE A DIFFERENCE IN ACHIEVED TRESHOLDS?

R. W. Gatehouse, Wilma Veenhof
Psychology department, University of Guelph, Guelph, Ontario, Canada

SUMMARY

This study compared audiometric thresholds using TDH-39P and ER-3A transducers under different ambient levels as associated with different test venues. The findings indicated that TDH-39’s yielded significantly lower thresholds when noise was <35 dBA, while inserts were better at 55 dBA. No threshold variability was noted for the two transducers. Practically however, the differences were so small and within the Hughson-Westlake (+-5 dB) error range which may render them inconsequential.

INTRODUCTION

Since the advent of "insert" earphones (Killion, 1984) studies have attempted to determine if they provide reliable, equivalent, lower/ or more accurate thresholds than Standard's accepted supraaural/ or other headphones in various noisy backgrounds (Clark & Roesser, 1988; Berger & Killion, 1989; Lindgren, 1990; Stuart et al., 1993). Some studies have also addressed a number of other audiometric practice issues some of which have been reviewed in Zwillocki et al., 1988, and Sataloff & Sataloff, 1993. As examples, issues such as attenuation characteristics of headphones (Killion et al., 1985; Hosford-Dunn et al., 1986; Sklare & Denenberg, 1987), and step sizes in threshold testing (Jervall & Arlinger, 1986) have been looked at. Additionally, the issue whether valid and reliable threshold testing always needs to be done in "sound-proofed booths" given "permissible ambient " has resurfaced given that "inserts" are claimed to provide some degree of attenuation that might obviate that necessity. Whatever their foci, the results of such studies have been somewhat inconsistent, and with some exceptions (e.g Clark & Roesser, 1988; Stuart et al., 1993) treatment conditions (orders; earphone fitting procedures) have not always been clearly detailed, and in some the statistical evaluations have not been rigorous. The present study was yet another attempt to consider some of these things under as rigorous conditions as possible.

METHOD

SUBJECTS

Eight females and 10 males, selected to encompass a wide age range (21-67; mean 34.0 years). Three of the older males exhibited untreated mild to moderate high frequency hearing loss from 2.0 to 6.0 kHz. Eight subjects indicated they had worked in noisy environments but none had used any hearing protection.

APPARATUS

ER-3A, 10 ohm insert tubephone (Etymotic Research) fitted with E-A-RLINK foam eartips (Cabot Safety Corp.), and TDH-39P (10 ohm) supraaural earphone fitted with P/N S1 cushions and held in a spring loaded headband (Telephonics Inc.). The headphones, were alternately connected according to the test condition to a PC PAD (FCAD I audiometer) controlled by PAR™ software (Hear - Wall Technologies, Inc.) which, with no operator, delivers selected tones, adjusts levels according to S's' responses that they detect/not the stimulus, and records responses.

Background noise was generated by Lafayette (models 1432; 1421) white noise system whose outputs were fed to a
TDH-39P speaker mounted on a height adjustable stand which allowed alignment with seated subjects' ears. The stand was placed at 180° and at either 1.90m or 1.05m for the "medium" and "high" noise levels conditions respectively. Room ambient and WN levels (dBA) were measured prior to test conditions with a Pulsar meter (model 83) set for "slow" response.

Three test venues were used. For "low" (<30 dBA) ambient conditions subjects, the audiometer and white noise delivery system were in an Eckel booth. For "medium" conditions this equipment and Ss were in a sound-treated lab with ambient raised via white noise to approximately 35dBA. The computer and E were in ante-room areas in both of these conditions. For "high" noise backgrounds all equipment and the E were in an office, isolated from major traffic and other possible noise intrusions and background levels of 55 dBA were maintained.

**DESIGN AND PROCEDURE**

The study involved 2 earphone types by three test venues/noise conditions repeated measures factorial design i.e. all Ss tested in all six conditions. An "incomplete counterbalanced order" according to the balanced "pairwise sequences" method (Keppel, Saufley, & Tokunaga, 1992) was used to determine the sequences of test venue and transducer type orders for each S. Thus after completion of a noise exposure and otological history questionnaire, and instructions in a separate office area a S was taken to a predetermined test room, seated and fitted with the appropriate ER-3A or TDH-39P transducers. The developer's instructions for deep insertion of the ER-3A's tubephones and foam ear-tips were followed, and all other standardized (e.g. glasses and earring removals, hair pushed away from ears etc.) testing were observed.

Testing used Sataloff & Sataloff's (1993) frequency order i.e. 1.0, 2.0, 4.0, 6.0, 0.25, and 0.5 kHz, with a 1.0 kHz practice trial for each ear before threshold determinations. The left ear was tested first. Two, three threshold runs, separated by "no more than" one week were given to accommodate the volunteers schedules. Ss were asked as much as was possible to avoid "high noise areas and activities" during the inter-test period. Since instructions emphasized that Ss concentrate in order to achieve the lowest thresholds possible, an optional 5 minute break was offered between the second and third threshold run on each day.

**RESULTS**

A t-test on mean thresholds for each ear collapsed across frequencies was non-significant (t=0.65; df=17; p>.05). Subsequently, the frequency means across subjects were averaged and this value was used in the ANOVA. Table 1 shows that achieved mean thresholds and standard deviations varied little except for the supraaural headphones under high noise and that overall the headphone types varied by slightly less than 1.0 dB. (Inserts av. 11.28; TDH-39P av. 12.2). The ANOVA

<table>
<thead>
<tr>
<th>Noise Level (dBA)</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Mean by Headphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inserts (ER-3A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>10.79</td>
<td>10.42</td>
<td>12.64</td>
<td>11.28</td>
</tr>
<tr>
<td>SD</td>
<td>8.78</td>
<td>9.03</td>
<td>9.39</td>
<td></td>
</tr>
<tr>
<td>Supraaural (TDH-39P)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>9.11</td>
<td>8.74</td>
<td>18.76</td>
<td>12.20</td>
</tr>
<tr>
<td>SD</td>
<td>8.88</td>
<td>9.04</td>
<td>9.51</td>
<td></td>
</tr>
<tr>
<td>Mean by Noise Level</td>
<td>9.95</td>
<td>9.58</td>
<td>15.7</td>
<td>11.74</td>
</tr>
</tbody>
</table>

(Table 2) indicated no main Headphone effect (F<sub>1,17</sub> =0.81; p>.05) but significant Noise and a Noise by Headphone interaction effects (F<sub>2,34</sub> = 53.8; p<.01; F<sub>3,42</sub> = 37.8, p<.01). Further One Way repeated measures ANOVA's (not shown) to separate these effects (see means Table 1) indicated that under Low and Medium noise conditions supraaural thresholds were significantly lower than with inserts, but under High noise backgrounds were poorer. As well, the simple
noise effect was significant for both headphones types. Pairwise comparisons for Low and Medium noise for each headphone were not different, so the means were combined and compared to the High noise means. For both headphone types thresholds were significantly elevated under High noise (supraural F = 80.4; p < .01; insert F = 7.5; p < .05). That is, regardless of headphone type High noise produces significantly higher thresholds. Finally, separate t-tests on the two headphones at all frequencies affirmed the manufacturer supplied insert correction factors did equate thresholds to the TDH-39's i.e. from +5.6 at 0.25 kHz to -8.2 at 6.0kHz.

Table 2. Summary Table for Repeated Measures ANOVA on Headphone Type and Ambient Noise Level.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headphone</td>
<td>9.252</td>
<td>1</td>
<td>9.252</td>
<td>0.807</td>
<td>0.382</td>
</tr>
<tr>
<td>error</td>
<td>194.849</td>
<td>17</td>
<td>11.462</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise</td>
<td>919.404</td>
<td>2</td>
<td>459.702</td>
<td>53.822</td>
<td>&lt; .01 **</td>
</tr>
<tr>
<td>error</td>
<td>290.398</td>
<td>34</td>
<td>8.541</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Headphone * Noise</td>
<td>418.029</td>
<td>2</td>
<td>209.015</td>
<td>37.836</td>
<td>&lt; .01 **</td>
</tr>
<tr>
<td>error</td>
<td>187.822</td>
<td>34</td>
<td>5.524</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** indicates significant

CONCLUSIONS

In ambient of 35 dBA or less, supraural headphones yielded significantly lower thresholds than inserts, but at 55 dBA the reverse was found. Two factors lessen the impact of these observations. First, threshold variability, as in several previous studies e.g. Stuart et al., (1991) for the headphones was not different. Second, in practical terms the mean differences under Low and Medium backgrounds was very small (i.e. < 2 db HL), while under High noise it was still only about 6 dB HL with an advantage for inserts.

These observations seem to be relevant to several threshold testing practice issues. First, the standard practice of determining thresholds via the Hughson-Westlake method may preclude a determination of the superiority of particular headphone types unless the background noise levels are considerable. Exactly where any threshold advantage passes the ± 5 dB error range implicit in that method was not examined here since testing was not done at levels between 35 and 55 dBa backgrounds but it is currently being studied. Nonetheless, even here the mean insert advantage (approx. 6.1 dB HL) under the highest background is still only minimally outside such a method imposed 5 dB error and it is difficult to strongly endorse that it is an "advantage". In the same vein, these data seem to once again raise the issue raised as far back as Harris & Meyers (1954) and again by Harris(1978) whether the Hughson-Westlake method itself is viable, and it is apparent that some in the audiological community still interpret 10 dB shifts and differences as being significant.

Second, the issue whether sound-proofed enclosures are required for reliable/ valid pure tone audiometry would seem to follow directly from the foregoing. Simply, the obtained mean threshold differences (Table 1) were only outside the ± 5 dB step under "high" ambients. Such ideas are implicit in the articles by Jervall & Arling.1986 and Stuart et al., 1991. Finally, the results here are in sharp contrast to Clark & Roesser's (1988) where ER-3A advantages of 11 - 20 dB SPL across the same frequency range were reported, although the comparison was to TDH-50's. Explanation of two studies's results differences is not readily seen since presumably those researchers tested under permissible ANSI standards ambients whereas the ambients used in this study should have caused upward shifts especially at lower frequencies where ANSI permits least noise. In fact, St here (exception being "high" noise supraural condition) achieved thresholds as low as -10dB HL for the frequencies below 1.0kHz. Thus both headphone types are able to attenuate ambients to permit testing down to -10dB HL except at very high levels, and maybe ANSI permissible ambients need not be as low as they are.
REFERENCES


EXAMINATION OF OBJECTIVE EVALUATION OF BINAURAL ANALYSIS

Klaus Genuit
HEAD acoustics GmbH, Kaiserstr. 100, 52134 Herzogenrath, Germany.

SUMMARY

Aurally adequate sound measurement technology makes use of both present psychoacoustic knowledge, e.g. loudness, roughness, fluctuation, sharpness and so forth, and Artificial Head measurement technology with transmission characteristics comparable to human hearing. By taking into account psychoacoustic evaluation parameters very often good results for judging sound events with regard to sound quality, as compared to subjective impressions, can be achieved. If a sound situation is relatively complex, i.e. if it consists of various single sound sources at different spatial positions, significant level and phase differences between the left and right ear occur which - in comparison with a monaural evaluation - can yield different results. Such effects have been observed for some time already. Speech intelligibility in a noisy environment, e.g., depends on the positions of the sound sources. Furthermore, investigations into noise in workplaces showed that binaural recordings as compared to monaural recordings caused significantly different physiological reactions of the test subjects. Moreover, it became evident that noise annoyance of single sound sources in a complex mixture of sounds does not only depend on absolute parameters such as A-weighted SPL, loudness etc., but also on their localization. In the interior of a car, for instance, the individual engine orders can produce significant amplitude and phase differences between the left and right ear, due to wave and multipath propagation, influencing the subjective evaluation. There are no objective measurement procedures available at present. In the last decades a lot of scientific basic research was done on binaural signal processing. Due to the complexity of signal processing in human hearing there have been no complete models yet which simulate binaural signal processing in a simple way and thus constitute a basis for the objective determination of sizes, derived from a binaural measurement to describe subjectively perceived sound quality. Therefore, investigations aiming at the selection of certain signal components from a complex mixture of sounds were carried out, based on additional information such as directional sound incidence of the sound sources under test or engine r.p.m. for selecting individual engine orders. On the basis of known models to predict speech intelligibility in a noisy environment, binaural models for the objective determination of sound quality, based on these simple application examples and Artificial Head recordings, were developed.
INTRODUCTION

Acoustic recordings are made binaurally with the Artificial Head. The psychoacoustic evaluation of the recorded sounds has so far been made individually for each channel, not considering the data of the respective other channel for calculations. This procedure contradicts the requirements of aurally-accurate analysis technique. Man does not distinguish between left and right sensation, but perceives only one sound coming from both ears. Therefore, a method is required to combine both channels of the recording. To this end, a very simple method is being investigated at present, based on the addition of both channels. As usually the delay times of the signals differ, a corresponding cancellation is added to one channel, aiming at a constructive superposition of the ear signals. This method was used for recording engine sounds.

ORDER ANALYSIS INSIDE OF A VEHICLE

The following is an analysis of a recording of the noise in the interior of a vehicle made using the artificial head measuring system located at the front right-hand seat. An acceleration from 1,000 rpm to 6,000 rpm was carried out in third gear. All test engineers made the observation that in the range above 4,000 rpm the noise in the interior of the vehicle clearly changed for the worst. Fig. 1 shows the A-weighted sum level, the loudness and the roughness over rpm. The analysis is not able to give any significant explanation of the fact that the noise caused by the engine in the vehicle interior in the range between 4,000 and 5,000 rpm was evaluated as particularly unpleasant.

Fig. 2 highlights the individual order levels of the left and right ear for the 2nd and 4th order within the rpm range objected to. An examination of Fig. 2 reveals clear differences between the A-weighted level of the 2nd and 4th order, as well as very clear left/right differences not apparent in the sum level in Fig. 1. Extreme left/right differences of this kind create a particularly unpleasant effect on the ear at levels of up to approx. 15 dB in the frequency range below 500 Hz. This is because in free-field exposure such level differences cannot occur. In general, our auditory sense reacts to such unaccustomed auditory events by producing an unpleasant pressure on the ears. In the interior of a vehicle, however, a sound field distribution can be produced which shows nodes and bulges in the sound pressure distribution, resulting from, on the one hand, standing waves due to air excitation, but also to the overlay of air excitation with the simultaneous solid-borne sound from the drive unit to vehicle body coupling. Depending on the geometry of the vehicle these nodes may be created at specific frequencies in such a way that they become located at the mid-point of the head. In this case, the signals left/right of the ear would be in phase opposition. This can produce level differences of approx. 17 dB, depending on the position of the head and the structure of the nodes and bulges in the standing waves.
Fig. 2: Interaural level differences of the 2nd and 4th order

To confirm our hypothesis about the extent to which these effects are relevant in subjective evaluation, we next carried out the following auditory tests. Firstly, the 2nd, 4th and 6th order were completely removed. Next, only the signal components of the individual orders were selected out from the original signal and, after appropriate processing through adjustment of the level differences, re-mixed with the signal minus orders. Thus, three signals were produced: firstly, the original signal; secondly, the original signal without the 2nd, 4th and 6th order, and thirdly, the original signal again, but with reduced the interaural differences for the 2nd, 4th and 6th order. Subjective evaluation placed the 3rd noise unmistakably between the two extremes, i.e. the original noise including all interaural differences, and the original noise minus the 2nd, 4th and 6th order. This signifies that the presence of extreme level, and also phase, differences in the individual spectral ranges results in an unfamiliar auditory impression and is auditively considered as having unwanted noise quality.

BINAURAL PROCESSING MODEL

The principal idea of binaural signal processing is shown at fig. 3. The incoming signals of the left and right ear will be received by the human internal filter system which has an adaptive characteristic to change the midfrequency of the filter to the significant spectral components of the signal. We can assume /1/ that the human hearing is able to process around 22 such filters to cover the whole frequency range from 20 Hz to 20 kHz. The question is how the human hearing is able to optimize the signal to noise ratio at each frequency group. The human hearing is able to improve the signal to noise ratio depending on the incidence of sound. A good approach is given by VOM HÖVEL /2/ based on the EC (equalization and cancellation) model of DURLACH /3/ and is shown in fig. 4.

Fig. 3: Overview block diagramm of binaural signal processing

Here the principal binaural processing at one of the 22 receiver filters is demonstrated. The idea is that the left and right ear signal will be manipulated by changing the interaural level and phase in such a way that the signal to noise ratio will be increased. The model consists of a variable time delay and a variable attenuation in combination with a randomized error for the phase and amplitude. Thus it is possible to consider the fact that when you listen to two different signals with headphones where one signal is the same on both ears and the other signal is completely uncorrelated the human hearing is able to increase the signal to noise ratio only to a maximum of around 12 to 15 dB /4/.
The sound described at fig. 1 was analyzed with the model according to fig. 4 and the result is shown in fig. 5 in comparison to the A-weighted level of the left and right earsignal only. The influence of the level and phase differences at the different orders to the subjectively received level is clearly shown. The curve of the binaural level correlates well with the subjective evaluation of the sound by listening tests. That means the binaural model which is well known for the calculation of speech intelligibility in a noisy environment is a good first approach for a binaural order analysis of sound inside cars based on an artificial head measurement.

LITERATURE

COMPUTATIONAL MODELLING OF PSYCHOACOUSTIC COMBINATION TONES AND DISTORTION-PRODUCT OTOACOUSTIC EMISSIONS

Christian Giguère 1, Hans Kunov 1, Guido F. Smoorenburg 2
1 Institute of Biomedical Engineering, University of Toronto, Toronto, Canada
2 Dep. Otorhynolaryngology, University Hospital Utrecht, The Netherlands.

SUMMARY

This paper reports on the modelling of psychoacoustic and otoacoustic auditory distortion products elicited by two closely-spaced stimulus tones $f_1$ and $f_2$. The model reproduces the general level behaviour observed in psychoacoustic experiments for the $2f_1-f_2$ combination tone, in particular the higher levels found with a cancellation-tone rather than a loudness-matching procedure. The growth of the $2f_1-f_2$ otoacoustic emissions at the eardrum of the model also has a shape significantly different from that of the two corresponding psychoacoustic measures. The results may imply a complex relation between these different types of measures.

INTRODUCTION

Auditory distortion products of the type $f_1-n(f_2-f_1)$ have traditionally been studied as combination tones (CTs) measured psychoacoustically in cancellation-tone or loudness-matching experiments (e.g. [1,2]). More recently, they have been studied mostly as distortion-product otoacoustic emissions (DPOAEs) measured non-invasively in the ear canal using sensitive electroacoustic equipment [3]. Both phenomena are widely believed to have a common cochlear origin and appear to be closely related to the functioning of the outer hair cells (OHCs) [3].

The computational synthesis of both CTs and DPOAEs requires a peripheral auditory model with bidirectional outer ear, middle ear and cochlear stages. A recently published model [4] meets this criterion and contains distributed nonlinear elements derived from the fast motile responses of the OHCs. A parametric analysis of the model nonlinearity was conducted with the purpose of reproducing CT and DPOAE experimental data. This paper presents the results of a typical configuration of the model. The configuration is kept fixed throughout the paper in order to highlight the similarities or dissimilarities between the various methods of CT measurements, and between CTs and DPOAEs.

MODEL

The model is shown in Fig 1 and is described in detail in [4]. The circuit elements and sources to the left of the figure form an equivalent-circuit approximation for the sound diffraction effects of an acoustic wave laterally upon the human head. The incident wave $P(t)$ is then processed through a series of cross-coupled concha/ear canal, middle ear and cochlear stages. The latter is a nonlinear 1-D transmission line approximation of basilar membrane (BM) motion, discretized into $N=320$ channels from base to apex. The circuit elements $R_n$, $L_n$ and $C_n$ are analogous to the effective acoustic resistance, mass and compliance of the cochlear partition in
channel \( n \). The current \( I_n(t) \) is equivalent to the BM volume velocity. The model is implemented numerically in the time domain using the technique of wave digital filtering [6]. The sampling rate of operation is 71680 Hz.

In each BM channel, a voltage source \( V_{n \text{ohc}}(t) \) produces a nonlinear and saturating pressure, assumed to originate from the OHCs, as follows:

\[
V_{n \text{ohc}}(t) = G(I_n) \ast R_n \ast I_n(t),
\]

where \( G(I_n) = G \ast (1 + |I_n(t)|/I_0)^{-0.5} \)

and \( I_0 \) is an input scaling constant (3.6x10^{-3} cm/s) while \( G \) is the maximum feedback gain (0.99).

The choice of the nonlinear gain \( G(I_n) \) is made on the basis of previous psychoacoustic and modeling studies of CT generation [1,2,5], which indicated a signal-compressing type of nonlinearity such as a power-law with exponent between 0 and 1. Functionally, \( V_{n \text{ohc}}(t) \) reduces the damping of the BM at low levels and leads to level-dependent BM tuning curves.

RESULTS

Combination tones (CTs)

Figure 2a presents an example of the BM excitation pattern for a two-tone stimulus \( f_1, f_2 \). The horizontal axis is the BM place from base to apex. The excitation patterns to the stimulus components \( f_1 \) and \( f_2 \) greatly overlap near their place of resonance where they give rise to a relatively broad peak. The CTs of the type \( f_1-n(f_2-f_1) \), \( n = 1...4 \), have lower frequencies and thus travel further towards the apex where they are resolved as clear peaks at their place of resonance. The most prominent peak \( (n=1) \) corresponds to the commonly studied 2f1-f2 distortion product. The CTs of the type \( f_2-m(f_1-f_2) \) are also generated by the model but are not resolved on Fig 2a. They have frequencies higher than the two stimulus components and are masked by the excitation patterns due to \( f_1 \) and \( f_2 \).

Psychoacoustically, two main methods have been used to measure the level of combination tones [1,2]. In the cancellation-tone method, a probe tone at the frequency of the CT to be measured is presented externally and simultaneously with the stimulus components \( f_1 \) and \( f_2 \), and the task of the subject is to adjust its amplitude and phase until the pitch sensation of the CT becomes inaudible. The common interpretation of this method is that, at the cancellation point, the probe tone has the same amplitude as the internally-generated CT but is 180 degrees out of phase. Cancellation-tone experiments are reproduced in the model by adjusting the amplitude and phase of the probe tone until the peak in the BM excitation pattern to the corresponding CT frequency would disappear completely. As in psychoacoustical experiments, the model adjustments are quite
Figure 2: Combination tones from the model. (a) BM excitation pattern for stimulus \( f_1 = 1400 \text{ Hz} \), \( f_2 = 1680 \text{ Hz} \), \( L_1 = L_2 = 60 \text{ dB SPL} \). (b) Growth of the 2f1-f2 CT \((n=1)\) for \( L_1 = L_2 \).

sensitive to small variations in amplitude and phase, so that in general this procedure leads to a clear cancellation point. Typical model accuracy is better than 0.5 dB and 5 degrees.

A second psychoacoustic method consists of matching the loudness of the CT with that of a probe tone of the same frequency presented externally but non-simultaneously, e.g., by use of the pulsation-threshold technique [1,2]. This method is referred to here as the comparison method. It is easily reproduced in the model under the assumption that, at the matching point, the peak of the BM excitation pattern due to the probe tone is equal to that of the CT during the stimulus \( f_1, f_2 \) presentation.

Figure 2b presents the growth of the 2f1-f2 CT in the model with increasing stimulus level \((L_1 = L_2)\) using both measurement methods. Overall, the slope of the cancellation data is very close to 1.0 and tends to decrease slightly at higher stimulus levels in agreement with psychoacoustic data [2]. The model cancellation levels also quantitatively agree with the psychoacoustic data at mid and higher stimulus levels for an equivalent frequency ratio \( f_2/f_1 \). However, the slope of the cancellation data is slightly higher than 1.0 at lower stimulus levels (<40 dB SPL) and thus the cancellation levels tend to be underestimated by 5-15 dB in this region. The model comparison levels are lower than the model cancellation levels and tend to saturate faster at high stimulus levels as found psychoacoustically [2].

Distortion-product otocoustic emissions (DPOAEs)

Figure 3a presents an example of a spectrum of the model sound pressure \( V_d(t) \) at the eardrum position for a two-tone stimulus \( f_1, f_2 \). The distortion products of the types \( f_1-n(f_2-f_1) \) and \( f_2-m(f_1-f_2) \) can now be both clearly seen. Their amplitude relative to that of \( f_1 \) and \( f_2 \) is a function of the stimulus level.

Figure 3b presents the growth of the 2f1-f2 and 2f2-f1 DPOAEs in the model \((L_1 = L_2)\). Overall, the growth of the 2f1-f2 DPOAE shows a significantly lower slope (about 0.6) and is more irregular than the growth of the corresponding CT in Fig 2b. The 2f1-f2 DPOAE eardrum levels are about 60 dB below that of the stimulus components \( f_1 \) and \( f_2 \) for stimulus levels between 30 and 50 dB SPL. The difference increases at lower and higher stimulus levels. The model also shows a significant difference between the growth patterns of the 2f1-f2 and 2f2-f1 DPOAEs.
CONCLUSIONS

The model reproduces the difference observed psychoacoustically between cancellation-tone and loudness-matching estimates of the CT, and successfully predicts an increasing gap with stimulus level for a fixed frequency ratio f2/f1. Further analyses of the BM patterns have revealed that the difference between the two methods is closely related to suppression effects by stimulus f1.

The growth curve of the 2f1-f2 DPOAE is significantly different from that of the two corresponding CT measures. Further analyses are necessary to investigate whether the differences imply that DPOAEs and CTs represent nonlinear activity generated in different parts of the BM or whether they are due to different modes of propagation after generation. Extension of the model to the active case G > 1 is also considered to further improve the agreement between model and experimental data for CTs and DPOAEs.

REFERENCES


A STUDY INTO ASYMMETRIC HEARING LOSS AMONG A TRUCK DRIVER POPULATION ON THE SUNSHINE COAST IN QUEENSLAND

Beno Groothoff
Division of Workplace Health and Safety, Brisbane, Australia.

SUMMARY
The results of the noise exposure measurements in this study indicate that for all but one of the 20 rides exposures were in excess of the exposure limit. The drivers' right ears were exposed to higher noise levels which, in six instances were more than double the sound intensity of their left ears. These higher exposures were caused by wind noise and suggest that given enough exposure time this would lead to asymmetric hearing loss.

No pre-employment audiometric records were made and only recent records of 9 out of the 20 participating drivers were available. They could however not be used to determine to what extent the right ear was worse affected than the left ear as they were made after a working day on the truck and temporary threshold shifts would probably have occurred.

The results of the noise exposure measurements indicate that the potential for asymmetric hearing loss is definitely present. Statistical analysis by means of matched pair comparison tests and t-test of the left and right ear exposure measurement results confirmed the potential for asymmetric hearing loss as the difference between left and right ear exposures was found to be significant, (P<0.05) and SD 1.54486.

Truck drivers normally have longer working hours than most other working people, which in order to maintain an exposure which does not exceed $L_{eq, 8h} 85 \text{ dB(A)}$ reduces the allowable exposure level to 84 dB(A) for ten hour shifts and 83 dB(A) for 12 hour shifts. Although the sample size in this study was small with only 20 drivers, the excessive noise exposures obtained, despite relatively new trucks, causes concern for the participating drivers and for drivers of trucks of an older design which would be noisier.

All but one of the rides resulted in excessive noise exposures, either outright or when the duration of the exposures were taken into account. Considering that the company in this study operates predominantly with more modern and therefore quieter trucks this finding is considered a cause for concern as it can be directly related to the larger trucking community.

INTRODUCTION
The aim of this study was to investigate the possibility of asymmetric hearing loss occurring in truck drivers and whether this could be attributed to external noise, such as the wind rushing past the open drivers side window at high speeds. As road traffic in Australia drives on the left hand side of the road their right ear was expected to be worse affected than their left ear. The study examined a truck driver population of 20 male
drivers of a company which operates a fleet of 42 Tonne sand and gravel trucks. The company is based on the Sunshine Coast in Queensland. The drivers normal duties involved a fair proportion of highway driving at speeds up to 100km/hour. For the investigation a reliable method had to be developed which enabled the noise exposures at the right and left ears to be measured simultaneously given that the right ear was expected to be exposed to potentially high wind velocities and associated noise.

Trucks are considered workplaces under the Workplace Health and Safety Act and noise exposure of truck drivers has become even more important with the introduction on 3 January 1994 of a lower noise exposure limit specifying among other things, a level of noise of 85 dB(A) as an average over an eight hour period which should not be exceeded.

**METHOD AND RESULTS**

The method consisted of two simultaneously operating noise dosemeters attached to a driver with the microphones positioned at the top of the driver’s left and right shoulders in accordance with the findings of a study by Muldoon (1973). The representativeness of this position in relation to the ear was checked by conducting a test under highway driving conditions with one microphone at the right ear and one microphone at the right shoulder position. The results of the two noise dose meters over a period of 2 hours and 13 minutes of predominantly driving under highway conditions were within 0.1 dB(A) of each other. The use of the shoulder position was therefore validated.

The literature dealing with wind effects on the microphones of sound level meters indicates that the integrity of the microphones becomes affected by wind at speeds above 5 m/s (20 Km/h), (AS 2659). According to studies by Hassall & Zaveri (1988) these wind effects start to take place at about 40km/h. Wind speed measurements in the truck cabins ranged between 0.55 and 1.62 m/s (1.98 - 5.83 km/h) at the position of the driver’s right ear with higher wind speeds at the back wall of the cabin. Where the wind enters the cabin speeds ranged between 1.43 and 3.70 m/s (5.15 - 13.3 km/h). Measured along the back wall wind speeds ranged between 1.88 and 2.70 m/s (6.76 - 9.72 km/h). All results were well below both the above mentioned limits. In addition the microphones were equipped with windscreen to protect the sensitive microphone diaphragms against quarry dust and possible interference with the correct operation of the microphone. The use of the shoulder position was therefore again validated.

Audiometric test records of only 9 out of the 20 drivers were available for this study and found to be highly questionable due to the fact that the tests were conducted after a days work on the trucks and, based on the results of the noise exposure surveys which indicated excessive noise exposures, temporary threshold shifts would probably have occurred. It was therefore decided that these results could not be used in this study. Because of this decision the extent to which the right ear is worse affected than the left ear, could not be established.

Not being able to use the audiometric data had also altered the method of statistical analysis from multiple regression analysis of the information from the noise exposure surveys, audiometric tests and questionnaires, to matched pair comparison analysis and t-test of the information from the left and right ear exposures obtained during each ride.

The results of the statistical analysis concluded that there was a significant difference between the left and right ear exposures and therefore the null hypothesis that there was
no difference in hearing loss potential had to be rejected. This finding confirmed the results of the measurements showing excessive noise exposures on all but one occasion.

Field data was collected during 20 rides lasting up to 12 hours each per shift during which the author was present as an observer for the full duration of each shift every time. Observations consisted of the taking of notes on truck and driver activities occurring. The noise dose meters were switched on at the same time at the beginning of the shift and switched off at the end. This would enable identification and comparison of all events occurring for the left and right ears. After each survey the data stored in the noise dose meter was downloaded via an interface into a personal computer where it was analysed.

The $L_{Aeq}$ dB(A) results of the noise level exposures were used as the parameter for the determination of any differences between left and right ear exposures and compared with the noise exposure limits of the Workplace Health and Safety Regulation 1989. The Standard Deviation was calculated to indicate any spread of the readings. Next matched pair comparison tests were conducted on the 20 sets of left and right ear exposures and finally the t-test was applied to determine the level of significance for the differences between the right and left ear exposures.

Of the 20 sets of data of left and right ear exposure results only 1 indicated non excessive noise exposure at the right ear. The left ear was predominantly exposed to cabin noise and during 6 rides exposures remained below the limit for excessive noise exposure of $L_{Aeq}$, h 85dB(A) or its equivalent when the duration of the rides is taken into account. The frequent use of air operated parkbrake systems caused levels of impulsive noise in the cabins up to 101.9 dB(A) Fast or 124.0 dB Peak (Lin) every time they were activated. The left ear noise exposure level results ranged between $L_{Aeq}$ 80.2 dB(A) and $L_{Aeq}$ 88.3 dB(A), with an average noise exposure level of $L_{Aeq}$ 85.07 dB(A).

The right ear exposure consisted mainly of external noises predominantly from the wind rushing past the open drivers side window and, to a lesser extent, sources as exhausts, tyres and other traffic. Right ear noise exposure level results ranged between $L_{Aeq}$ 83.4 dB(A) and $L_{Aeq}$ 89.2 dB(A), with an average noise exposure level of $L_{Aeq}$ 86.78 dB(A).

The maximum difference in which the right ear exposure was higher was 9 dB(A) and the minimum difference was 0.9 dB(A), with an average difference of 1.67 dB(A). The results of the comparison between the left and right ear exposures are represented in figure 1 below.

![Figure 1: Comparison of Left and Right Ear Noise Exposures](image)

- 243 -
Apart from the chance of hearing damage, excessive noise exposure may affect the truck driver's general health, and task performance during driving. Stressors such as excessive noise make demands upon a driver's information processing abilities (Johnson 1980).

CONCLUSION
The results of this study confirm the findings of earlier studies by Nerbonne and Accadri, (1975); Johnson et al (1980); Kai Hong Lam (1980); Hessell et al (1982) and Dufresne et al (1987) who all came to the conclusion, or remarked on the fact, that asymmetric hearing loss was present in the subjects of their studies. Dufresne et al in their study observed a bilateral similarity for most occupations and concluded that seldom does an occupation induce hearing loss confined to, or predominantly in one ear, but truck drivers may well be the exception. The possibility of asymmetric hearing loss occurring in truck drivers is a very real one with ramifications for the wider trucking community.

It can be concluded that despite the fact that the audiometric data could not be used the main aim of this study; to investigate the possibility of asymmetric hearing loss occurring in truck drivers and attributed to external noises such as the wind rushing past the open drivers side window, has been met.

REFERENCES
A RESONANT TECTORIAL MEMBRANE WITH AN ELECTROMECHANICAL OUTER HAIR CELL AS THE MECHANISM FOR COCHLEAR FREQUENCY SELECTIVITY

Anthony W. Gummer, Werner Hemmert, Serena Preyer, Hans-Peter Zenner
Section of Physiological Acoustics and Communication, Dep. Otolaryngology,
University of Tübingen, Silcherstraße 5, Germany.

SUMMARY
Based on vibration measurements of the tectorial membrane (TM) and basilar membrane (BM) and electrophysiological measurements from isolated outer hair cells (OHCs), it is concluded that a resonant TM acts in concert with an electromechanical OHC to reduce the impedance of the cochlear partition and thereby produce cochlear frequency selectivity.

INTRODUCTION
There is overwhelming evidence that OHCs are a crucial component of the cochlear tuning mechanism: electrical stimulation of the efferent fibres of the crossed olivocochlear bundle, which terminate mainly on OHCs, causes reduction of the tuned responses of the afferent fibres (Wiederhold and Kiang, 1970), the inner hair cells (Brown and Nuttall, 1984) and the BM (Dolan and Nuttall, 1994). To date, the most plausible mechanism for cochlear tuning mediated by the OHCs is the electromechanical model. The OHC can respond to controlled changes of its electrical or chemical environment by contraction or elongation of its cell-body (Ashmore, 1987; Brownell et al., 1985; Dallos et al., 1993; Santos-Sacchi, 1989; Zenner et al., 1985). The forces produced by such motility are supposed to act on the cochlear partition to partially cancel its frictional losses or, more concisely expressed, to reduce its mechanical impedance. In the same way that the middle-ear serves to match the low-impedance motion of the air particles to the high-impedance motion of the cochlear fluids, so too is a mechanism required to match the fluid impedance to the still higher impedance of the cochlear partition. Accordingly, the appropriate intracochlear impedance matcher might be found in the electromotility of the OHCs. Whether the electromechanical force is sufficient to significantly influence the motion of the cochlear partition in vivo has not been experimentally determined. Not only must the force be of sufficient magnitude, but it must be applied at the correct moment in the stimulus cycle. For example, if the OHC were to produce a contractile force at the moment when the direction of the BM motion changed from scala vestibuli to scala tympani then the OHC would actively attenuate BM motion - the effective damping of the cochlear partition would be increased, resulting in a loss of tuning. This moment is defined by the phase of the electromechanical force relative to the displacement of the OHC stereocilia, as well as by the phase of the stereocilia displacement relative to the remainder of the cochlear partition - in particular, relative to the BM and to the TM. Neither the relative magnitudes nor the phases are known. Therefore, present experiments in our laboratory are designed to characterize the BM and TM micromechanically and the isolated OHC electrophysiological, with the aim of understanding the interaction of these three cochlear structures.
METHODS

For the electrophysiological experiments, the receptor potential of isolated OHCs was measured in response to direct mechanical stimulation of the hair bundle (Preyer et al., 1994). The electrical input impedance of the cell was evaluated by measuring the change in membrane potential in response to intracellular current injection with the cell clamped to zero current (Preyer et al., 1994). Micromechanical experiments were conducted in the apical turns of the guinea-pig cochlea, where the TM is optically accessible at its upper surface. Vibration measurements in the transverse direction were made with a laser Doppler velocimeter coupled into an epifluorescence microscope (Gummer et al., 1993). Vibration measurements in the radial direction, approximately parallel to the reticular membrane, were made with a self-calibrating photodiode mounted on the microscope parallel to its focal plane (Hemmett et al., 1995). For both types of experiment, the frequency response was determined with a noise stimulus to reduce the recording time. For the purpose of this short communication, the data are summarized in idealized form in Fig. 1.

RESULTS

Irrespective of their cochlear origin, the frequency response of the OHC receptor potential relative to displacement of the hair bundle resembled that of a first-order low-pass filter: i) the amplitude was constant up to a corner frequency, above which it decreased at 6 dB/oct; and ii) the phase was zero at low frequencies, decreased to 45° at the corner frequency and tended to 90° at higher frequencies. The frequency response of the receptor potential was identical to that of the electrical input impedance, implying that the frequency response is defined exclusively by the electrical input impedance of the OHC. The corner frequency decreased monotonically with increasing cell length, which when plotted on a logarithmic frequency axis, amounted to a decrease of 0.58 oct per 10 μm increase of cell length, beginning at 546 Hz for the shortest recorded OHC of length 20 μm (Preyer et al., 1995). Therefore, if the extremes of cell length were derived from OHCs from the extremes of the cochlea, the corner frequencies for the apical-most and basal-most OHCs were about 2.6 oct and 6.2 oct below their respective place frequencies. For a membrane filter with slope of 6 dB/oct, 6 oct amounts to an amplitude attenuation of 36 dB relative to the low-frequency value. Not only is the drive to the electromotor in the OHC wall thereby severely compromised, but such a relatively low corner frequency means that the phase lag, relative to hair-bundle displacement, is 90° for the frequency region relevant for enhanced cochlear tuning by the electromechanical action of the OHCs.

The functional significance of a delayed OHC response can only be understood in conjunction with the micromechanics of the BM and TM. The motion of the BM and TM were similar in the transverse direction, implying that both structures moved in unison for this vibrational degree of freedom. However, the motion of the TM in the radial direction was tuned to a frequency about 0.5 oct below the resonant frequency of the BM, whereas the radial component of the BM motion was just an attenuated version of its transverse component. In other words, the TM possesses a second degree of vibrational freedom, which is not present in the motion of the BM.

DISCUSSION

A complete dynamic description of the interaction between OHC, BM and TM also requires information about the frequency response of the electromechanical - or reverse transduction - pathway. For our purposes there are two phases to be considered: i) the phase of OHC motility relative to the receptor potential and ii) the phase of the OHC force relative to the motility. The
Fig. 1 Comparison of the frequency responses of the basilar membrane (BM), tectorial membrane (TM) and the isolated outer hair cell (OHC). The curves are idealized from data measured in the apical region of the guinea-pig cochlea. The vibration responses are for displacement relative to constant pressure, in the transverse direction for the BM, but in the radial direction for the TM. The receptor potential is relative to stereocilia displacement. A: Amplitudes are relative to their low-frequency values. B: Phase is positive for transverse motion toward scala vestibuli, for radial motion toward the spiral limbus and for hair-bundle motion toward the longest stereocilia. Frequency (oct) is relative to the BM resonant frequency. The phase delay of 90° in the OHC receptor potential and the inertial motion of the TM in the tuned region of the BM response ensures that OHC electromotility reduces the effective resistance of the cochlear partition.

The former phase difference appears to be zero for frequencies up to at least 24 kHz (Daltos, 1994) and, likewise, the latter is zero if the OHC is compliant in the functionally relevant frequency region. Therefore, if compliant, the OHC force must lag stereocilia displacement by 90° at these frequencies. To understand the relevance of this delay, consider what would happen if the TM were elastic in the tuned region of the BM frequency response. For maximal BM displacement toward scala vestibuli, the stereocilia would be deflected away from the spiral limbus, in the depolarizing direction, and the OHC force would be maximal when the BM moved back from scala vestibuli to scala tympani, namely at that instant when the velocity of the BM is maximal. By definition, the viscous force opposing BM motion is maximal at the instant of maximum velocity.
Therefore, the OHC force would be maximal at the same instant and in the same direction as the viscous force. That is, for compliant TM motion the OHC force would attenuate BM motion. Nature appears to have resolved the dilemma by placing a mechanical resonance in the TM, tuned to about 0.5 oct below the BM resonant frequency. In this situation, TM motion at frequencies in the tuned region of the BM response lags TM motion at very low frequencies by about 180°, implying that its motion is mainly inertial at these higher frequencies. Consequently, OHC depolarization is maximal for maximum BM displacement in scala tympani and the OHC force is maximal when the BM moves from scala tympani to scala vestibuli. Therefore, the OHC force is maximal at the same instant as, but in the opposite direction to, the viscous force, and thereby has the appropriate phase to amplify the motion of the cochlear partition.

CONCLUSION

Only when TM motion is inertial can the electromechanical force of the OHC act to improve the sensitivity of the cochlear partition.

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft, SFB 307, Teil project C10.

REFERENCES


ISOMETRIC MAMMALIAN MIDDLE EARS: THE EFFECT OF SIZE ON AUDIOGRAMS

Simo Hemilä, Sirpa Nummela*, Tom Reuter*
Laboratory of Physics, Helsinki University of Technology, Otakaari1, Espoo, Finland
*Department of Zoology, University of Helsinki, Finland

SUMMARY

We first analyze the signal transmission through a generalized isometric mammalian middle ear. When the size of this idealized ear varies, the frequency of lowest threshold \( f_l \) will vary; however, when we use scaled frequencies \( \bar{f}_l \) all predicted audiograms become identical. The high frequency hearing limit \( f_H \) of this idealized mammalian hearing apparatus is proportional to \( 1/\sqrt{m} \) (\( m = \) ossicular mass). This prediction is in approximate agreement with experimental data from land mammals and seals. Next a simple equivalent circuit model of the mammalian middle ear is presented. When three measured quantities of each middle ear are used in scaling, \( f_H \) can be more accurately predicted. In small rodents \( f_H \) seems to be lower than predicted by the model.

INTRODUCTION

Our investigation is based on anatomical middle ear data and behavioural audiograms from 28 species including both small and large mammals. Nummela (1995) has measured the tympanic membrane and oval window areas (\( A_1 \) and \( A_2 \)), the lever arm lengths (\( l_1 \) and \( l_2 \) of malleus and incus, respectively), and the ossicular masses (\( m \)). Behavioural audiograms of the same 28 species have been determined by a number of authors (summarized by Fay, 1988). The middle ears of the land mammals are fairly isometric; in seals \( m \) is about ten times larger than expected on the basis of \( A_1 \). In this work we pursue further the consequences of middle ear isometry.
STRICTLY ISOMETRIC MAMMALIAN MIDDLE EARS

Our comparison of mammalian audiograms is based on five assumptions, all of them being approximatively valid:

1. The shapes of all middle ear structures are isometric and the size of each ear is defined by a characteristic length $\Lambda$.

2. In ears of different size, the densities, the Young’s moduli, and other material parameters are equal.

3. All middle ears are equally good impedance matching devices. Thus, at the frequency of lowest threshold, $f_i$, the power transmission factors are equal.

4. The incident sound wave in the auditory meatus is treated as a plane wave.

5. We assume that, at different frequencies and in different animals, the threshold of hearing corresponds to a definite intensity $I_{\text{min}}$ of the sound wave entering the cochlea.

The mammalian middle ear is a mechanical device which consists of a pressure transformer mechanism, solid elastic components and gas spaces, inertial masses, and resistances. Such a device can be represented as an equivalent circuit. When the corresponding components of ears of different characteristic lengths $\Lambda$ are compared, it appears that the compliances $C$ (both solid elastic structures and gas spaces) are inversely proportional to $\Lambda$, while the inertial mass components $L$ are proportional to $\Lambda^3$ and the dominant resistance, the inner ear resistance $R$, is proportional to $\Lambda^2$.

We compare any impedance $Z$ of an ear of size $\Lambda$ with the corresponding impedance $Z_0$ of a reference ear of the size $\Lambda_0$. Let us denote $w \equiv \Lambda/\Lambda_0$. Moreover, we compare these impedances not at the same frequency but at equal scaled frequencies $f/w$. Then it appears that all mechanical impedances are scaled by the same factor $w^2$:

$$Z_C(f/w) = w^2 Z_{C0}(f), \quad Z_L(f/w) = w^2 Z_{L0}(f), \quad R = w^2 R_0$$

Let us consider a situation where equal-intensity sound waves at frequencies $f$ and $f/w$ enter these middle ears. When the incident pressures are equal, the corresponding input forces are scaled by $w^2$. Therefore, all the corresponding pressures and velocities in these two ears are equal, and the input intensities of the inner ears are equal, too. Taking into account the five assumptions concerning isometric ears, we obtain the final result:

*When compared at scaled frequencies $f/w$, the audiograms of isometric ears are equal.*
Indeed, the absolute thresholds are roughly equal (0 dB SPL), the shapes of all audiograms are fairly similar (excluding the smallest mammals), and with increasing middle ear size the audiograms move to lower frequencies.

As a consequence of this result, the high frequency hearing limit $f_H$ should be proportional to $\Lambda^{-1}$. Because the inertia of the middle ear ossicles is an important limiting factor, we chose $\Lambda = \sqrt[3]{m} / \rho$ where $\rho$ is the bone density. Indeed, in mammalian ears of different sizes, $f_H$ is approximately inversely proportional to $\sqrt[3]{m}$.

A THREE-PARAMETER MODEL OF THE MIDDLE EAR

Fig. 1A shows a simplified mechanical middle ear model and a corresponding simplified equivalent circuit. The components were chosen so that the acoustic input impedance as a function of frequency reasonably well fitted the measured human function. Then scaling rules for these components, based on the measured $A_1$, $l_1$ and $m$ values, were derived (for details, see Hemilä et al., 1995). Thus predictions of $f_H$ values was possible (Fig. 1B). For morphologically extreme middle ears this model improved the predictions based on strict isometry.

![Diagram A](image)

**Fig. 1.** (A) A mechanical middle ear model and the equivalent circuit corresponding to this model. $J$, the inertial moment; $C$, the compliances (elastic components); $R$, the resistance of the cochlea. (B) The behaviourally determined high frequency limit of hearing plotted against the theoretical limit predicted by the three-parameter model.

- , rodents; $\triangle$, cat; $\Delta$, dog; $\square$, chimpanzee; $\blacksquare$, harbour seal; $\square$, man; o, other mammals.
Predictions obtained with this model indicated that only about half of the increase in threshold intensity level at $f_H$ is due to the ossicle inertia.

As seen in Fig. 1B the measured $f_H$ values of the smallest mammals are smaller than predicted by our model. Possibly some limiting factors in cochlear transduction increase the threshold for hearing at frequencies above 100 kHz. In addition, these smallest *microtype* middle ears deviate from the general type in two respects (Fleischer, 1978). Typically a mouse and bat malleus has a rather rigid connection to the tympanic bone. Thus $C_s$ in the model is reduced and the (logarithmic) frequency range of the audiogram becomes narrower, in agreement with measured audiograms. Secondly, the center of mass does not always coincide with the rotational axis. This increases the inertial moment, represented by $m$ in the model, and leads to a decrease of $f_H$, contributing to the deviation seen in Fig. 1B.

REFERENCES


NEW METHOD OF HEARING AIDS FITTING

E. Hojan, R. Drazazga
Institute of Acoustics of the Adam Miczkiewicz University, Poland,
60-769 Poznań, ul. Matejki 48/49.

SUMMARY

New procedure of fitting hearing aids basing on the attribute "loudness", albeit referring to natural sounds with the sound level value varying in time. The results of introductory experiments, for 21 patients with the hearing aids fitting to the new method are described as natural and more pleasant than to the classical one.

INTRODUCTION

The results of tests of hearing aids conducted according to the recommendations of the standards IEC 118 or DIN 45605 differ from those of appraise as of the same devices in clinical tests, [1]. The necessity of applying natural signals while fitting a hearing aid arises from the attempts at simultaneous testing of most psychoacoustic characteristics of hearing which determine the quality of perception, [2], [3]. Research focusing on allowing an optimal acoustic comfort for hearing-impaired people showed, that the most stable attributes appeared to be the "sharpness" and "loudness" of a sound [4]. Based on the attribute "loudness", referenced to natural sounds with amplitude changeable in time, a new method (HGJ method - Hojan, Geers, Jeziorska) is proposed for fitting hearing aids, [5, 6, 7].

HGJ : A NEW METHOD OF FITTING HEARING AIDS

The starting point of HGJ method is Steven's famous argument stating that there is a absolute correlation between the values of the scale and subjective sensations, identical for people of normal and impaired hearing. The HGJ method partly bases on the procedure of Würzburger Hörlfeldskalierung (WHS), [8]. In the HGJ method, the acoustic signal used for the fitting of the hearing aid is sampled with frequency 41.4 kHz and the resulting values of the amplitude are stored by a computer. Each interval of the signal is submitted to the following processes: the calculation of the FFT from 4096 samples reflecting the momentary values of the acoustic pressure, the division of the frequency range within which a FFT analysis was conducted into 0<n<5 user-defined subranges. The computer program TRAX, [5] use the data as a pointer to switch to
calculating successively all included frequency ranges. TRAX'S activity for each listener is the same:
- first it registers in the computer memory his/her reaction time made by moving the mouse according to the impression of loudness of the given signal
- second, it seeks the most-suitable curve for the responses in the physical sound pressure level-subjective loudness domain, i.e...in [dB SPL]-[KU] domain.

All existing curves for the given subject are taken into account during calculating his/her average response function. An individual listening test consists of registering by computer all changes of the subject's response to the sound intensity as he/she is hearing the acoustical signal by headphones. The listener transmits his/her reaction to sound loudness change by moving a mouse pointer along the loudness scale on the computer screen. Then he can asses the loudness in the range from 0 to 53 [KU]-Fig.1

![Fig. 1 The scale of categorical estimations of signal loudness.](image1)

Listener's responses are examined in the whole frequency area of the given signal as well as in its limited ranges. The curves of the loudness changes in sound, registered according to an objective (upper curve) and a subjective method (lower curve) - Fig. 2, were sampled at 98.84 ms intervals and compared.

![Fig. 2 The curves of the changes in signal loudness, registered according to an objective (upper curve) and a subjective method (lower curve).](image2)

The resulting diagram is presented in Fig. 3.
Fig. 3  Categorical estimation of signal loudness as the function of its level of acoustic pressure for the entire frequency range.
Curve 1: 10 healthy persons, 2: single patient E.S., 3: single patient K.N.

Fig. 3. reflect the results of 10 healthy persons-curve 1 for a fragment of the Ravel Piano Concerto. Plotting onto Fig. 3 the results of patient with a hearing impairment (curves 2 and 3) allows the determination of the target value, i.e. that of the signal amplification at the output of the hearing aid, based on the known value of the signal at the input, Fig. 4.

Fig. 4  The level of acoustic pressure at the output of a hearing aid as the function of the level of pressure of the acoustic signal at its input - the Ravel Piano Concerto; for the entire frequency range. Curves 2, 3 - marked as in Fig. 3.

From these curves it is possible to read out the hearing aids parameter: output gain, transmission characteristic, AGCi and AGCo, PC point, compression ratio,

PRELIMINARY RESULTS
The presented procedure was tested through a series of introductory examinations. The examinations covered 20 observers of normal hearing, aged 25 through 40, and 21 observers with the hearing impairment; all observers had been submitted to audiological examinations. The presented natural sounds were a fragment of the 1st
Movement of the Piano Concerto in G major by Ravel, lasting 170 s, Leq = 73 dB, with 28 dB dynamic of local changes and a fragment of the Fugue in B minor by J. S. Bach, lasting 188 s, Leq = 75 dB with dynamic of local changes up to 36 dB. Total dynamics of both signals was the same and equalled 40 dB. After investigating the group of patients whose hearing was impaired, 14 patients had been provided with hearing aid according to the criteria of HGJ method and methods of tonal audiometry. Patients were then examined with questionnaires to decide in which a case the prescribed hearing aid was more acceptable. Treating results of this comparison as entirely preliminary an evaluation of practical application of HGJ method, following general observation have been made:

1. HGJ method requires an exact explanation of calibration procedure of the loudness of the signals with a computer and mouse system
2. It is necessary to make a trial scaling (calibration) of the loudness using a natural sound
3. Patients consider the examination as very natural, pleasant one
4. Acoustic signals perceived with the hearing aid adjusted to the HGJ data are described as natural and pleasant but more silent then the signals perceived from the hearing aid set up according to classical methods
5. Surprising ease in scaling the volume (loudness) of the sounds was encountered with 3 patients who had cochlea implants.

CONCLUSIONS
The presented method of fitting hearing aids allows to determine automatically some parameters of hearing aids in full and some frequency ranges, based on a subjective estimation of loudness and objective measurement of the sound level of the natural sounds. The results of introductory experiments, show that the described HGJ method are described as natural and more pleasant then to the classical one.

REFERENCES
VOLUME FLOWS IN THE INNER EAR WINDOWS

Magne Kringlebotn
The University of Trondheim, Norwegian Institute of Technology,
Division of Physics, N-7034 Trondheim, Norway.

SUMMARY
The equality of volume displacements in the inner ear windows is commonly assumed. In the present work this assumption is experimentally verified. The stapes is given a known displacement. The volume displacement of the round window is determined by measuring the sound pressure set up in a tube cemented to the round window. Inner ears of pigs have been used in the investigation. Supplementary measurements on one human temporal bone have been performed. The equality of the volume flows in the inner ear windows is also supported through an analysis of earlier measurements of the round window displacement for a given sound pressure level at the eardrum. A more comprehensive presentation of the work will appear in The Journal of the Acoustical Society of America.

INTRODUCTION
Are the volume displacements of the oval and round windows equal? The answer has important implications both for the structure of middle ear models (Peake et al., 1992) and for understanding hearing thresholds in patients with disarticulated middle ear ossicles (Shera and Zweig, 1992). The equality of volume displacements in the inner ear windows is also a condition for determining the input to the inner ear by measuring the volume displacement in the round window, for instance when measuring the frequency characteristics of the middle ear (Kringlebotn and Gundersen, 1985).

In conflict with the common assumption of equal volume flows in the inner ear windows, Békésy (1960a) suggested the presence of additional volume flows in the inner ear due to the streaming of fluid in blood vessels and capillaries. Direct measurements of the volume velocities in the oval and round windows have been reviewed by Nedzelntsksy (1974). His conclusion is that the volume flows in the inner ear windows are equal, but only to within an uncertainty perhaps as large as ± 10 dB in magnitude.
METHOD

![Experimental setup for determining the volume flows in the inner ear windows. The numbers are B&K type numbers.](image)

A known movement is mechanically connected to the stapes. The volume displacement of the oval window is the displacement of the footplate times its area. The volume displacement of the round window membrane is calculated from the sound pressure measured at the end of a small glass tube cemented to the round window, taking into account the pressure variations due to standing waves in the tube. Pig ears and one human ear have been used in the investigation. When observed from the brain side, the inner ears may easily be localized and broken loose by means of a screwdriver. The middle ear ossicles chain will then usually disconnect at the incudostapedial joint.

RESULTS

The measured ratios of the volume displacements in the round and oval window, $S_r/S_o$, are shown in Fig. 2 and 3. Average results for 8 pig ears are shown in Fig. 2, and the results for one human ear in Fig. 3.

![Graph showing ratio of volume displacements in the round and oval window of 8 pig ears. Shown are average results and standard errors in mean values.](image)
Fig. 3. Ratio of volume displacements in the round and oval window of a human ear. The two data sets are for different equilibrium positions of the stapes.

DISCUSSION

In average for 8 pig ears and for frequencies up to 2400 Hz the volume flows are found to be equal within ±1 dB, the standard errors in mean values being about 0.4 dB up to 1000 Hz and larger above. The assumption of equal volume displacements in the two inner ear windows is thus confirmed within the measuring accuracy.

Fig. 2 indicate a rapid decrease in the measuring accuracy at higher frequencies. Also, high frequency measurements are supposed to be unnecessary, as the effective fluid path between the two windows decreases with increasing frequency. This is so because the maximum volume displacement of the basilar membrane moves towards the inner ear windows as the frequency increases. The fluid takes the easiest and shortest possible way. With increasing frequency, other possible pathways are thus excluded, or made less accessible by the increase in input impedance of capillaries leading out of the inner ear.

The conclusion that the volume flows are equal in the two inner ear windows may also be supported from earlier measurements on human ears (Kringlebotn and Gundersen, 1985). For a given sound pressure $p_d$ at the eardrum, having area $A_d$ and mean velocity $v_d$, its volume velocity is given by:

$$v_d A_d = \frac{p_d}{Z_d}$$

(9)

where $Z_d$ is the acoustic impedance at the eardrum. Let the velocity at the umbo be $v_u$ and at the stapes $v_s$, and the area of the footplate $A_s$. The volume displacement in the oval window per unit sound pressure at the eardrum then becomes:

$$S_o = \frac{1}{j\omega} \cdot \frac{v_s A_s}{p} = \frac{1}{j\omega} \cdot \frac{1}{Z_d} \cdot \frac{v_u}{v_d} \cdot \frac{v_s}{v_u} \cdot \frac{A_s}{A_d}$$

(10)

where $j = \sqrt{-1}$ and $\omega = 2\pi f$. 

- 259 -
At low frequencies the eardrum impedance is stiffness controlled and given by:

$$Z_d = \frac{\rho_o c_o^2}{\sigma_{oV}} \quad (11)$$

where \( V \) the equivalent volume of the eardrum.

Substituting \( V = 0.7 \text{ cm}^3 \) (Rabinowitz, 1981), \( A_d = 60 \text{ mm}^2 \) (Kringlebotn, 1988), \( v_d/v_u = 1.15 \) (Tonnorf & Khanna, 1972), \( v_u/v_s = 1.3 \) and \( A_s = 3.2 \text{ mm}^2 \) (Békésy, 1960b):

$$S_o' \text{ (dB re } 10^{-5} \text{ mm}^3/\text{Pa}) \approx 24.9 \quad (12)$$

For the 20 best (presumably most normal) human ears out of 68 examined (Kringlebotn and Gundersen, 1985), it was found experimentally that at low frequencies the volume displacement of the round window per unit sound pressure at the eardrum is given by:

$$S_r' \text{ (dB re } 10^{-5} \text{ mm}^3/\text{Pa}) = 21.5 \pm 1.8 \quad (13)$$

Eqs. (12) and (13) gives for the volume flow ratio:

$$S_r'/S_o' \text{ (dB)} = -3.4 \pm 1.8 \quad (14)$$

The close agreement with ratio data in Fig. 3 is accidental, but considering the measuring errors, we here have a mutual support of the two investigations. The earlier supports the conclusion that the volume flows in the two windows are equal, while the present supports the reasonableness of the formerly obtained data.

REFERENCES

Briel & Kjær (1961). Artificial Ear Type 4153. Instructions and Applications, Appendix B.
AN EQUAL-LOUDNESS CONTOUR FOR UNDERWATER ACOUSTIC SIGNAL AND ITS DEPTH DEPENDENCY

Kazuoki Kuramoto, Kensei Oimatsu, Shin‘ya Kuwahara, Shizuma Yamaguchi*
Maritime SAFety Academy, Wakaba-cho, Kure, 737 Japan
* Faculty of Engineering, Yamaguchi University, Tokiwadai, Ube, 755 Japan

SUMMARY

One of the most simple and effective ways for giving the instruction to divers and/or for preventing the diving accidents is to transmit an audio acoustic signal directly in water. Although it is well known that we have an auditory sensation in water, the appropriate expression of loudness in hearing are scarcely studied. In this paper, an equal-loudness contour and a minimum audible field in water are obtained. We discuss the relation of loudness in hearing between in water and in air and investigate the variation of these contours against the water depth.

INTRODUCTION

From the necessity of communication in underwater operation and/or from the insurance of safety in leisure diving, we have proposed the "underwater transmission system" used by an audio frequency signal [1–3]. This is the most simple and effective way because the divers usually have no communication apparatus and the instruction can be given to the large number of people. In such case, it is fundamentally important to investigate an auditory sensation in water, i.e., how the audio signals radiated from the underwater transducer can be heard with diver's ear. The auditory sensation in water has been studied by some investigators until now [4–6]. However, the measurements were mainly about the minimum audible level but scarcely about the loudness level. Therefore, it is necessary to investigate the auditory sensation in water, in particular about the loudness in hearing, for a practical purpose of the construction of the "underwater transmission system".

In the present study, the equal-loudness contour and the minimum audible field in water are obtained by hearing tests for two subjects in the pool or the water tank. Secondly, we discuss the relation of loudness in hearing between in water and in air through two additional experiments. Finally, effects of water depth for these contours are investigated as well.

EXPERIMENTAL PROCEDURE

Equal-loudness level
Measurements for the equal-loudness level were carried out in the pool (25 × 15m). Two subjects, whose auditory sensation in air is normal, dive about 0.3 m under the water with an open-circuit SCUBA equipment and hear the underwater sound radiated from the underwater loudness speaker. The ear canals of the subject's ear were filled with water and the pressure balance was kept with the middle ear. Arrangement for the measurement in the pool is shown in Fig.1.
Experimental procedure in the present work was based upon the experimental guide which has been proposed at ISO TC43/WG1 (7), that is, 1) experiments are made by constant method, 2) time sequence of a stimulus pulse is 1 [s], and 3) measurement frequency must be used in the center of 1/3 octave band between 250 Hz and 4 kHz, and so on. Time pattern of the stimulus sound pair is indicated in Fig. 2. The sound level of the reference signal (1 kHz) was fixed at a constant level (122 and 142 [dB re 1 µPa]).

On the other hand, the sound levels of the test signals were determined as follows. The center level with the same loudness as the reference is briefly selected in advance and eleven numbers of test levels were set up in 2 dB divisions within the limits of 10 dB around the center. The test sound (one of the 11 levels) and the reference sound (122 dB or 142 dB) are presented at random. Each subject decides which is louder, reference or test, and notifies his judgment to the experimenter by a buzzer. The sound pressure level of the test signal was read from the data sheet on the level recorder which records the underwater sound level all the time. Uncertain judgment was never permitted. Then, only one equal-loudness level in water is attained by means of the method of maximum likelihood from the 110 stimulus pulses (11 levels × 10 sessions).

**Minimum audible level**

As the background noise level in the pool is about 87 dB, which is fairly larger than the value expected as a minimum audible level, hearing test is performed by using the water tank in the laboratory (1 × 1 × 2 m; background noise level is about 73 dB). The subject puts one's head in the water and hears the underwater sound. Measurements are made by a limited method. Namely, the sound signal is presented by increasing or decreasing 1 dB step from the initial level which is 20 dB lower or higher than the predicted one. Each test was repeated five times and a mean value was employed as a minimum audible level.

**RESULTS AND DISCUSSION**

Equal-loudness level (122 dB and 142 dB) and minimum audible level obtained in this study are indicated in Fig. 3. The two equal-loudness contours and the minimum audible field have a same tendency to incline upward in the right side, which means that the underwater sound with high frequency is so hard to hear for the diver. This is fairly different in the case of the air, for example, the Robinson-Dadson contour (8) which

![Fig. 3 Experimental results (equal-loudness level and minimum audible level) and modified Robinson-Dadson contour.](image-url)
shows a tendency that the sound with low frequency is so difficult to hear.

**Relationship between in water and in air**

In order to grasp quantitatively the differences of hearing between in water and in air, two additional measurements about the loudness level were done as below. Measurements were performed by means of the method of adjustment in the water tank.

At first, the sound pressure level in air was obtained at several frequencies when the sound in air indicates the same loudness as the sound of 142 dB in water. Figure 4 shows the results of the first measurement. We can find two facts as follows,

(a) The sound pressure level in air $A \ [\text{dB re } 20 \mu \text{Pa}]$, which was the same loudness as the sound of 142 $[\text{dB re } 2 \mu \text{Pa}]$ in water, is roughly expressed against the frequency $f \ [\text{Hz}]$,

$$A = -24.6 \log (f/1000) + 63.0 \quad (1)$$

(b) There is a certain transmission loss caused by the propagation path between the water and the ear. At 1 kHz, for example, the sound of 142 $[\text{dB re } 20 \mu \text{Pa}]$ in water corresponds to the sound of 63 $[\text{dB re } 2 \mu \text{Pa}]$ in air. This fact implies that the transmission loss among the two media is 17 dB since the sound of 142 dB in water is physically equivalent to the sound of 80 dB in air considering that the standard of sound level and the acoustic impedance are different at each medium.

Secondly, the loudness tests similar to the first measurement were done at 1 kHz but against several sound pressure level. Results of the second measurement is shown in Fig. 5. It is found that relation of the sound pressure level between in water $W \ [\text{dB re } 1 \mu \text{Pa}]$ and in air $A \ [\text{dB re } 20 \mu \text{Pa}]$ can be expressed by the straight line inclined 45 degree as,

$$W = A + 79.0 \quad (2)$$

This means that the difference of the sound pressure level in each medium is independent of the pressure level, that is, 10 dB interval between the equal-loudness contours is invariable among in water and in air. We can notice immediately that the minimum audible level of 4 dB in air corresponds to the level of 83 dB in water by an extrapolation of the straight line in Fig. 5 (the position is denoted by cross-shaped broken line). It is notable that the predicted level of 83 dB is close enough to the level of 84 dB obtained in our experiments.

**Modified Robinson-Dadson contour**

On the basis of the equations of (1) and (2)
obtained above, we make an attempt to draw an equal-loudness contour in water from the Robinson-Dodson contour (R-D contour) in air and to compare with our experimental results. As a trial, let us rotate the R-D contour in a counter-clockwise direction corresponds to the slope of 24.6 dB in eq.(1) and let us add the pressure level to the R-D contour corresponds to the constant of 79 dB in eq.(2). Attained results is indicated by solid lines in Fig.3. A good agreement between the experimental values and the modified R-D contour is observed including the minimum audible field. These facts suggest that the results and the methodology in the present work are surely proper.

Depth dependence

In order to examine the effect of the equal-loudness contour and the minimum audible field against the water depth, hearing test was also made at the bottom of the pool with 5 m in depth. It is so-called that the depth of 4.5 m is so sufficient to bring about an impediment in hearing. Experimental procedure is almost the same as the previous tests except that measurements are made by the method of adjustment. The subjects have difficulty in breathing to obtain the sufficient time in measurements.

Figure 6 shows the depth dependencies of the equal-loudness contour and minimum audible

field. A significant difference is not observed between the results in 0.3 m and in 4.5 m.

CONCLUSION

1) Equal-loudness contour and minimum audible field in water were obtained by the tests of hearing for two subjects. As a result, it is found that the underwater sound in the higher frequency region is so hard to hear, which is fairly different in air.

2) Our results of underwater auditory sensation can be expressed by the modified Robinson-Dodson contour which derived through the two additional measurements about the loudness level among in water and in air.

3) From the experiment of the depth dependence, a significant difference was not observed in equal-loudness contour and minimum audible field as far as the water depth is 4.5 m.

The reason why the auditory sensation shows a fairly different aspect between in water and in air is maybe attributable to the physical mechanism of sound propagation process which will be discussed in the next paper.

References

AUDITORY VIRTUAL ENVIRONMENT: SIMPLIFIED TREATMENT OF REFLECTIONS

Hilmar Lehnert, Marcus Richter
Lehrstuhl für allgemeine Elektronik und Akustik, Ruhr-Universität,
IC 1/132, D-44780 Bochum, Germany.

SUMMARY
One of the key methods used in auditory virtual environment is the individual binaural auralization of the direct sound(s) and the early reflections. With respect to the feasibility of this technology it is essential to keep the computational effort of the auralization - namely the number of FIR coefficients of the head related impulse responses (HIR) - as small as possible. Various investigations, employing a single source in an anechoic environment, into this topic have been carried out.

To study the required HTF filter accuracy for reflections, simulations were made, where the direct sound was auralized with full filter length and the binaural treatment of the reflections was systematically simplified. A-B comparisons between the simplified and the original simulations were carried out in listening tests for various kinds of auditory environments. The results indicate that the number of coefficients can be limited to 20 at a sampling frequency of 44.1 kHz, corresponding to an impulse response length of about 0.45 ms.

INTRODUCTION
Auditory Virtual Environment refers to a technology which is capable of shifting a listener auditorily into a different environment which only exists in the form of a computer model (Lehnert & Blauert 1991). Amongst other things, this includes reproducing the spatial properties of the simulated sound field correctly. This is done by means of Binaural Technology, i.e. incoming sound field components are filtered with the Head Transfer Functions (HTF) of the listener and, after proper equalization, reproduced over headphones (see e.g. Blauert 1983, Møller 1992).

The most natural impression can be achieved when the technology provides interactivity and immersion. These requirements form a great technological challenge for the simulation system: The actions of the subjects (e.g. head movements) have to be measured constantly and the simulation needs to be updated in such small intervals, that processing delay and the quantization of movements with respect to time are not perceivable. The corresponding perceptual thresholds are approximately between ten and 100 milliseconds.

In the course of the basic research project SCATIS (Spatially Coordinated Auditory/Tactile Interactive Scenario) which is funded by the European Union inside the program ESPRIT a virtual environment generator for the auditory and tactile modalities is under development (Lehnert 1993). The auditory subsystem is based on the following method (Lehnert & Blauert 1991): Direct sound(s) and early reflections are traced in real-time and auralized using individual auralization units. Late reflections and binaural reverberation are directly generated by parametric algorithms. The auralization units are realized using a network of digital signal processors DSP. Each of the DSPs carries out a task which is shown in fig. 1.

After a simple delay, three separate filters (PRF) are used to perform some prefiltering and two FIR-filters are in charge of the binaural processing, i.e. the filtering with the HTFs. The three pre-filters are used to model source directivity and reflectivities of surfaces. In the example with three pre-filters it is possible to render a second order reflection. This should be sufficient for most cases, since it is probably neither meaningful nor feasible to compute precise reflections of an order higher than two.
Figure 1: Structure of a single auralization unit

The structure shown in figure 1 allows the assignment of computational resources to each block according to its psychoacoustical significance. If, for example, the direct sound of a primary source is to be auralized, no reflection filters are required. Consequently, the blocks PRF-2 and PRF-3 can be skipped, and more power, in terms of FIR coefficients, can be assigned to the binaural filters to obtain precise spatial imaging. For a second order reflection, all pre-filters are involved. Given that the processing power of an auralization unit is constant, this results in less accurate binaural filtering since, more processing power is allocated to the pre-filters. Thus the structure of the auralization unit automatically provides task adequate assignment computation power.

The objective of the present study was to determine the number FIR coefficients required to auralize a reflection. For a direct sound rendered in an anechoic environment a FIR filter length of about 1.6 ms (72 coefficient at 44.1 kHz) is sufficient to represent the impulse response measured at the blocked entrance of the ear canal (Pötsch et al. 1986, Sandvig & Hammershøi 1994). The number of coefficients to auralize reflections was expected to be significantly less.

SHORTENING IMPULSE RESPONSES

Several methods were tried out to shorten the impulse responses including, several window techniques and minimizing the magnitude error of the frequency responses. The final algorithm was a fairly primitive one, which has the advantage that it can be applied easily in real-time for any given length, N, of the impulse response. The starting point of the shortened impulse response is given by the first sample, the magnitude of which exceeds a value computed as the maximum magnitude of the impulse response, multiplied with a threshold factor, k. Starting with this sample, a rectangular window with a length of N samples is applied to the impulse response. Beside the impulse response length N, this threshold factor k is the only parameter of this algorithm. In some pilot studies it was found that the choice of k is not very critical. The main reason for this effect is, that the impulse responses have a relatively sharp onset so that, for a wide range of k, most of the impulse responses start at the same sample.

LISTENING TESTS

The objective of the listening tests was to measure the required impulse response length, N. It was expected that this length would depend both on the type of test signal and on the type of the auditory environment, i.e. size, geometric shape, reverberation times, position of listener and sound source etc. Therefore an attempt was made to identify the setup which is most sensitive to the simplification. To this end three sets of listening tests were carried out. In the first set the volume of the room was varied, in the second the geometric shape, and in the third set the position of source and listener were changed. Each set continued with the configuration that had turned out to be the most critical in the previous set.

The stimuli were generated by a computer model. The reference stimulus was produced by using full-size impulse responses. For the test stimulus, the same configuration was simulated but with simplified binaural impulse responses. It should be emphasized here that in all simulations, the direct sound component was always the same in the reference and the test stimulus, i.e. only the binaural treatment of the reflections was simplified and the direct sound was left untouched.

In all the tests A-B comparisons were carried out using different psychophysical procedures. The different procedures were used to find one which is optimal in terms of the spread of the results and test duration. In all the tests the probability of a subject detecting the simplified version, p_{50%}, was calculated on a confidence level of 95%. The detection threshold was accordingly 50%.
In the first set of tests a lecture room was simulated at three different volumes. The volumes were 150 m$^3$, 1500 m$^3$ and 15000 m$^3$. The corresponding reverberation times were 0.9 s, 1.8 s and 4.0 s at 1 kHz. Figure 3 shows the detection probability as a function of the number of FIR coefficients. In the graphical representation of the setups, the position of the receiver is marked with the letter "R" and the position of the source with an "S". Their orientation is given by the local coordinate cross assigned to them. The X-axis represents the front direction, and the Z-axis the vertical direction.

Figure 3: Detection probability $p_{95\%}$ as a function of the number of FIR coefficients for pink noise (■) and speech signals (♦) at three different room volumes.

It can be seen that the room with the medium volume (1500 m$^3$) shows the highest sensitivity to the simplification of the reflection treatment. In all cases where the detection probability is above the threshold of 50%, the effect is more noticeable for noise rather than for speech signals. This suggests that the main perceptual effect was a change in the timbre of the sound rather than changes in the spatial or temporal structure of the sound field.

In the next set three different geometries were simulated, each at a volume of 1500 m$^3$ with a complexity of between 24 and 250 surfaces. Due to the different ratios of the volume to surface in these spaces, the reverberation times varied between 1.5 s and 1.8 s. This time, only the cases for $N = 10, 15, 20$ were investigated and the detection probability was only in few cases above the threshold. Although the lecture room (HIC) seems to be the most sensitive configuration, there was little dependency on the room geometry in general.

The third set of experiments was carried out with the lecture room at 1500 m$^3$. This time, position of the source and receiver were varied. In setup a) the receiver was placed very close to a wall, in setup b) the receiver was located very close to the sound source and in setup e), the distance between source and receiver was fairly large. The configurations and the corresponding results are shown in Figure 4:

As one would expect, the most sensitive situation is the one where the receiver is close to the surface, since in the reflection has about the same energy as the direct sound in this situation. The detection probability was always higher for the male voice than for the female voice, where it was almost always zero. Since the main difference between the male and the female voice is the lower boundary frequency, this could be an indication that the effect of shortening the impulse responses is mainly audible at low frequencies. The smoothing of the fine structure at the higher frequencies does not seem to cause any audible effects.
CONCLUSION

It is possible to greatly simplify the binaural treatment of room reflections. In all simulated setups no configuration could be found where the detection probability was above the threshold of 50% for $N = 20$. Even for $N = 15$ audible artefacts occurred only in rare cases. The main effect of the simplifications seems to occur at low frequencies.

The algorithm used to simplify the impulse responses was a fairly primitive one, which can easily be implemented in real time. Since the audibility of the artefacts mainly emanates from the low frequency area, the chance of more sophisticated algorithms producing better results for a given $N$ is not very high. The length of the impulse response limits the frequency resolution in principle. For $N = 20$ at a sampling rate of $44.1$ kHz, this resolution is only $2205$ Hz! In practice this leads to filters which are completely flat below $1$ kHz. Therefore the whole low frequency area is represented by only one value. Apparently, the limiting factor here is simply the filter length and not the way the filter is built.

Admitted, only a very limited number of configurations have been treated in this investigation. Although an attempt was made to vary the most important factors in room acoustics in a reasonable manner, a really systematic approach is definitely not possible. Therefore it is difficult to make a generalization about the results. For example, in the case of a listener located between two closely spaced parallel rigid walls, the simplified model may not work correctly. However, it seems reasonable to assume that 20 coefficients are sufficient in nearly all relatively natural cases.

REFERENCES

MOST COMFORTABLE LISTENING LEVELS FOR LISTENING THROUGH EARPHONES AND IN A SOUND FIELD

Tomasz Letowski, Diana W. Emanuel*
U.S. Army research Laboratory, Aberdeen Proving Ground, MD 21005
*Towson State University, Towson, MA 21003.

SUMMARY

Most comfortable listening (MCL) levels for connected speech were measured for: (1) monaural earphone listening, (2) binaural earphone listening, (3) monaural sound field listening, and (4) binaural sound field listening. Two groups of normal-hearing listeners, age 18-26 (Group 1) and 58-75 (Group 2), participated. Three types of earphones were used: Telephonics TDH-49 supra-aural earphones (both groups), Sennheiser HD-250 circumaural earphones (Group 1 only), and Etymotic ER-3A insert earphones (Group 1 only). In all listening conditions and for both groups of subjects, the binaural MCL levels were lower than monaural levels by about 4 to 6 dB. Soundfield levels were lower than the TDH-39 earphone levels by approximately 3 to 6 dB. The MCL levels for older adults were similar to those for younger adults for the same speech material. They went up, however, about 4 to 5 dB when the speech material was replaced with a TV movie. Results of this study may be useful in selecting and predicting MCL levels and in comparing MCL data obtained in various listening conditions.

INTRODUCTION

Most comfortable listening (MCL) level is frequently used as the level of sound presentation in audiology, speech science, psychoacoustics, and sound quality studies. There is, however, a lack of consensus on how the MCL level depends on the listening conditions, instruction, age of the listener, or test procedure. For example, the MCL levels for monaural earphone listening to speech signals have been reported as low as 48 dB SPL (Ventry, Woods, Rubin, and Hill, 1971) and as high as 85 dB SPL (Richards, 1982). It is generally accepted that MCL is a range of levels rather than a single level, but this range must be relatively narrow to have a practical value.

The purpose of this study was: (1) to measure and compare MCL levels for connected speech presented in a sound field and with a variety of earphones and (2) to determine whether MCL level changes with age. The direct impetus for this study was the authors' attempts to use the MCL level for speech in noise as a measure of tolerance to noise and a criterion for attenuation of speech by hearing protectors.

METHOD

Two groups of subjects, age 18-26 (Group 1) and 58-75 (Group 2) participated in the study. All listeners had hearing threshold levels 20 dB HL or better at audiometric frequencies from 250 Hz to 4 kHz. The difference between hearing thresholds in right and left ears never exceeded 10 dB.
at a given frequency. All subjects in both groups were female. The subjects in Group 1 were students of The Pennsylvania State University while the subjects in Group 2 were volunteers selected from over 200 respondents to a newspaper advertisement.

All tests were conducted in a listening booth (Suttle B1) having ambient noise level suitable for ears open testing (ANSI S3.1-1991). The subject was seated in the center of the booth facing a loudspeaker at a 1.5 m distance (sound field listening) or wearing earphones (earphone listening). In the sound field, MCL levels were measured for binaural (both groups) and monaural (Group 1 only) conditions. The subject was asked not to move during testing and to hold her head in one position. Three sets of earphones were used in earphone listening tests: Telephonics TDH-49 supra-aural earphones, Sennheiser HD-250 circumaural earphones, and Etymotic ER-3A insert earphones. Subjects in Group 2 used only the TDH-49 earphones. In the monaural condition the preferred ear was used and the other ear was occluded with an EAR \textsuperscript{TM} plug and a Bilsom earmuff (sound field listening) or covered with an inactive earphone (earphone listening). During the tests with ER-3A earphones, the foam tips were inserted 12 to 15 mm from the entrance to the ear canal.

The speech signal was a commercial recording (Auditec, St. Louis) of a story read by a male talker. The talker had a general American accent and read the text with flat inflection. The signal was played from a CD player through a Beltone 2000 audiometer and delivered to either the loudspeaker or the earphones. In addition, the subjects in Group 2 listened to the soundtrack of the movie "The Princess Bride" and watched the movie on the TV screen. The movie was played from a VCR and the TV was located in front of the subject and below the loudspeaker level. The soundtrack was delivered through the loudspeaker or the earphones, depending on the listening condition.

The MCL levels were obtained using a 2dB-up/10dB-down ascending variant of the method of limits. Four seconds of speech were presented at each level. The subject's task was to indicate whether the speech signal was too soft or too loud to be comfortable. The optimum listening level was obtained three times and the median value was recorded as the MCL value.

The Beltone 2000 audiometer and all transducers were calibrated prior to data collection (ANSI S3.6-1989). The calibration signals were an octave noise band centered at 1000 Hz and the speech signal used in this study. The calibration accuracy was verified at the end of the study and the calibration error did not exceed 1 dB for any set of transducers. The TDH-49 and HD-250 earphones were calibrated using an IEC 318 coupler with a Type 1 Adaptor (HD-250). The ER-3A earphones were calibrated using an ANSI HA-2 coupler. The ER-3A foam plugs were 12 mm long.

The data collection was part of a larger study on speech perception in noise. Each subject participated in one listening session. The duration of the session was approximately 1.5 hours but the MCL testing contributed to less than one-third of the session. The order of listening conditions was semi-counterbalanced across subjects.

RESULTS AND DISCUSSION

Means and standard deviations for MCL levels obtained for young (Group 1) and older adults (Group 2) in various listening conditions are summarized in Tables 1 and 2, respectively. The mean monaural and binaural MCL levels for young adults and the sound-field listening situation
were found to be about 64 and 60 dB SPL, respectively. These values agree well with the level of conversational speech, which is typically around 65 dB SPL. They also agree with binaural MCL levels reported in our previous studies (Letowski, Magistro, and Ritter, 1994; Letowski, Burstein, Clark, Romanowski, and Sevec, 1955).

Table 1. Means (M) and standard deviations (SD) for MCL levels reported by young adults (Group 1).

<table>
<thead>
<tr>
<th></th>
<th>Monaural Listening</th>
<th>Binaural Listening</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (dB SPL)</td>
<td>SD (dB)</td>
</tr>
<tr>
<td>Sound Field</td>
<td>64.4</td>
<td>6.1</td>
</tr>
<tr>
<td>TDH-49</td>
<td>70.3</td>
<td>5.8</td>
</tr>
<tr>
<td>HD-250</td>
<td>65.1</td>
<td>9.4</td>
</tr>
<tr>
<td>ER-3A</td>
<td>61.0</td>
<td>7.6</td>
</tr>
</tbody>
</table>

Table 2. Means (M) and standard deviations (SD) for MCL levels reported by older adults (Group 2).

<table>
<thead>
<tr>
<th></th>
<th>Monaural Listening</th>
<th>Binaural Listening</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (dB SPL)</td>
<td>SD (dB)</td>
</tr>
<tr>
<td>Speech</td>
<td>Sound Field</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>TDH-39</td>
<td>66.9</td>
</tr>
<tr>
<td>TV Movie</td>
<td>Sound Field</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>TDH-39</td>
<td>71.9</td>
</tr>
</tbody>
</table>

The differences between monaural and binaural MCL levels measured for young adults were very similar for all listening situations (loudspeaker, earphones) and stayed within 4 to 6 dB. The differences between the soundfield and earphone MCL levels for the same listening condition (monaural, binaural) were in the 3 to 6 dB range (see Table 1). Please note, however, that these differences are only true for the calibration procedures described in the previous section.

Comparison of the MCL levels for subjects in Group 1 and Group 2 for the same speech signal is shown in Figure 1. This comparison reveals that MCL levels for older adults are somewhat lower than for young adults. This difference is not, however, statistically significant (F=1.08, df=1.18; p>0.3). Nevertheless, it indicates that the MCL levels for normal-hearing older adults are not higher than for young listeners. The need for higher speech levels for older people seems only to result from acquired hearing loss and not from increased age.

The story used as the speech signal in this study was related to the metric system and was read without inflection by the talker. Such text was probably not very exciting for the older adults. As a result, they might have done less than they could to follow the text and were concerned with its loudness rather than its intelligibility. During the second task, where the subjects were asked to set their MCL levels for the soundtrack of an interesting movie, their MCL levels went up by about 3 to 5 dB. The greatest increase was observed for monaural earphone listening whereas the smallest was found for binaural soundfield listening. The average size of this increase corresponds well to the difference of 4 to 5 dB between the MCL level for loudness and MCL level for the intelligibility reported by Hochberg (1975). Please note, however, that the increase in the MCL level for a TV program could be also caused by different mode of presentation (audio-video).
CONCLUSIONS

The results reported in this study demonstrate that age alone does not affect MCL level for connected speech. This level, however, can be easily affected by the listener's interest in the content of speech. The difference between monaural and binaural MCL levels for speech seems to be about 4 to 6 dB and independent of the listening situation.

ACKNOWLEDGMENTS

This study was supported by a grant from the AARP Andrus Foundation to the first author.

REFERENCES


AUTOMATIC AND AURAL-PERCEPTUAL SPEAKER VERIFICATION IN THE PRESENCE OF MASKING NOISE

Wojciech Majewski
Institute of Telecommunication and Acoustics, Technical University of Wroclaw, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland.

SUMMARY

The method and the results of automatic and aural-perceptual speaker verification in the presence of masking noise are presented and discussed. For both methods similar values of verification errors were obtained. Diminishing of S/N ratio had generally a small influence on speaker verification errors. A proper selection of signal parametrization and decision strategy in automatic method and a proper selection and training of a listener in aural-perceptual method may result in zero errors. Thus, it is difficult to state which one of the two presented methods is absolutely superior.

INTRODUCTION

Speaker recognition encompasses two basic procedures, i.e. identification and verification. Speaker identification is defined as the process of identifying an unknown speaker from samples of his or her voice. Similarly, speaker verification refers to a decision-making-process leading to an acceptance or rejection of claimed identity of a speaker.

Speaker recognition may be carried out by means of subjective or objective methods. In subjective methods decisions relating to speaker recognition are performed by a human being, in objective methods such decisions are made by a computer. Subjective methods may be either visual or perceptual. Visual methods are based on visual comparisons of speech spectrograms, whereas the perceptual ones are based on aural comparisons of speech samples with the utilization of long- or short-term memory. Nowadays, the visual method is loosing its significance, while aural-perceptual methods constitute still actual subject of investigations [1], similarly like objective, automatic methods [2], which are very close to practical applications.

There is a number of publications [3] related to the comparison of effectiveness of perceptual and automatic methods and in this thrust the present paper is placed. Specifically, the results of aural-perceptual and automatic speaker verification obtained under similar conditions are presented and compared.

EXPERIMENTAL MATERIAL.

As a phonetic material a Polish word "anonimy" (Eng. anonyms) was utilized. It is a key word of confirmed usefulness for the purpose of automatic speaker recognition [4], consisting of voiced sounds considered as good carriers of the individual voice features.
In the experiments the voices of 10 male speakers exhibiting no speaking defects were used. Each speaker produced the key word 20 times under average acoustic conditions (computer laboratory, S/N=42 dB) and five times for each of three S/N ratios (18, 9 and 3 dB), obtained by masking the utterances with the aid of white noise radiated by a loudspeaker placed one meter from the microphone under 45° angle to the axis speaker-microphone.

The utterances produced during a single session were converted to digital form, registered in computer memory and utilized next in the experiments of automatic and aural-perceptual speaker verification.

AUTOMATIC SPEAKER VERIFICATION (ASV)

Utilizing an universal procedure of automatic speaker recognition in open sets, that has been worked out in our Institute [5,6], a series of experiments has been carried out [7] in order to find out the influence of external masking noise on speaker verification errors. To create patterns of key word utterances three following parameter sets were used: speaking fundamental frequency contours (F₀) approximated by Gramm polynomials of five coefficients (parameter space dimension P=5), distributions of time intervals between zero-crossings (ZC) into eight time channels (P=8) and mean spectrum (MS) in one-third octave bands of central frequencies ranging from 160 to 3150 Hz (P=14).

Depending upon the adopted strategy of operation of ASV system, especially depending upon the selection of learning sequence and decision thresholds, different results of speaker verification were obtained. To obtain satisfactory results it was necessary to include in the learning sequence all potentially possible conditions of speech transmission. False rejection error α and false acceptance error β obtained under such conditions for assuring α=β optimal decision thresholds established for no noise condition (S/N=42 dB) and established individually for each value of S/N are presented in Table 1.

Table 1. α and β errors (in %) for optimal decision thresholds established for S/N=42 dB (a) and for thresholds established individually for given S/N ratio (b).

<table>
<thead>
<tr>
<th>S/N (dB)</th>
<th>F₀</th>
<th>ZC</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>α</td>
<td>β</td>
<td>α = β</td>
</tr>
<tr>
<td>42</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
</tr>
<tr>
<td>18</td>
<td>0.0</td>
<td>9.5</td>
<td>6.0</td>
</tr>
<tr>
<td>9</td>
<td>6.0</td>
<td>8.0</td>
<td>6.5</td>
</tr>
<tr>
<td>3</td>
<td>22.0</td>
<td>8.5</td>
<td>13.0</td>
</tr>
</tbody>
</table>

It may be seen from the data presented in Table 1, that including in the learning sequence all conditions of speech transmission, it is possible to obtain speaker verification scores practically independent on external disturbances, if decision thresholds may be adjusted to particular conditions of transmission (columns "b"). If such possibility does not exist (columns "a"), the
verification errors become generally larger for smaller values of S/N ratios. The parameter ZC was in such a case the most sensitive to the change of S/N ratio, what is quite obvious considering the definition of this parameter. It is worth to note that for parameter MS, regardless of the manner of establishing the decision thresholds, verification errors equal zero were obtained for all S/N ratios, what only partially may be explained by the largest dimension of this parameter (P=14 in comparison to P=5 and P=8 for the other two parameters).

AURAL-PERCEPTUAL SPEAKER VERIFICATION

Experiments on perceptual speaker verification [8] were performed by a group of five listeners of normal hearing, who were comparing subsequent pairs of key word utterances. The utterances were reproduced from the computer and D/A converter by means of headphones. The task of the listeners was to make a decision, whether a given pair of stimuli was produced by one speaker or by two different speakers.

The listening session were carried out under average acoustic conditions (computer laboratory) and the listeners' answers were registered by the computer, which classified them properly. Each of the subject took part in eight listening sessions of two hours duration each. After a short, few minutes training each subject made 1275 judgements during each session. Every session concerned the comparisons of two sets of stimuli of definite value of S/N ratio. α and β errors made by particular listeners in subsequent listening sessions are presented in Table 2.

Table 2. α and β errors (in %) made by particular listeners.

<table>
<thead>
<tr>
<th>S/N (dB)</th>
<th>Listener</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Set</td>
<td>TM</td>
<td>GJ</td>
<td>MK</td>
<td>JS</td>
<td>JZ</td>
<td>Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 42</td>
<td>0.7 16.6</td>
<td>17.3 2.5</td>
<td>8.0 8.8</td>
<td>1.3 12.5</td>
<td>3.3 2.0</td>
<td>6.1 8.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 18</td>
<td>22.0 1.7</td>
<td>40.0 3.3</td>
<td>17.3 6.3</td>
<td>15.3 1.8</td>
<td>14.7 1.3</td>
<td>21.9 2.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 9</td>
<td>29.3 1.8</td>
<td>50.0 9.2</td>
<td>24.7 5.3</td>
<td>15.3 5.0</td>
<td>18.0 2.4</td>
<td>24.7 4.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 3</td>
<td>14.0 1.1</td>
<td>42.7 6.0</td>
<td>15.3 4.5</td>
<td>5.3 4.5</td>
<td>16.0 2.5</td>
<td>18.7 3.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 18</td>
<td>4.0 0.0</td>
<td>7.1 5.2</td>
<td>1.3 0.9</td>
<td>0.7 0.7</td>
<td>2.7 1.7</td>
<td>3.2 1.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 9</td>
<td>5.3 0.0</td>
<td>16.0 1.5</td>
<td>7.3 0.1</td>
<td>1.3 0.8</td>
<td>5.3 1.2</td>
<td>7.0 0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 3</td>
<td>0.0 0.0</td>
<td>19.3 2.3</td>
<td>12.0 1.0</td>
<td>1.3 0.9</td>
<td>4.0 1.9</td>
<td>7.3 1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 42</td>
<td>0.0 0.0</td>
<td>19.3 0.8</td>
<td>7.3 0.2</td>
<td>2.7 0.2</td>
<td>2.7 1.8</td>
<td>6.4 0.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results presented in Table 2 indicate large differences in the number of errors made by particular listeners, what causes that the mean values of errors averaged over the listeners (presented in the last two columns) are of tentative nature. These large differences in the listeners' judgements reflect different abilities of the listeners in discriminating speakers' voices and their different experience in speaker verification tasks (listeners JS and JZ had such experience and they knew most of the speakers, so they could additionally utilize a long-term memory in their
judgements, listeners TM and GJ were the graduate students in acoustics at our Institute and the
listener MK was a naive subject with no experience in such experiments).

The influence of listener's training on speaker verification scores may be also observed. The
experienced listeners JS and JZ obtained similar results and the differences in their judgements for
the first and the last session, that were run under the same conditions of transmission, are rather
small, while the remaining listeners were improving their results in the course of the experiments.
A very good ability to speaker recognition was indicated by the listener TM, who during last two
sessions, i.e. after intensive practice, did not make any mistake.

Finally, it is worth to note, that false rejection errors (x) were much larger, when the utterances
under comparison were recorded for different S/N ratios. Thus, the external noise of different
levels makes the task of positive verification more difficult for the subjects.

CONCLUSIONS

The comparison of the results obtained by means of automatic and aural-perceptual method is
difficult, since the procedures utilized in both cases were somewhat different and it was not
possible to perform both experiments under exactly the same conditions. Generally, however, it is
possible to state, that the values of errors in both methods are similar, that diminishing the S/N
ratio had relatively small influence on the speaker verification results and that the proper selection
of signal parametrisation and decision strategy in automatic method and the proper selection and
training of a listener in perceptual method may result in zero errors in both methods under
comparison.

On the basis of the performed experiments it is difficult to state which one of the two presented
method is absolutely superior. Thus, a conclusion presented in one of the previous studies [3], that
perceptual method provides better result than the automatic one, has not been confirmed.

REFERENCES

1. J. Kreman, G. Papcun, Comparing discrimination and recognition of unfamiliar voices, Speech
2. S. Furu, Speaker-dependent-feature extraction, recognition and processing techniques, Speech
Communication, 10, 5-6 1991, 505-520.
3. A. Federicio et al, Comparison between automatic methods and human listeners in speaker
vol 2, Belgrade 1989, 427-430.
5. C. Basztura, W. Majewski, J. Jurkiewicz, Automatic voice recognition in open sets, Archives of
Acoustics, 13, 3-4 1988, 205-218.
6. C. Basztura, Experiments of automatic speaker recognition in open sets, Speech
7. C. Basztura, W. Majewski, W. Barycki, Influence of external masking noise on the effectiveness
of automatic speaker verification, Archives of Acoustics (accepted for publication).
8. G. Jasinski, Subjective verification of the results of automatic voice recognition, Master's
Thesis, Institute of Telecommunication and Acoustics, Technical University of Wroclaw,
THE MIDDLE EARS OF BIRDS AND MAMMALS: A COMPARISON OF STRUCTURE AND FUNCTION

Sirpa Nummela, Simo Hemilä*, Tom Reuter
Department of Zoology, University of Helsinki, P.O. Box 17, Helsinki, Finland
*Lab. of Physics, Helsinki University of Technology Otakaari 1, FIN-02150 Espoo

SUMMARY

We are interested in the general question how animal size limits the size and information processing capacity of sense organs. We measured the following structural parameters: skull mass, ossicular mass, areas of the tympanic membrane (TM) and the oval window (OW). In both birds and mammals we found that the size of the middle ear is negatively allometric to animal size (skull mass). When a bird and a mammal have equal TM areas, the masses of their columellae (birds) and stapes (mammals) are equal, indicating that these homologous ossicles - despite their different shapes - are similarly optimized in relation to TM induced stresses. The TM/OW area ratios are also similar in birds and mammals. The two additional mammalian ossicles, the malleus and the incus, produce a hinge-system with lever arms further enhancing the pressure at the level of the OW. These ossicles clearly increase the mass inertia of the ossicular chain, and the role of this hinge must be critically pondered.

INTRODUCTION

The middle ears of land-living vertebrates (tetrapods) act as an impedance matching device between the air and the inner ear fluid. The pressure enhancement is mainly created by a considerable TM/OW area ratio.

Birds have one vibrating ossicle, the columella, between these two membranes, while mammals have three ossicles, the malleus, the incus and the stapes. The malleus-incus hinge moves the stapes which acts as a piston against the oval window. As a result of a difference in lever arm length in the malleus-incus hinge, an additional pressure enhancement is obtained. Similar lever arm effects are in fact possible to obtain in the bird middle ear, too (Gaudin, 1968; Saunders and Henry, 1989).

It is thus obvious that in the middle ears of birds and mammals somewhat different ossicular arrangements meet similar functional needs. In this short review we make a quantitative comparison of these functionally corresponding structures.

A previous study has already revealed some constant size relations for the mammalian hearing apparatus (Nummela, 1995): the combined mass of the ossicles is negatively allometric to the skull mass (seals excluded), and the relation between the mass of the stapes and the combined mass of malleus and incus is strictly isometric, as is the relation between a two-dimensional measure of the ossicles (mass$^{2/3}$) and the TM area.
Further the high-frequency hearing limit of mammals is inversely proportional to the cubic root of the ossicular mass (Hemila et al., 1995), suggesting that the inertia of the ossicles is an important factor limiting high-frequency hearing.

In this study we use the following symbols: \( M = \) malleus mass, \( I = \) incus mass, \( S = \) stapes mass, \( \text{Col} = \) columella mass. \( A_1 = \) TM area, \( A_2 = \) OW area.

MATERIAL AND METHODS

Skulls and middle ear ossicles from 63 mammalian and 17 avian (bird) species were obtained from various Zoological museums. Masses were determined using appropriate balances. The area of the pars tensa region of the tympanic membrane was obtained from the bony ridge surrounding it (sulcus tympanicus), and the area of the oval window was acquired by measuring the stapes footplate area. The mammalian data have been presented in a previous paper (Nummela, 1995).

RESULTS

For describing the size of the middle ear as a function of the animal size, we plotted the ossicular mass against skull mass. For mammals we used the sum of the masses of malleus and incus, \( M+I \). The mass of stapes is less than 10% of the \( M+I \) mass, and further we know that the mass of stapes is strictly isometric with that of malleus-incus. By excluding stapes we could include a few species for which the stapes was missing in our material.

The allometric equations when plotting ossicular mass (mg) against skull mass (g) were for birds and mammals respectively:

- **Birds:** \[ y = 0.143 \times 0.673^{x} \ (r=0.95) \]
- **Mammals:** \[ y = 1.373 \times 0.513^{x} \ (r=0.95) \]

The true seals (Phocidae) deviate from all the others by having ten times heavier ossicles than other mammals with similar skull mass (Fig. 1).

Isometric objects have the same shape although their size may vary. For isometric bones having the same density, a given cross-sectional area is related to the mass of the bone as mass \( ^{2/3} \). Such cross-sectional areas describe (at least theoretically) the relative compressive strengths of the ossicles.

Next we plotted some generalized two-dimensional (cross-sectional) values of the ossicles against the corresponding TM areas. In all three cases, \( (M+I)^{2/3}/A_1, S^{2/3}/A_1 \) and \( \text{Col}^{2/3}/A_1 \), the relationship was strictly isometric (for the two latter plots, see Fig. 2). The regression analyses give the equation \( y = 0.016 \times 0.990^{x} \ (r=0.94) \) for the avian columella, and \( y = 0.02 \times 0.987^{x} \ (r=0.96) \) for the mammalian stapes (in relation to TM area).

These two groups of data points do not differ, i.e., they could be samples from the same population of normally distributed data points. Again, the true seals were excluded from the mammalian regression analysis.
Fig. 1. Mass of columella (birds) and combined mass of malleus and incus (M+I, mammals), plotted against skull mass. Symbols and abbreviations in Figs. 1-3: ●, birds; ○, mammals (seals excluded); ■, true seals. Species: B, Bat; C, Cat; Cc, Chinchilla; Cn, Canary; E, Elephant; G, Gerbil; M, Man; Md, Mallard duck; S, Shrew; St, Starling; Sw, Swan.

Fig. 2. Two-dimensional value of columella (birds) and stapes (mammals) plotted against the TM area. Regression lines for birds (Bi) and mammals (Ma) are shown. For symbols and abbreviations, see Fig. 1.

Fig. 3. Oval window area plotted against tympanic membrane area for birds and mammals. For symbols and abbreviations, see Figs. 1 and 2.
Finally we plotted OW area against TM area (Fig. 3). The regression analysis for birds gives the equation \( y = 0.068 x^{0.897} \) \((r=0.96)\), and for mammals \( y = 0.032 x^{1.047} \) \((r=0.97)\). Again, the coefficients of these two groups do not deviate significantly from each other.

DISCUSSION

In both birds and mammals the middle ear size is negatively allometric to animal size. This reflects a general principle for sense organs in the animal kingdom: the need of information about the outside world is not related to animal size in the same direct way as the need of food and oxygen is related to body size.

It is also important to note that the information receiving and processing capacity of a sense organ is rather a function of its absolute size, not of its relative size. Thus it is worthwhile for some small animals to invest altogether 5-10 % of their body mass in sense organs. Large animals can often allow themselves very efficient sense organs which still occupy a much smaller fraction of their mass and body surface.

A large TM area always implies an interesting compromise: it allows efficient collecting of acoustic energy and a high signal/noise ratio at the level of the inner ear. It is indeed very efficient in low-frequency hearing. However, as Fig. 2 reveals, it also results in massive middle ear ossicles and impaired high-frequency hearing.

The similar area ratios in Fig. 3 offer no surprise; they indicate that mammalian and avian middle ears perform a very similar impedance matching. Fig. 2, however, presents an entirely novel and surprising observation for which we have no functional interpretation. Although an avian columella and a mammalian stapes have rather different shapes, these ossicles have the same mass when they are driven by equally wide tympanic membranes and coupled to equally large oval windows.

REFERENCES


PSYCHOACOUSTIC EVALUATION OF SPEECH ENHANCEMENT USING THE PARTRAN CONCEPT

P. Rubak, K. Hermansen, F. K. Fink, U. Hartmann
Aalborg University, Dept. of Communication Technology, DSP-group.
Fredrik Bajers Vej 7, DK 9220 Aalborg Ø., Denmark

PARTRAN CONCEPT

The PARTRAN concept is aimed at flexible speech signal manipulation without introducing unwanted signal artefacts. See figure 1 for a symbolic diagram of the principle, and ref. [1] for a more detailed description. The signal manipulation is based on a separation of the speech signal into a model describing quasistationary parts, and a transient part. The quasistationary part is related to the vocal tract and the transient part is primarily related to excitation signal (pitch signal) and the fast transient sounds as the stop consonants.

This separation makes it possible to manipulate the quasistationary and transient parts individually, which provide us with a high degree of flexibility in the signal-processing.

The quasistationary part is described by an 12. order LPC model. The 12 coefficients are calculated every 5 msec, based on a 30 msec frame-length. The LPC model also determine an inverse filter (whitening filter). The speech signal is inverse filtered, and the output is the residual signal (represents the speech information not included in the vocal tract model). Primarily the residual signal relates to the pitch signal and the stop consonants.

Transformation possibilities

After separation of the speech signal, manipulation can take place either in the vocal tract model or in the residual signal or in both. Manipulation in the vocal tract model uses a pseudo decomposition principle. We separate the LPC-spectrum into a number of 2. order sections (each characterised by estimated poles from the spectrum and padded with zeros at z=-1 and z=1). The 2. order sections are described by three parameters: the resonance frequency, the bandwidth (or Q-factor) and the power. These 2. order sections forms the basis for the flexible manipulation of the vocal tract spectrum. Manipulation takes place by changing these key parameters for each section, according to the purpose of the speech processing.

Resynthesis

The last step in the signal processing is to filter the residual signal through a resynthesis filter, which is based on the transformed 2. order sections. The best results are obtained by using a FIR filter (change of filtercoefficients in IIR filters can produce spurious sounds).
Software implementation
The PARTRAN-concept has up till now been implemented in the program language C and run on a Personal Computer. For the time being we have a first version of a real-time implementation with the ADSP 2101 signal processor (16 bit fixed point) as the hardware platform.

Figur 1: PARTRAN concept for speech transformation.

SPEECH ENHANCEMENT

There are several applications of the PARTRAN concept. In this paper we will consider speech enhancement through spectral sharpening. The PARTRAN concept provide an efficient formant based processing scheme to improve the signal-to-noise ratio for the formants. The level differences between formant peaks and the "noise controlled" valleys can be enhanced in a flexible way. The basic idea is to transform the Q-factors obtained from the LPC analysis. For speech signals, contaminated with background noise, an increase of the Q-factors will result in a better signal-to-noise ratio.

Selection of Q-factors
The selection of Q-factors represent a compromise between improvement in the signal-to-noise ratio and unwanted distortion of the temporal envelope. Onset disparities are important for speech perception. According to Summerfield [3] a broadening of the formants (synthesized CVC was presented in quiet) had deleterious effects on identification accuracy. These results were obtained for both normal and impaired listeners. It could be interesting to look at the Q-factors in the auditory system, we know that this system work well, and represent a good compromise between frequency- and time resolution. According to Evans [2] the effective bandwidth of the auditory filter, is about 1/6 octave (filter Q-factor about 10, from 1-10 kHz), which is substantially narrower than the critical bandwidth. It seems reasonable, as a starting point, to use about the same Q-factors as occurring in normal speech. But higher Q-factors will result in a better signal-to-noise ratio.

Comparison with other methods
In the literature several methods to improve the frequency resolution and the signal-to-noise ratio are described. The method described in [4] is based on a convolution, in the frequency domain, with a "Mexican-Hat function" (difference of Gaussians). In this way small peaks in the spectral envelope pattern can be amplified substantially.
To remove irrelevant spectral details, the excitation pattern corresponding to the short-term magnitude spectrum is calculated before the "peak-amplifying convolution" is performed. Small but consistent improvement in speech intelligibility for hearing impaired persons is reported. The basic idea behind this processing method is fundamentally the same as lateral suppression used in the auditory processor. The theoretical background for the PARTRAN method is not far from this "convolution" method.

**SUBJECTIVE ASSESSMENT OF SPEECH QUALITY**

The purpose of the subjective listening test was an assessment of the possible degradation of the speech quality caused by the signal processing in the PARTRAN system. A basic requirement is that a 1:1 transformation must result in a good speech quality. To verify this we have conducted a listening experiment, according to CCIR Rec. 562-2.

**Rating scale & Speech material**
We selected the five-grade impairment scale:
1=Very annoying, 2=Annoying, 3=Slightly annoying, 4=Perceptible (but not annoying) and 5 imperceptible.

The original speech signal (unprocessed) was used as the reference signal. The test signals was the speech signals after the processing in the PARTRAN-system. Four different speakers was used, two male and two female. Two different sentences was selected, resulting in 8 different testsignals. The duration of each sentence was about 5 sec.

**Psychoacoustic paradigm**
For tests using the five-grade impairment scale an ABAB paradigm is recommended. Each program sequence is presented in the following order: 1. original sequence, 2. same sequence (impaired), 3. original sequence (repeated) and 4. same sequence (repeated).

**Listening panel & Sound reproduction**
We used 12 listeners, age between 20 and 50 years, some of the listeners had a mild high frequency hearing loss. (The applications of the PARTRAN system are intended to include both normal hearing and hearing impaired persons). We used headphone reproduction, AKG K 240 DF was selected. By using a 6 channel headphone amplifier Behringer (Model HA 903) was it possible to use 6 subjects simultaneously.

**RESULTS**

Normally we use a frame length of about 30 msec in the PARTRAN system. We believe that this parameter is crucial. Therefore we decided to investigate how the frame length will influence the speech quality. Frame lengths between 20 msec and 100 msec were investigated.

The mean-ratings were in the range from 3.8 to 4.2 (perceptible, but not annoying) for frame lengths between 20 msec and 60 msec. A small decrease in quality was observed for frame lengths in the range from about 60 msec to 100 msec.
The frame shift is normally about 5 msec. The importance of this parameter was also investigated. This parameter is not critical in the range from 5 msec to 15 msec.

PRELIMINARY ASSESSMENT OF SPEECH ENHANCEMENT

The PARTRAN system provide us with a flexible method to improve the signal-to-noise ratio, by increasing the Q-factors for each 2. order section. We know that increase in Q-factors will improve the signal-to-noise ratio. But unfortunately increasing Q-factors are expected to decrease the sound quality (caused by spurious sounds). To investigate the optimal selection of Q-factors we carried out a number of informal listening test (only the authors). We investigated also the benefits obtained by increasing the frame length in the case of noise added to the speech signals.

Experimental conditions
In the following experiments we added white noise, transmitted through a 6 dB/octave filter, to a clean speech signal. The signal-to-noise ratio was selected to 6 dB. We increased the order of the LP-model from 12 to 20. The FIR re-synthesis filter was of order 300. We used a constant absolute bandwidth for each transformed 2. order section. (Q-factors proportional with the center frequencies). All formants with bandwidth below 240 Hz were discarded.

RESULTS
As expected the signal-to-noise ratio was improved when the bandwidth decreased, but we also got increasing problems with disturbing spurious sounds. A bandwidth about 50 Hz seems to be a reasonable compromise between improvement of the signal-to-noise ratio and an acceptable disturbance by spurious sounds.
We also observed that increasing frame length had a beneficial influence on the spurious sound level. Our preliminary listening experiment suggest that a frame length about 200 msec will be appropriate (6 db signal-to-noise ratio).

REFERENCES


-284-
SPATIAL RESOLUTION AND INTERACTIVE CHANGES OF HEAD-RELATED TRANSFER FUNCTIONS FOR VIRTUAL REALITY

Jesper Sandvad, Dorte Hammershøi
Acoustica Laboratory, Aalborg University, Aalborg Ø., Denmark

SUMMARY

The problems of necessary spatial resolution in Head-Related Transfer Function (HRTF) measurements and runtime exchange of HRTF filters in Virtual Reality are addressed in this paper. Objective analysis suggest that a very high spatial resolution is necessary, while experience with an implemented experimental system is, that a moderate resolution is sufficient. A crossfade method artificially increases resolution during exchange of HRTF filters, and experience with the implemented system supports this approach.

INTRODUCTION

The auditory part of VR systems is usually generated by headphone simulation. The technique used for simulation is to convolve a sound signal with the transfer function of the human head and ears, the Head-Related Transfer Function (HRTF). An HRTF is specific for the angle of incidence of the sound, which means that each sound wave that is to be simulated (direct from source to listener, or reflected by walls etc.) requires a specific set of HRTFs. The principles are well known, and more thorough explanations of the technique can be found in the literature (from our own laboratory e.g. [1], [2]).

The term Auditory Virtual Environment, abbreviated AVE, is often used to describe the auditory part of the VR system specifically. It is used for the remaining part of this paper to describe the headphone simulation.

In AVEs all directions of sound incidence are possible, and HRTFs should therefore - in principle - be known for all directions. HRTF measurements can only be carried out for a finite number directions, but if the directions included in the measurements are sufficiently close in space, directions in between can be predicted with reasonable accuracy. How close this needs to be is the first issue of this paper.

In the VR system the (virtual) sound sources and the subject are usually allowed to move, and the sound must be generated accordingly. This means that not only must the HRTFs in the VR system be close enough in space to fully represent all directions, they must also be close enough to enable a smooth transition from one to another. The problem of exchanging filters during runtime is the second issue of the this paper.

The present paper only introduces the problems of spatial resolution and the replacement of HRTFs during runtime. The approach is to discuss properties of the HRTFs objectively. An experimental implementation of the headphone simulation has been made, and observations from pilot experiments is made in the final part of the paper.
Figure 1 Amplitude and phase of dummy head measurements shown as function of azimuth angle for three different frequencies. Each point corresponds to a measurement.

SPATIAL RESOLUTION OF MEASURED HRTFs

It is known that HRTFs vary between subjects (e.g. [3]), and further that the best localization is achieved with one's own 'ears' [4]. An analysis of the spatial resolution should thus be performed for individual HRTFs, but measurements of sufficiently fine resolution for humans are extremely difficult. The spatial resolution of HRTFs is therefore in the present section examined using measurements of fine spatial resolution from an artificial head.

The measurements were performed for directions in the horizontal plane, with an angular resolution of 2.8°. Figure 1 shows the amplitude and phase of the measurement as a function of azimuth angle for 0.9 kHz, 3.9 kHz and for 8.1 kHz. It can be seen that for 0.9 kHz, the amplitude and the phase varies only slightly between directions. It can thus be assumed that the spatial 'sampling rate' is sufficient. For 3.9 and 8.1 kHz rather large and abrupt differences in amplitude are seen for azimuth angles between -120° and -45° (the 'shadow side'). These differences are caused by constructive and destructive interference as the sound travels along different paths around the head. The curves for phase are continuous for all frequencies.

While the phase functions are in general very smooth, and could be sampled with a lower resolution with little or no loss of information, it can be seen that for frequencies above a few kHz, we cannot be sure that the amplitude functions are sufficiently sampled on the 'shadow side'. This means that we cannot, without a priori knowledge on the structure of the HRTF's, recreate the amplitude for missing HRTFs from measurements as close as 2.8°.

Figure 2 illustrates the impact of measuring with a poorer resolution (22.5°), and predicting HRTFs for directions in between by linear interpolation on a decibel basis. It can be seen that the amplitude is
predicted excellently up to more
than 10 kHz by the linear interpo-
lation for the ear closest to the
sound source, but poorly above just
a few kHz for the other ear.

These results do not look encourag-
ing, but the angular area with poor
prediction is also an area with poor
angular resolution of the human ear:
The localization blur at 90° sound
incidence is in the order of 10°.
This suggests that the ear does not
use the fine structure of the 'sha-
dow side' HRTF for localization.
The coarse interpolation shown in
Figure 2 may therefore be non-
obstructive for localization abilities
and naturalness in an AVE.

RUNTIME EXCHANGE OF HRTF
FILTERS

In the real world, HRTFs change in
a continuous manner with the direc-
tion to the sound source. This can
be simulated in a digital signal
processing system by sampling the
position of the virtual sound sour-
ces in time (as the subject moves
around in the virtual environment),
and exchange HRTF filters accord-
ingly. If the simulation has a suf-
fi ciently high resolution in time and
space, it will be perceived as con-
tinuous, but if the HRTF filters are too diverging spectrally, discontinuities in time may occur. The
discontinuities are perceivable as transients, or 'clicks', especially in narrowband sound signals. In Figure 1 it is seen, that even in the mid frequency range, the amplitude changes can be abrupt even at small
direction changes. Even if this is not perceivable as change in direction, the instantaneous change in
spectral content may be perceived as a transient. A spatial resolution of 2.8° thus seems far from enough.

Demands for spatial resolution much higher than shown in Figure 2 is however not realistic with current
technology, and a workaround must be found. If it is assumed, that any HRTF between two measured
HRTFs can be generated by a linear interpolation in time, the spatial resolution during a filter exchange
can be increased artificially by generating a large number of intervening HRTFs during the change. This
corresponds to performing a crossfade between the output of consecutive filters. As an effect, all
discontinuities disappear. This is illustrated in Figure 3.

It should be noted that the above assumption is generally not valid, and is only a reasonable approximation
for spatially close HRTFs. The crossfade corresponds to a linear interpolation in time, which is the same
as a linear interpolation of complex frequency values (but not the same as the interpolation of amplitude
in Figure 2).
Figure 3 Illustration of the effect of crossfading on a 1000 Hz sinusoid. The bottom panels show the sum of the upper panels, left side: no crossfading, right side: 0.8 ms crossfading.

The length of the transition period should in principle depend on the speed of head movement. The speed of head movement can only be estimated from previous position measurements, and a time variable delay is therefore unavoidable.

CURRENT IMPLEMENTATION

An experimental AVE system capable of free field simulation of a single sound source at any direction has been implemented with off-the-shelf DSP hardware in a PC. The system contains several HRTF sets measured in 22.5° resolution and interpolated to 3.5° covering the whole sphere. Crossfading filters is implemented with a length of up to 256 samples, corresponding to 5.33 ms. No transient noise is perceived when the crossfading is active, and the movement of virtual sound sources are perceived as smooth and continuous.

ACKNOWLEDGEMENTS

Economic support from the Danish Technical Research Council and the European Communities (ESPRIT Basic Research project #6358, SCATIS) is greatly acknowledged. The dummy head measurements were performed by Clemen Boje Jensen, Sofus Birkedal Nielsen and Claus Vestergaard Skipper.

REFERENCES

COMPARISON OF EIGHT PSYCHOMETRIC METHODS FOR MEASURING THRESHOLD OF HEARING

Michael Friis Sørensen, Morten Lydolf, Peder Chr. Frandsen, Henrik Møller
Acoustics Laboratory, Aalborg University, Aalborg Ø, Denmark

SUMMARY

The aim of this investigation was to examine the significance of the psychometric method used for measuring thresholds of hearing. Eight psychometric methods were studied, of which six were adaptive. The group of adaptive methods consisted of the method of adjustment, the PEST method, and four methods of limits, including the ascending and bracketing methods as described in ISO 8253-1, a descending method, and a fixed-frequency Békésy method. Two non-adaptive methods, methods of constant stimuli, were included in the investigation, one in the classical form with direct detection, and one implemented as a 2-AFC procedure.

24 otologically normal persons participated, and hearing thresholds were determined in a free field for pure tones at the frequencies 500, 1000 and 2000 Hz. Three repetitions were made. The results remain to be analyzed at the time of submission.

INTRODUCTION

ISO 226 [1] contains reference data for the binaural hearing threshold of pure tones in a free field. Even in the latest version from 1987, the free field data are exclusively based on investigations made by Robinson and Dadson in 1956 [2]. In 1988 ISO Working Group TC 43/WG 1 “Threshold of hearing” decided to revise the reference threshold data, and new values based on data from ISO 226 and from 6 recent studies were issued in 1994 as ISO 389-7 [3]. However, the differences between threshold data from different studies are so large that they cannot be explained from statistical variation. One possible reason for discrepancies is the psychometric method. The aim of the present investigation was to examine the significance of the psychometric method used for measuring hearing thresholds.

METHOD

Eight psychometric methods were included, of which six were adaptive. For each subject, the threshold was determined at the frequencies 500 Hz, 1000 Hz and 2000 Hz, using the six adaptive methods. Following this, the two non-adaptive methods were used to determine the threshold at 1000 Hz. These methods need a fair estimate of the threshold as input, and for this purpose the individual mean of the thresholds obtained in the adaptive methods was used.

The order of the adaptive methods was balanced in a latin square design, while the order of the two non-adaptive methods was randomized. Threshold determination with the whole range of methods was carried out three times, two times at one day and one time a few days after. New latin squares for the adaptive methods and new random choices of the order of the non-adaptive methods were used for the repetitions.

- 289 -
Some of the methods looked identical to the subjects, as their task was the same. Each time a method with a new task was used with a subject, a task familiarization was carried out comprising a complete threshold determination at 1000 Hz. The result of this was not used. In all methods each threshold determination commenced with a tone familiarization, implemented by starting the presentations 35 dB above the threshold of ISO 226. Data from the tone familiarization provided adequate inputs to start some of the adaptive methods. In general, the methods were implemented in rather long versions, probably longer than needed for a satisfactory threshold determination. This was done to facilitate a study of various ways of threshold estimation, including maximum likelihood estimation. Examples of the course of the methods are given in Figure 1.

**Bracketing method.** The tone familiarization included a decent and an ascent, both having a step-size of 10 dB. The ascent started 7.5 dB below the first level in the descent, at which the subject did not respond. The bracketing method itself comprised 6 sets ascents and descents, both having a step-size of 5 dB. Ascents started 5 dB below the first level in the preceding descent, at which the subject did not respond, and descents started 7.5 dB above the first level in the preceding ascent, at which the subject responded.

**Ascending method.** The tone familiarization was as in the bracketing method. The ascending method itself comprised 12 ascents with a step-size of 5 dB. Each ascent started 7.5 dB below the level, at which the first response occurred in the preceding ascent.

**Descending method.** The tone familiarization was as in the bracketing method. The descending method itself comprised 12 descents with a step-size of 5 dB. The descents started 7.5 dB above the first level, at which no response occurred in the preceding descent.

The bracketing and ascending methods were implemented nearly as described in ISO 8253-1 [4]. However, a step-size at or below 2.5 dB was specified in preferred test conditions given by WG 1 [5]. Preliminary experiments showed, though, that the resulting low rate of change in level might appear undesirable for the subject. As it was believed that the low step-size was specified by WG 1 as a means of obtaining a high accuracy, it was decided to use a step-size of 5 dB, but to maintain presentations with 2.5 dB intervals by interlacing the ascents and descent by using the "odd" steps of 7.5 dB after a descent or an ascent.

**Parameter Estimation by Sequential Testing (PEST).** The method was originally described by Taylor and Creelman [6], but a modified version was used as described by Findlay [7]. The step-size commenced at 10 dB, then repeatedly halved down to 0.3 dB. With the step-size of 10 dB, W was 0.5 to obtain the tone familiarization, thereafter W was increased to 1.

**Method of constant stimuli with direct detection.** The tone familiarization consisted of a descent with four stimuli and a step-size of 10 dB. 91 levels were equally distributed in the interval ±12.5 dB around the threshold estimate. The selected levels were randomized before presentation.

In the five methods described above the subjects were instructed to press a button, each time they heard a tone. The duration of the tone was 1 s, and the response accept period was from tone start to 0.5 s after tone stop. The delay between presentations was randomized in the interval 0.75-4 s, if response occurred commencing after the response or at the termination of the tone, whichever was the latest, or if no response was given commencing at the
termination of the accept period. However, in the ascents of the bracketing and the ascending methods, and in the ascents of the tone familiarization of the descending method, presentations were only delayed 0.25 s, since the unheard tones were believed to be perceived by the subjects as a delay.

Method of constant stimuli with two alternative forced choice (2-AFC) detection. The tone familiarization was as in the method with direct detection. 61 levels were equally distributed in the interval ±12.5 dB around the threshold estimate. Levels below the threshold estimate were duplicated, resulting in a total of 91 levels. The selected levels were randomized before presentation. Tones with a duration of 1 s were presented during either a red or a green period, indicated by lights and separated by a pause of 0.5 s. The order of the lights was fixed. 0.25 s after the green period, both lights turned on, and the subjects was to answer.

Békésy method. The method was implemented as a fixed-frequency method in accordance with ISO 8253-1 [4]. The tone familiarization consisted of a descent with a step-size of 2.5 dB. The method itself consisted of 8 ascents and descents with a step-size of 1.25 dB. The tone duration was 250 ms, followed by a pause of 250 ms. The corresponding rates of level change were 5 dB/s and 2.5 dB/s. The subjects were instructed to keep the answer-button pressed, when the tones were audible.

Method of adjustment. In this method the tone was continuous. The subjects were instructed to use a volume control (a multi-turn dial) and turn down the level below audibility, then up and down to encircle their threshold and mark by pressing a button, when the tone was just audible. The procedure was carried out six times.

The tones were presented in an anechoic room with the subjects facing a loudspeaker placed behind a curtain. The sound field, harmonic distortion, frequency stability etc. were in accordance with the preferred test conditions [5]. 10 female and 14 male subjects participated. The age range was 19 to 25 years with an average of 23.4 years. Questionnaires were answered by all subjects to ensure that none had a history that would suggest an impaired hearing. Standard audiometry was made for documentation. The subjects were asked not to be exposed to noise the last two days before the experiments were made. All subjects were found otologically normal in a medical examination with otoscopy and tympanometry on the days of experimentation. The subjects were paid for their participation.

The results remain to be analyzed at the time of submission.

5. ISO/TC 43/WG 1/N 122: Preferred test conditions for the determination of the minimum audible field and the normal equal-loudness level contours. 1988.
Figure 1. The course of the 8 psychometric methods, shown for one subject, first repetition at 1000 Hz. For the first 5 methods '+' indicates a response and 'o' indicates no response. In the 2-AFC method '+' indicates a correct response and 'o' symbolizes a wrong response. In the Békésy method '+' indicates, when the subject presses the button, and 'o' indicates, when he releases it. An asterisk in the adjustment method symbolizes the subject's threshold mark. The scale of the abscissa is not the same for all methods.
DIRECTIONAL DEPENDENCE OF LOUDNESS AND BINAURAL SUMMATION

Michael Friis Sørensen, Morten Lydolf, Peder Chr. Frandsen, Henrik Møller
Acoustics Laboratory, Aalborg University, Aalborg Ø, Denmark

SUMMARY

Present noise measurements are based on a single microphone with equal weighting of sound from all directions. This technique does not take into account directional dependence in the complex human binaural sound perception. The importance of this problem is not known, since the variation of loudness with direction to the sound source has not been thoroughly examined. Calculations based on knowledge about sound transmission to the ears and the assumption that the brain summarizes the binaural signals on a simple power basis, indicate that more than 10 dB deviation in sensitivity relative to the front direction could be expected. In the present ongoing investigation binaural threshold and loudness data for the frequency range 500-12500 Hz will be determined for 6 directions to the sound source. Both monaural and binaural thresholds and loudness will be determined for 12 test subjects enabling examination of the binaural summation. Altogether, these data will enable an evaluation of the deficiency of the present noise measurement technique, when spatial aspects are considered, including an assessment of the necessity of introducing artificial heads for future measuring standards.

INTRODUCTION

Standards for measuring noise for estimation of its annoyance or loudness prescribe use of a single microphone with equal weighting of sound from all directions. This technique does not take into account that humans perceive sound with two ears, each performing a complex weighting of the sound depending on direction to the sound source. However, the significance of this is not known precisely, since the variation of loudness with direction to the sound source is only very briefly described in the literature [1], [2]. As a result of this, the present standard concerning equal loudness contours, ISO 226 [3], only contains data for frontal and random sound incidence. In the audiometric standard, ISO 8253-2 [4], reference hearing thresholds for 45° and 90° azimuth are given only in terms of corrections derived from the physical transmission to the closest ear. Thus, the binaural summation is not described, and the true binaural thresholds are not known.

It is the aim of the present investigation to obtain data for the directional dependence of hearing thresholds and perception of loudness.
ESTIMATION OF DIRECTIONAL DEPENDENCE OF LOUDNESS

An estimate of the directional dependence can be calculated assuming that the hearing summarizes the signals at the ears on a simple power basis - corresponding to summation of uncorrelated signals. The calculation requires knowledge on the physical sound transmission to the ears from various directions of sound incidence, that is the Head-Related Transfer Functions (HRTFs). HRTFs were previously measured on 40 subjects at our laboratory [5].

![Graph showing directional dependence of loudness](image)

**Figure 1**

The thin curves show HRTFs for both ears of a single subject. The dashed curve is the binaural sum calculated on a simple power basis. The bold curve shows this power sum relative to that for the front direction. For each set of curves the direction of sound incidence is indicated.

Figure 1 shows the HRTFs for both ears of a single subject with sound from various directions (thin lines). As it is seen, there is a considerable directional dependence of HRTFs for each ear, with up to 40-50 dB difference between ears at high frequencies. The directional dependence of the calculated binaural power sums (dashed lines), is much less. For the directions in front and above the HRTFs are nearly identical at the two ears, and it is obvious that the power sum corresponds to the assumption that we are 3 dB more sensitive when

- 294 -
listening binaurally as compared to monaural listening. For the left direction the difference between the ears is large, especially at high frequencies. In this case the power sum is dominated by the ear with the highest level corresponding to the assumption that the sensitivity equals the sensitivity of the ear nearest the sound source.

Also shown in Figure 1 is the binaural power sum relative to that for the front direction (bold lines). From this it is seen that for the frequency range 6-10 kHz we could expect a sensitivity around 10 dB higher with sound coming from left and above than from front. On the other hand at 2-5 kHz and above 12 kHz the front direction is the most sensitive of the directions presented.

Data for 40 subjects have been inspected and the characteristics described are valid in general.

PSYCHOACOUSTIC EXPERIMENTS

It is interesting that, with the assumptions described in the previous section, there may be directions of sound incidence, where we are 10 dB more sensitive than for frontal incidence - at certain frequencies even more. This means that the annoyance of noise from these directions may be underestimated using descriptions of hearing sensitivity based on frontal sound incidence. This is an obvious reason for collecting psychoacoustic data in order to test, whether the assumptions about the binaural summation are valid.

In the ongoing investigation binaural thresholds and 60 phon loudness data will be determined for 12 test subjects and for 6 directions. In order to test the described assumptions about binaural summation the HRTFs as well as monaural thresholds and loudness will be determined. To eliminate differences in sensitivity of the subjects' two ears, monaural data for both left and right ear are required. As seen in Figure 1 the interesting frequency range is above 1 kHz. Therefore, 9 pure tones in the range 500 to 12500 Hz are included in the study.

![Figure 2](image)

*Figure 2*

Preliminary threshold data for a single subject with sound from the left side and from the front. The asterisks indicate the difference in pure tone thresholds, while the solid curve shows the calculated difference in binaural power sum, based on measured HRTFs.

Figure 2 shows the result of a pilot experiment where the binaural thresholds were determined for a single subject. The figure shows thresholds for sound from the left relative to that for
the front direction (asterisks). In addition, the calculated binaural power sum relative to that for the front direction is shown, based on the subject's HRTFs. For this single example there seems to be a fair agreement between the change with direction in sound transmission and the corresponding change in sensitivity at threshold.

However, if the binaural power assumption is to be evaluated, precise and efficient methods for determining thresholds and equal-loudness are needed. Therefore, the choice of psychometric threshold procedure for the experiments will be based on a thorough investigation including 8 different methods [6].

CONCLUSION

Based on measurements of the physical transmission to the ears from a certain direction [5] the binaural sum has been calculated using simple power summation. Assuming that the human brain summarizes the signals at the ears in a similar way, these calculations can be used to estimate the directional dependence of the binaural hearing sensitivity. The calculation have revealed that for frequencies above 4-5 kHz more than 10 dB difference could be expected relative to frontal sensitivity. A brief pilot experiment has verified this result.

Therefore, binaural as well as monaural threshold and loudness data will be determined for different directions of sound incidence enabling an evaluation of the spatial aspects of the present noise measuring technique.

REFERENCES

REFERENCE EQUIVALENT THRESHOLD SOUND PRESSURE LEVELS FOR NEW EARPHONES

Hisashi Takeshima, Tomoki Hiraoka, Masazumi Kumagai,
*Toshio Sone, Yōiti Suzuki
Sendai National College of Technology, Kamiyashi, Aoba-ku, Japan.
*Research Inst. Electrical Communication, Tohoku University, Sendai, Japan.

SUMMARY

Equivalent threshold sound pressure levels, ETSPL, for four audiometric earphones, TDH-39P, TDH-49P, HDA-200, and NEDO-H3 were measured to provide data for revision of ISO 389. The ETSPL's of TDH-39P and TDH-49P were found to agree well with the RETSPL specified in ISO 389, while the ETSPL of NEDO-H3 differed from it below 400 Hz. The ETSPL of HDA-200 did not remarkably differ from that obtained by PTB.

In addition, to examine the dependence of the ETSPL on the earphone, transformations from the SPL in an artificial ear to the SPL in a real ear, as well as the noise level in a subject's ear caused by the wearing of an earphone, were measured. The ETSPL translated to the SPL in a real ear showed good agreement with the RETSPL above 500 Hz. This indicates that the major factor of the dependence of the ETSPL on the earphone was the difference of the transformation level from the SPL in the artificial ear to the SPL in the real ear.

INTRODUCTION

The threshold of hearing in terms of the sound pressure level calibrated with an acoustic coupler or an artificial ear is called Equivalent Threshold Sound Pressure Level, ETSPL. ETSPL of a normal person is specified in the international standard ISO 389[1] as Reference Equivalent Threshold Sound Pressure Level, RETSPL.

In ISO 389, RETSPL's are specified for three kinds of earphones, namely, TDH-39, DT-48, and other supra-aural earphones. The intension of ISO/TC 43/WG 1 is to specify RETSPL's for other earphones.

To provide data for the revision of ISO 389, we measured ETSPL's for four audiometric earphones, TDH-39P, TDH-49P, HDA-200, and NEDO H3. In this paper, ETSPL's obtained hitherto are described. In addition, to discuss the causes of the dependence of the ETSPL on the earphone, we measured the transformation from SPL in an artificial ear to the SPL in a real ear as well as the noise level in a subject's ear.

MEASUREMENT OF ETSPL

Experimental conditions

The earphones used in this study were Telephonics TDH-39P, TDH-49P with MX41/AR cushion, Sennheiser HDA 200, and Ashida Sound NEDO-H3. TDH-39P, TDH-49P, and NEDO-H3 are supra-aural type earphones, while HDA-200 is a circum-aural type earphone.

The experimental conditions met the preferred test conditions proposed by ISO/TC 43/WG 1[2]. The bracketing method[3] was used to measure the threshold of hearing. The
sound pressure levels of the stimuli were changed at 2 dB steps. Descending runs and ascending runs were carried out seven times alternately. The last six threshold levels in the seven trials were averaged and adopted as the threshold of hearing of the subject.

The stimuli were pure tones of 1 second duration inclusive of 50 ms rise/decay time. The frequencies were the 1/3 octave series specified in ISO 266 and audiometric frequencies between 125 Hz and 8 kHz for TDH-39P, TDH-49P, NEDO-H3, and such frequencies between 125 Hz and 16 kHz for HDA-200.

There were 19 subjects, 10 males and 9 females, 18 to 23 years of age. They all had monaural hearing thresholds with hearing levels below 10 dBHL for the frequencies 125 – 8000 Hz and normal tympanograms with peaks within ±50 daPa. Only one ear was measured in each subject.

Results

Figure 1 shows averaged ETSPLO's for the TDH-39P earphone calibrated by the IEC 303 acoustic coupler. From the figure, we can see that the averages agree well with the RETSPL for TDH-39 specified in ISO 389.

Figure 2 shows averaged ETSPLO's for supra-aural earphones, TDH-39P, TDH-49P, and NEDO-H3. The ETSPLO's of TDH-39P and TDH-49P are in good agreement with the RETSPL for supra-aural earphones specified in ISO 389, while the ETSPLO's for NEDO-H3 are obviously different from the RETSPL below 500 Hz.

Figure 3 shows averaged ETSPLOs for a circum-aural earphone, HDA-200. There is no remarkable difference between our ETSPLO's and those obtained by PTB.

MEASUREMENT OF SPL IN A REAL EAR AND NOISE LEVEL IN A SUBJECT'S EAR

Sound pressure level in a real ear

We measured the transformation from SPL in an artificial ear to SPL in a real ear. A small-size electret condenser microphone, RION EU-22 (3.6mm × 3.6mm × 2.1mm), was set at the entrance of the ear canal for measurement of the sound pressure level in the real ear. Random noise was presented through the earphones, then the sound pressure level was measured at the entrance of the ear canal and in an IEC 318 artificial ear. The SPL at the entrance of ear canal was measured five times in each subject after refitting the earphone at each measurement. The transformation in each subject was obtained by subtracting the SPL's in the artificial ear from the average of the five SPL's at the entrance of the ear canal. Eleven of the 19 subjects participated in this measurement.

Figure 4 shows means and standard deviations of the transformations for the four types of earphones. The transformations in TDH-39P and TDH-49P decrease for 600 Hz and below. The transformation in NEDO-H3 also decreases in the same frequency range but the decreases are smaller than those in TDH-39P and TDH-49P. However, the transformation in HDA 200 is almost zero dB below 1 kHz. Moreover, standard deviations in HDA-200 and NEDO-H3 are smaller than those in TDH-39P and TDH-49P below 600 Hz.

Noise level in a subject's ear

We measured the noise level in a subject's ear caused by the wearing of an earphone. The noise levels were measured at the entrance of the ear canal with an EU-22 microphone when the subject was wearing an earphone and when he/she wasn't. Seven subjects participated in this measurement.

Figure 5 shows the averaged 1/3 octave band noise levels with and without the earphone in place, system noise levels of the microphone system, and the ambient noise in the anechoic room measured with a low noise B&K 4179 microphone. In the frequency range below 100
Figure 1: ETSPL for TDH-39P calibrated with the IEC 303 acoustic coupler.

Figure 2: ETSPL for the supra-aural earphones, TDH-39P, TDH-49P, and NEDO-H3, calibrated with the IEC 318 artificial ear.

Figure 3: ETSPL for the circum-aural earphone, HDA-200, calibrated with the IEC 318 artificial ear.

Figure 4: Transformation levels from SPL in the artificial ear to SPL at the entrance of the real ear canal.

Hz, there are clear differences between noise levels when the subject was wearing the earphone and when he/she wasn’t. From these results, we concluded that the measurements of the noise levels caused by the wearing of an earphone succeed below 100 Hz.

**DISCUSSION**

The results of ETSPL measurement shown in Fig. 2 and Fig. 3 indicate that the ETSPL depends on earphones. This dependence of the ETSPL can be regarded as being caused by the following two reasons. (a) The transformation from the SPL in the artificial ear to the SPL in the real ear depends on the type of earphone used. (b) The amount of masking caused by noise in the subject’s ear also depends on the type of earphone used. From Fig. 4, obvious differences can be seen in the transformations between the four earphones. From Fig. 5, it is apparent that the noise levels also depend on the earphones.

ETSPS’s for each subject were converted into thresholds of hearing in the SPL at the entrance of the ear canal by use of the transformations from the SPL in the artificial ear.
Figure 5: Noise levels in the subject’s ear. Figure 6: Threshold of hearing in SPL at the entrance of ear canal.

to the SPL at the entrance of ear canal. Figure 6 shows the averages of the converted thresholds of eleven subjects. We can see that the thresholds with all earphones are in good agreement, especially above the frequency of 500 Hz. This indicates that the major factor of the dependence of the ETSPL on the earphone above 500 Hz is the transformation from the SPL in the artificial ear to the SPL in the real ear. Below 400 Hz, on the other hand, we can see some systematic differences of about 4 - 10 dB. These differences are attributable to the differences of the noise in subject’s ear, because the noise levels in the ears of subjects with HDA-200 and NEDO-H3 are 5 - 10 dB higher than those with TDH-39P and TDH-49P. This evidence is insufficient, however, to explain the differences in the converted thresholds, since we don’t know whether the noise level in the subject’s ear below 100 Hz affects the thresholds of hearing up to 400 Hz or not.

CONCLUSION

We measured ETSPL’s for four audiometric earphones. The ETSPL’s of TDH-39P and TDH-49 agreed well with the RETSPL, while that of NEDO-H3 was different from the RETSPL below 400 Hz. There were no remarkable discrepancies between the ETSPL’s of HDA-200 and the results obtained by PTB.

In addition, we examined the dependence of ETSPL on the earphone. It was confirmed that the major factor of the dependence above 500 Hz was the transformation from the SPL in an artificial ear to the SPL in a real ear, and it was suggested that the factor of the dependence in the low frequency range was the noise level in subject’s ear caused by the wearing of earphone.

ACKNOWLEDGMENTS

We wish to thank Mr. Ikuo Ohira of Ashida Sound Co., Ltd., and Mr. Eiich Nonomura of Matsushita Communication Industrial Co., Ltd., for supporting this research, and Mr. Seiji Hori of Rion Co., Ltd. for providing the small size microphones.

REFERENCES
SOME EXPERIMENTAL RESULTS FOR A FULL-SCALE REVISION OF EQUAL-LOUDNESS LEVEL CONTOURS

Hisashi Takeshima, Masazumi Kumagai, Yōiti Suzuki*, Hideyuki Fujii*, Toshio Sone*, Takeshi Fujimori**
Sendai National College of Technology, Sendai, Japan
*Research Inst. Elec. Communication, Tohoku University, Sendai, Japan
**Electrotechnical Laboratory, 1-1-4, Umezono, Tsukuba 305, Japan

SUMMARY

To revise the equal-loudness level contours specified in ISO 226, data of five experimental studies, including that conducted by the present authors, have been reported recently. This paper describes an estimation of new equal-loudness level contours based on these experimental data. A model for equal-loudness level contours derived from a loudness function is herein proposed.

INTRODUCTION

The equal-loudness level contours and the thresholds of hearing for pure tones in a free field are specified in the international standard ISO 226. Since a new finding which was slightly different from ISO 226 at around 400 Hz was presented at a meeting of ISO/TC 43/WG 1 in 1985, a full-scale revision of ISO 226 was decided at ISO/TC 43/WG 1. We have been conducting a series of experiments to obtain data from Japanese subjects since 1986[1,2]. Similar experiments have been conducted by Betke and Meller(1989)[3], Fastl et. al.(1990)[4], Watanabe and Møller(1990)[5], and Poulsen and Thøgersen(1994)[6].

At the final stage of this project, we should be able to determine the contours that best describe equal-loudness levels at any loudness levels and at any frequencies. In this paper, we propose a model for equal-loudness level contours based on the loudness function. Moreover, we attempt to estimate equal-loudness level contours applying this model to our data as well as to all the recent data mentioned above.

MODEL OF EQUAL-LOUDNESS LEVEL CONTOURS

The data obtained by the experiments are discrete and include some errors. To estimate the equal-loudness level contours from the discrete data, therefore, interpolation is needed. Moreover, to estimate reliable contours from limited data, it is important to apply good knowledge of loudness perception to the data processing.

As to the representation of the relation between sound pressure of a pure tone and its loudness, there have been several proposals [7–12]. For example, Lochner and Burger[8] proposed a loudness function as follows:

\[ S = k(I^\alpha - I_0^\alpha), \]  

where \( S \) is the loudness of a pure tone, \( I \) is its intensity, \( I_0 \) is the threshold of hearing in terms of intensity, and \( \alpha \) and \( k \) are constants which are functions of the pure tone frequency.

In this paper, equation (1) is used as the loudness function to describe full-scale equal-loudness level contours. The equation agrees well with the experimental results [11] in spite of its simple form.
It is assumed in this paper that the exponent $\alpha$ is approximately 0.27 for a 1 kHz pure tone [8]. Hence, the loudness of a 1 kHz pure tone, $S_n$, is represented by

$$S_n = k_n(I_n^{0.27} - I_{0n}^{0.27}),$$

(2)

where $I_n$ is the intensity of a 1 kHz pure tone and $I_{0n}$ is its threshold of hearing in intensity. On the other hand, the loudness of an $f$ Hz pure tone, $S_f$, is represented by the following equation:

$$S_f = k_f(I_f^{0.27} - I_{0f}^{0.27}),$$

(3)

where $I_f$ is the intensity of an $f$ Hz pure tone and $I_{0f}$ is its threshold of hearing in intensity. When the loudness of an $f$ Hz pure tone is equal to that of a 1 kHz pure tone, the following equation stands:

$$k_f(I_f^{0.27} - I_{0f}^{0.27}) = k_n(I_n^{0.27} - I_{0n}^{0.27}).$$

(4)

This equation can be rewritten as

$$I_f = [K_f(I_n^{0.27} - I_{0n}^{0.27}) + I_{0f}^{0.27}]^{1/\alpha_f},$$

(5)

where

$$K_f = k_n/k_f.$$

If these intensities are converted to sound pressure levels, equation (5) represents the relation between sound pressure level for an $f$ Hz pure tone and its loudness level. In other words, this is a proposed model for describing equal-loudness level contours.

**ESTIMATION OF EQUAL-LOUDNESS LEVEL CONTOURS**

Equal-loudness level contours were estimated according to the following procedure:

1. The sound pressure levels at the threshold of hearing are smoothed with B-spline functions.
2. $\alpha_f$ and $K_f$ are estimated by applying equation (5) to the experimental data of the equal-loudness levels with the method of non-linear least squares.
3. The $\alpha_f$ estimated in the previous process are smoothed with B-spline functions.
4. $K_f$ are re-estimated with the method of least squares, after substituting $\alpha_f$ obtained in the previous step.
5. Equal-loudness level contours are calculated from equation (5) in which the estimated $\alpha_f$, $K_f$, and smoothed threshold of hearing curve are used.

Two sets of equal-loudness level contours were derived with the procedure described above. One of them was estimated with only the data we obtained [1,2]. The other was estimated using the data obtained by various researchers [3–6], including us, who all are engaged in the revision of ISO 226.

Figures 1 and 2 show the results: (a) and (b) are estimated $\alpha_f$ and $K_f$ in equation (5), respectively, and (c) illustrates the equal-loudness level contours up to 90 phon as well as the thresholds of hearing. Figure 3 shows an example of the relation between sound pressure level and its loudness level.

Figure 4 shows the equal-loudness level contours estimated above as well as those specified in the present ISO 226. There are few differences between the estimated two sets of contours, while large discrepancies are clearly seen between the estimated contours and the present standard, especially in the frequency range below 1 kHz.
Figure 1: Estimation of the equal-loudness level contours that was applied to our experimental data. (a) Estimated $\alpha_f$ and its smoothed curve. (b) Estimated $K_f$ and its smoothed curve. (c) Estimated equal-loudness level contours as well as experimental data with 95% confidence limits.

Though we did not use a model to describe frequency characteristics, Figures 1(a), 1(b), 2(a) and 2(b) show that $\alpha_f$ and $K_f$ change smoothly according to frequency. Therefore, we think it reasonable to join these parameters by a smooth curve using B-spline functions.

CONCLUSION

We have proposed a model for describing equal-loudness level contours based on the loudness function shown by equation (1). Moreover, equal-loudness level contours were estimated by applying the model to our data as well as to all the data obtained recently. While the two sets of contours closely resemble each other, both of them are very different from the contours specified in the present ISO 226. The present results show that ISO 226 should be revised.

REFERENCES
Figure 2: Estimation of the equal-loudness level contours that were applied to the data obtained by five researchers. (a) Estimated $\alpha_f$ and its smoothed curve. (b) Estimated $K_f$ and its smoothed curve. (c) Estimated equal-loudness level contours as well as experimental data obtained by five researchers.

Figure 3: An example of the relation between sound pressure level of a pure tone and its loudness level applied estimated $\alpha_f$ and $K_f$ as well as experimental data.

Figure 4: Comparison of the equal-loudness level contours.
THE LOW DICHOTIC PITCH OF RANDOMLY LATERALIZED HARMONICS

A. Pjot van der Meulen, Frans A. Bilsen, Johan Raatgever
Delft University of Technology, Faculty of Applied Physics, Lab. of Perceptual Acoustics, P.O. Box 5046 2600 GA Delft, The Netherlands.

SUMMARY
Most modern models dealing with binaural hearing assume the generation of a two-dimensional activity pattern as a function of frequency and internal delay. The central spectrum model states that the pitch of a dichotic complex pitch stimulus is determined by grouping of harmonics across frequency for equal internal delay. In order to investigate whether the scanning is performed in this way, dichotic complex pitch stimuli with prominent peaks for several equidistant frequencies (harmonics) and random internal delays were made. If scanning takes place along a straight line (constant internal delay) the observed pitch is expected to be significantly different from the low pitch as a result of grouping of peaks for different internal delays. Pitch and saliency matchings have been carried out for several stimuli. The results of the pitch matchings suggest that it is possible to group peaks for different internal delays. The improved saliency for the pitch stimuli with peaks at a constant internal delay gives rise to the assumption that the ‘straightness’ of the line connecting the peaks is an additional cue for the binaural interaction system.

INTRODUCTION
Most recently developed theories about binaural interaction share some basic principles first described by Jeffres (1948). They assume that after peripheral band filtering internal activity patterns will arise, which are dependent on frequency and internal delay. The Central Spectrum (CS) theory, developed by Raatgever and Bilsen (1977, 1986), is based on linear systems theory. The theory supposes that the information at both ears is delayed internally after being filtered by identical peripheral band filters. Next, addition of the signals from both ears takes place. The signal power after the addition is a function of frequency and internal delay, and gives rise to a central activity pattern (CAP). This CAP is scanned by a central spectral pattern recognizer that selects spectra from it. In dichotic pitch, this selection is based on the pitch information in each spectrum. Each spectrum corresponds to a particular internal delay. An example of a CAP is given in Fig. 1A.

The MPS stimulus (Bilsen, 1976) consists of dichotic white noise with sharp interaural phase shifts of 2π at harmonic frequencies. For this stimulus the spectrum in the centre of the CAP
contains sharp peaks at harmonic frequencies. This spectrum is supposed to be selected and leads to a low pitch that is equal to the pitch of the fundamental, even when this fundamental has been filtered out. Consequently, the central spectrum model assumes that the pitch of a dichotic complex stimulus is determined by grouping across frequency for equal internal delay. This corresponds to scanning along a straight line (constant internal delay) in the CAP. Stern et al. (1988) use the term 'straightness' for the consistency of the internal delay over frequency.

In order to investigate whether this assumption is a valid one, a dichotic complex pitch stimulus is designed with harmonics more or less randomly scattered over the CAP. The central spectrum model expects that the central pattern recognizer selects spectra containing one of the harmonics. This results in an analytical pitch equal to the pitch of the harmonic. On the other hand, if the harmonics are grouped for different internal delays, the expected synthetical pitch is equal to the fundamental of the stimulus.

**PITCH MATCHINGS FOR DELAYED MPS STIMULI**

The dichotic complex pitch stimulus used is based on a MPS stimulus. In order to scatter the harmonics over the CAP, a large interaural (external) delay was applied to the MPS stimulus. By choosing a particular value for the external delay, the third and the sixth harmonic will be located in the centre of the CAP. The fourth harmonic will lie on one side of the CAP, while the fifth harmonic will lie on the other side. The other harmonics have been filtered out by a steep bandpass filter. The CAP of the first stimulus with a 100 Hz fundamental \( f_0 \) is shown in Fig. 1A. The external delay is 3.38 ms. The second MPS stimulus has a fundamental of 200 Hz and an external delay of 1.68 ms.

From the CS theory it follows that the spectra containing one of the harmonics have the best pitch information for this stimulus (type III). The expected analytical pitch will be equal to 3, 4, 5 or 6 times \( f_0 \). If peaks for different internal delays are grouped, the expected synthetical pitch will be equal to the fundamental \( f_0 \) of the MPS stimulus.

In order to stimulate the observer to listen in a synthetic mode, the stimulus was also presented without an external delay (type I). The expected low pitch is equal to the fundamental of the MPS stimulus. Another type of stimulus (type II) was an undelayed MPS stimulus with a fundamental (3\( f_0 \)) that is about three times (3\( f_0 \)=296 Hz, respectively 3\( f_0 \)=615 Hz) the fundamental of the stimuli of type I and III. In this case, the expected analytical pitch is equal to 3\( f_0 \) or 6\( f_0 \). All stimuli have been filtered by the same band filter and have been presented at a constant level of 50 dB SL. The three types of stimuli are schematically depicted in Fig. 2.

The pitch of the stimulus has been matched using a reference, consisting of a sinc wave of variable frequency added to uncorrelated white noise. This noise is in counter phase at both ears (antiphase). The percept of this reference is very distinct of the percept of the stimulus to be matched. A better reference seems to be a periodic impulse added to antiphase noise. This is not a valid option however, because it is possible that
the observer matches one of the reference components to one of the stimulus components. This reference, therefore, gives no definite answer that the observer listens synthetically. Typical results of the experiment are depicted in a histogram (Fig. 3), for two observers and a fundamental of about 100 Hz. The interval width is 1/8 f₀.

Observer PM matched f₀ for stimulus type I and III as well. For stimulus type II he matched mostly around 3 f₀, and sometimes 1.5 f₀. This, and other ratios between the expected and the matched pitch that are a power of two, can be explained by octave deafness. Observer FB could not always perform a matching, because of the faintness of the pitch. For stimulus type I he mostly matched around f₀ and 2 f₀. For stimulus type II he mostly matched around 3 f₀ and 1.5 f₀. Sometimes he matched 2 f₀. This may be caused by substituting two additional harmonics at 4 f₀ and 5 f₀. The few stimuli of type III he matched, he matched around f₀ and 2 f₀.

As can be seen from Fig. 1, the peaks of the harmonics are extended in the direction of the internal delay. This is a result of the periodicity of the CAP in this direction. The amplitude of these extensions diminishes for larger distances from the corresponding harmonic. The frequency of the extended peaks differs slightly from the frequency in the centre of the CAP. In a single spectrum, these extended peaks give rise to a low pitch, because their frequencies are almost harmonic. This low pitch is expected to be approximately equal to the fundamental f₀ of the MPS stimulus. This might also explain the low pitch found in the dichotic pitch matching experiments for stimulus type III. In order to find out what pitch a separate spectrum will lead to, 11 spectra of the CAP between -1 and 1 ms have been presented monaurally to the observers. For the delayed MPS stimulus (type III) with a 100 Hz fundamental these spectra are given in Fig. 1B. The monaural spectra can be constructed from this dichotic stimulus by adding the signals of both ears. The stimuli have been matched by a monaural reference consisting of a sine wave of variable frequency added to white noise. Typical results are depicted in Fig. 4 for two observers and a fundamental of 100 Hz.

The expectations for both analytical and synthetical listening are also presented in Fig. 4. The analytic expectations are based on prominent peaks around the dominant region of 600 Hz (Raatgever, 1980). In order to get a better correspondence between the expectations and the experimental results, the expectations are lowered or raised by one or two octaves. This is legitimate because of octave deafness.

The matched pitch values of observer PM indicate that he listened analytically. Apparently, this observer could not find enough information in the spectra to group some peaks. Observer FB matched both analytically and synthetically. He reported that the pitch was faint. This is the reason that he did not match for all presented stimuli delays.

The matchings are mostly below the expected pitch values. For the synthetic pitch matchings this can be explained by the fact that the pitch of complex tones with a low (lower than 1 kHz) fundamental is
lower than the pitch of a pure tone of the same frequency (Zwicker, 1990). The lower values of the analytic pitch matchings are probably caused by the fact that the frequency of the subjective half pitch value of a pure tone is smaller than the physical frequency of this pure tone divided by two (Zwicker, 1990). This implicates that the frequency of the subjective fourth pitch value of the expected analytic pitch is smaller than the physical frequency of the expected analytic pitch divided by four.

SALIENCY MATCHINGS FOR DELAYED MPS STIMULI

The 'straightness' of the line connecting the peaks might be an additional cue for the binaural interaction system. It possibly gives rise to a better saliency of the undelayed MPS stimulus (type I) versus the MPS stimulus with its harmonics scattered over the CAP (type III). For this reason also pitch strength or saliency measurements have been performed. The observer can adjust the signal-to-noise ratio of a reference to give it the same impression of pitch saliency as the low pitch of the MPS stimulus. The reference consists of a periodic impulse added to diotic white noise. The fundamental of the periodic impulse is equal to the fundamental of the MPS stimulus. The stimulus and reference are presented at a constant level of 50 dB SL. The fundamental of the reference and the stimulus is varied in a sequence: in the first interval the fundamental is presented, in the second interval the fundamental is raised by one third and in the last interval the fundamental is presented again. The adjusted signal-to-noise ratio is a measure for the saliency. On average, for a fundamental of 100 Hz, a difference of about 2 dB is found between the saliency of the undelayed MPS stimulus and the MPS stimulus with its harmonics scattered over the CAP.

CONCLUSIONS

The low pitch percept of the dichotic MPS stimuli in this experiment is faint because maximal 4 harmonics are present in the stimulus. Even experienced MPS listeners had difficulties in matching these stimuli with a pure tone. This causes the experiment to be difficult to perform. The results of the dichotic pitch matching experiments for the delayed MPS stimulus with its harmonics scattered over the CAP suggest that the observers were able to group peaks for different internal delays.

However, from the saliency experiments it turned out that the saliency of the delayed MPS stimulus with its harmonics scattered over the CAP is less (about 2 dB for a 100 Hz fundamental) than the saliency of the undelayed MPS stimulus with its harmonics on a straight line in the centre of the CAP. This suggests that the 'straightness' of the line, connecting the peaks in the CAP, is an additional cue for the binaural interaction system.

Finally, the results of the monaural experiments indicate that it is possible to extract a low pitch out of a single spectrum from the CAP of the delayed MPS stimulus. This low pitch corresponds to the (absent) fundamental of the MPS stimulus. However, this synthetic pitch percept was reported to be faint and most times an analytical pitch was matched. This low pitch could be an explanation for the low pitch found for the dichotic delayed MPS stimulus with its harmonics scattered over the CAP. In spite of this, the faintness of this low pitch makes it questionable that the pitch information in one of these spectra is enough for the binaural system to result in a low pitch percept for the delayed MPS stimulus.

REFERENCES

Raatgever, J. (1980), “On the binaural processing of stimuli with different interaural phase relations”.
Doctoral dissertation, Delft University of Technology, Delft.
ON SOUND DISTANCE PERSPECTIVE WITH BINAURAL HEARING IN ROOMS

Teruji Yamamoto, Touru Ichikawa
Faculty of Science, Toho University, Chiba, Japan.

SUMMARY
This paper describes multivariate analysis of the psychological scale obtained from subjective assessments for the sound distance perspective in rooms and compares the findings with corresponding physical values.

A hearing test by two channel headphone was carried out by the paired comparison method. Test sounds were reproduced through many loudspeakers set in two rooms and received by a dummy head microphone at a point in each room.

Data for assessment were constructed in three dimensional space by Kruskal's multidimensional scaling. The correlations between these and five physical values, including sound pressure level and the distance between the sound source and listening point were then studied.

The varimax method was used to extract the physical values which corresponded with three of the subjective measures as independently as possible. Coordinates of the measures were rotated in a space of three degrees of freedom. The structure vectors of each composite variable were calculated by multivariate analysis of the subjective measures and physical values. Excellent correlations were obtained for the following: (a) subjective measure II, distance between sound source and listening point, and (b) subjective measure III, listening sound pressure level.

INTRODUCTION
Visual sensations are important for judging sound distance perspective in rooms. It is said, however, that listeners can manage fairly well even without the help of vision. Some papers have reported that distance perspective depends on the loudness, reverberation and timbre, but these relationships have not been analyzed quantitatively.

In this study, we investigated how physical values in the sound field are related to subjective measures with regard to sound distance perspective in rooms. This paper describes the procedures and results of the multivariate analysis of the psychological scales obtained from the subjective assessments and also the corresponding physical values.

HEARING TEST
Hearing tests were first conducted in the open air to examine the test procedures and

[Diagram: Fig. 1 Block diagram of the experimental devices]
methods of analysis. Then, regular hearing tests were held in two rooms with respective spatial volumes of 152 and 58 cubic meters. Fig. 1 shows a block diagram of the experimental devices for the subjective assessments.

The original sound used was a male voice saying the words "satoh kun". The sounds were reproduced through many loudspeakers set up in the open air and in the rooms, and recorded through a dummy head microphone at one point in each setting. The numbers of sound sources in the open air and the two rooms were 12, 10 and 4, respectively. In order to imitate sounds from nearby sources, enlarged sounds from distant sources were included. The distances between sound source and the receiving points in the open air ranged from 3.0 to 10.0 meters. In the large room, the range was from 2.0 to 7.0 meters, and in the small room from 1.7 to 6.0 meters. Fig. 2 shows each arrangement of the loudspeakers in the large and small rooms.

The sounds were made in pairs for testing by the paired comparison method. The numbers of pairs in each condition were variously 144, 121 and 64. The hearing test was carried out by two channel headphone on pairs of stimuli compiled at random at 6 second intervals, with two postgraduate students selected as listeners. The assessments were made by the seven point bipolar scale, as shown in Fig.3. The correct judgment rates of the two listeners for 190 pairs of the identical stimuli were 97.3 % and 98.9 %, respectively, which we judged to be adequate for the experiment.

ANALYSIS OF THE RESULTS
Psychological Scaling

Data for each sound tested were constructed into a psychological scale using Kruscal's multidimensional scaling. In this method, "stress" is used to measure how well the calculated configuration matches the experimental data. In the open air, the configuration converged in two dimensions, and the stress was 6.8 %. This corresponded to a good fit. For all positions in the large room and one point in the small one, the values of stress were 4.7 % and 3.9 % in two and three dimensions respectively. The values did not vary much in more than three dimensions, so we judged that the subjective values in the rooms were configured in three dimensional space.

Fig.4 shows a scatter diagram between the values of subjective assessments, that is the dissimilarity, and psychological distance in the open air. This distance was calculated from the configuration of the test sounds scaled in two dimensional space. Fig.5 shows the scatter diagram for the rooms, as calculated from the configuration in three dimensional space. As seen in the figures, the psychological distances agreed fairly well with the dissimilarity.
Multivariate Correlation Analysis
The values corresponding to each configuration axis in the three-dimensional space were named subjective measures I, II and III, respectively, and the correlation between these and five physical values was then studied. The physical values were: listening sound pressure level L (dB), distance between sound source and listening point R (meters), angle of horizontal sound incident A (degrees), ratio of early to reverberant sound energy D (%) at 1 kHz, and reverberation time T (seconds) at 1 kHz. Each value is listed in Table 1.

As the configurations in three-dimensional space gave relative values, the coordinates were rotated into space of three degrees of freedom, and the correlations between the measures after rotation and the physical values were analyzed. The varimax correlation method was used to extract the physical values corresponding to the three subjective measures as independently as possible. That is, the weighting vectors and composite variables were calculated as far as possible so as to produce high correlation with some of the objective

Table 1 Subjective measures in three dimensions and physical values at the sound source positions.

<table>
<thead>
<tr>
<th>subjective measures</th>
<th>physical values</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>SI</td>
</tr>
<tr>
<td>1</td>
<td>2.91</td>
</tr>
<tr>
<td>2</td>
<td>-1.99</td>
</tr>
<tr>
<td>3</td>
<td>-1.94</td>
</tr>
<tr>
<td>4</td>
<td>-1.74</td>
</tr>
<tr>
<td>5</td>
<td>7.71</td>
</tr>
<tr>
<td>6</td>
<td>-3.66</td>
</tr>
<tr>
<td>7</td>
<td>-0.46</td>
</tr>
<tr>
<td>8</td>
<td>6.12</td>
</tr>
<tr>
<td>9</td>
<td>-6.16</td>
</tr>
<tr>
<td>10</td>
<td>-5.62</td>
</tr>
<tr>
<td>11</td>
<td>-7.12</td>
</tr>
</tbody>
</table>

Fig. 6 Structure vectors of each composite variable in the case of two rooms.
variables and low correlation with the remainder. Successive similar calculations were made for each composite variable orthogonal to the others.

Fig. 6 gives the structure vectors of each composite variable obtained, and Fig. 7 shows the configuration of subjective values at the stage in which the results shown in Fig. 6 were obtained. For reference, composite variables obtained from the open air results are shown in Fig. 8. This figure corresponds to the results shown in Fig. 5.

According to the results, each subjective measure was separately extracted in three composite variables. Subjective measure II has a high correlation with distance between sound source and listening point R, and some inverse correlation with the ratio of early to reverberant sound energy D. Furthermore, independently of this measure, subjective measure III has a high inverse correlation with listening sound pressure level L. Simple correlations between the subjective and physical values are shown in Figs. 9 and 10.

CONCLUSION

The space of psychological assessment for sound distance perspective in rooms is almost three dimensional. Each subjective measure and physical value forms a separate group. As shown in results both in the open air and in two rooms, listening sound pressure level and distance from the sound source have high correlations with the psychological assessments, independently of each other. It appears that these two physical values may be the main factors involved in sound distance perspective in rooms.

Fig. 9 Correlation between subjective measure II and distance from sound source R.

Fig. 10 Correlation between subjective measure III and listening sound pressure level L.
BIOACOUSTICS AND MEDICAL ACOUSTICS
THE RELATIVE ROLES OF THERMAL AND NONTHERMAL EFFECTS IN THE USE OF HIGH INTENSITY FOCUSED ULTRASOUND FOR THE TREATMENT OF BENIGN PROSTATIC HYPERPLASIA

Lawrence A. Crum, Wing Law*
Applied Physics Laboratory, University of Washington, Seattle, USA
* Focus Surgery, 225 Hammond Avenue, Fremont, CA 94539 USA

SUMMARY

Because of its seemingly benign and noninvasive nature, ultrasound has become a major tool in diagnostic medicine. Its use for fetal imaging has expanded to the extent that the overwhelming majority of pregnancies in the industrialized world are monitored with this modality. However, the use of high intensity ultrasound for therapy, although sought after for many years [1], has only recently been seriously considered for application. The development of a commercial system that utilizes high intensity focused ultrasound (HIFU) for the treatment of benign prostatic hyperplasia (BPH) has now resurrected this promising modality and offers the researcher opportunities to examine many interesting aspects of bioacoustics. This paper examines the relative roles of thermal effects and acoustic cavitation in the production of lesions by a specific HIFU device. It is found that under certain conditions, gas bubble generation may be induced by thermal effects, with the consequence that cavitation then becomes an important aspect of the continuing application of the therapy.

BACKGROUND

The interaction of ultrasound with biological tissues is usually categorized into two major groups: thermal and nonthermal mechanisms. Because temperature is normally a continuous variable over a region of interest, standard thermal diffusion models have been reasonably successful in predicting spatial and temporal values of this parameter for a given acoustic intensity [2].

The principal nonthermal mechanism with which an acoustic field interacts with tissue is thought to be acoustic cavitation. When a sound field of sufficient intensity propagates through a liquid medium, the acoustic rarefactions can place the liquid under considerable tensile stress. If this stress level exceeds the tensile strength of the medium, a cavity is formed, which fills with vapor and continues to grow so long as the (negative) pressure is less than the vapor pressure. Subsequently, however, the compressional portion of the acoustic field forces the bubble into an implosive collapse. Because the cavities formed are normally quite small (on the order of tens of microns), surface tension tends to maintain the cavity (or bubble) in a spherical shape. Thus, when these cavities collapse they tend to do so in some degree of spherical symmetry, which, together with the nature of an implosion, can result in energy density concentrations of twelve orders of magnitude [3]. Thus, acoustic cavitation can be an extremely violent process and can result in remarkably large values of the local temperature (kiloKelsins) and pressure (kilobars). In our ongoing research, we seek to ascertain the relative contributions of these thermal and nonthermal mechanisms in a particular application of HIFU: The treatment of Benign Prostatic Hyperplasia (BPH).
EXPERIMENTAL APPROACH

In the particular protocol used by the Sonablate device developed by Focus Surgery [4,5], a small transducer is used to image the prostate via the rectal canal. Once a particular region to be treated is identified, a 4-sec burst of HIFU at a local intensity of about 2000 W/cm² is applied to one sector of this region, a 12-sec rest period is then maintained, and another sector is automatically selected by the internal computer. This procedure is repeated until a region within the prostate that surrounds the urethra is treated. After each individual 4-sec treatment, the region is scanned and video updates presented of the insonified region. Often, but not always, discrete hyperechogenic regions appear in the scanned area. These echoes appear as white areas on the gray-scale image and are occasionally accompanied by a popping noise. Thus, these particular events have been named "popcorn" and are thought to be an important feature of the HIFU treatment. In many cases, the appearance of popcorn promises a successful treatment. The origin of popcorn is unknown and the principal subject of a recent investigation. We report here some preliminary results of our investigation.

We desired to perform some in vitro experiments that would elucidate the nature of popcorn and to determine if it had a thermal or nonthermal origin. Because of the limited space available in this report, we will present no block diagram of our experimental apparatus but will briefly describe our relatively simple system that we used to examine this phenomenon.

A section of turkey breast was positioned within a holder and immersed in degassed, distilled water maintained at 37 °C. The Sonablate transrectal probe that contained the imaging and therapy transducer was also immersed in the water very near the tissue. The imaging mode of the system was engaged and used to locate a region in the tissue to be treated. Coaxial with the focus of the transducer beam, but sufficiently downstream from the focus, was placed a broadband transducer that was used as a hydrophone. The signal from the hydrophone was amplified, filtered and input to both a chart recorder and a spectrum analyzer. A second channel of the chart recorder was used to display the electrical power supplied to the source transducer.

RESULTS AND DISCUSSION

Figure 1 shows a plot of the radiated (source) and transmitted (received) acoustic intensity as a function of time.

![Fig. 1. Plots of the radiated and transmitted acoustic intensity through turkey breast tissue as a function of time. At the position of the break, "popcorn" occurred. Note the rapid drop in transmitted intensity and the (reflected) noise in the radiated power. The slight offset in the occurrence of these events is due to the offset in the chart recorder pens. The vertical lines correspond to 2.5 s.](image-url)
It is seen in Fig. 1 that the appearance of popcorn resulted in a significant drop in the transmitted intensity; also, the popcorn affected the radiation impedance of the source transducer sufficiently so that oscillations also occurred in the radiated power. Furthermore, it is seen that after the appearance of popcorn, the transmitted intensity fluctuated wildly.

The most logical explanation for this behavior is that popcorn is the manifestation of noncondensible gas that has been generated near the focus of the source transducer. This gas must be noncondensible because it survives for periods of several seconds, even minutes. It is most likely the origin of the hyperechogenic region that is seen in the scanned region. Once this gas is introduced, it provides the site for intense cavitation-like behavior as the high intensity sound field continues to irradiate this region.

It is also of interest to speculate as to the origin of the gas--where does it come from and how does it appear so rapidly? Consider conditions applicable to those used for clinical operation: The volume of the focal region is quite small—the 3 dB points define a radius of less than 1.0 mm and a length of less than 1.0 cm—and the acoustic intensity is quite high—about 2000 W/cm². Under these conditions, one can calculate that the temperature rise of tissue in the focal region should be on the order of 500 K/s. Although those conditions were not duplicated in this in vitro experiment, it is likely that temperature rises can be so rapid that the local region of the tissue within the focus can be quickly driven to superheat conditions. If a region of the tissue becomes superheated, and then suddenly is nucleated, it is likely that an audible pop could be heard as the liquid in this region rapidly boils (literally explodes). Furthermore, since this area of tissue contains dissolved gas—in concentration equilibrium at a temperature of approximately 37 °C—it is reasonable to expect that noncondensible gas would diffuse into the many individual vapor cavities that would be generated.

We considered a test of this hypothesis: Consider Fig. 2, which shows a spectrum of the transmitted signal immediately after the appearance of popcorn.

![Fig. 2. Spectrum of the acoustic noise associated with the appearance of popcorn in an insonified region of a turkey breast. The main peak is at 4.0 MHz, the source frequency. Note the appearance of a significant subharmonic signal, and also evidence of a suprathermic at 6.0 MHz. In this figure, the zero offset is 0.1 MHz and the divisions are each 1.0 MHz. The ordinate is in dB with an arbitrary reference. Without popcorn, the bands at 2.0 MHz and 6.0 MHz are absent.](image-url)
Figure 2 provides considerable evidence in support of our hypothesis that popcorn is noncondensable gas remaining from an explosion of a superheated region of tissue. Note that there is a strong subharmonic, a weaker supraharmonic, and significant white noise. The broadband white noise located between 2-4 MHz is indicative of transient cavitation events that implode violently, creating impulsive sources that are represented as a broadband spectrum. The relatively large peak at 2 MHz, however, probably represents gas bubbles that are driven so nonlinearly that they radiate a significant subharmonic component. It appears as if the nonlinear response to be so large as to even induce a harmonic of the subharmonic--that is, a peak at 6.0 MHz.

CONCLUSIONS

We have examined various aspects of HIFU operation under a specific application--the treatment of BPH--for both in vivo and in vitro conditions. Some of our observations and deductions are as follows:

- In the specific application of HIFU to the treatment of BPH, a common observation is the production of "popcorn", a hyperechogenic region near the focus that is often accompanied by an audible "pop".

- In vitro examination of popcorn suggests that the hyperechogenicity is most likely produced by noncondensible gas.

- It is likely that the origin of popcorn is the superheated boiling of a localized region of tissue that is rapidly heated by the HIFU.

- Popcorn is acoustically very active, radiating numerous subharmonic and supraharmonic components, and is most likely acoustically opaque.

LIST OF REFERENCES


BIOLOGICAL EFFECTS AND MECHANISM OF ACTION OF EXTRACORPOREAL SHOCK WAVES

Michael Delius
Institut für Chirurgische Forschung, Ludwig-Maximilians-Universität, München, Germany

INTRODUCTION

Extracorporeal shock waves are pressure pulses of microsecond duration with peak pressures of 35-120 MPa followed by a tensile wave (Coleman and Saunders 1989). They are an established treatment modality for kidney and gallstone disease (Chaussy et al. 1980, Sackmann et al. 1988). More than 80% of stones in the kidney and ureter are nowadays treated by extracorporeal shock waves. Gallstones are still commonly treated by surgery, yet stone fragmentation by shock waves represents a treatment alternative for 20% of patients. Further applications of shock waves are pancreatic (Delhaye et al. 1992) and salivary (Iro et al. 1992) stones, as well as delayed fracture healing (Valchanou and Michailov 1991). These are either on their way to become established treatments or currently under investigation.

Extracorporeal shock waves are generated in water outside the body either electrohydraulically by underwater spark discharge between the tips of an electrode, or electromagnetically by sudden repulsion of a metal membrane overlying a flat coil, or piezoelectrically by 30-3000 piezoceramic crystals mounted on the inner surface of a spherical dish. Further details of shock wave generators, their clinical application and bioeffects have been reviewed recently (Coleman and Saunders 1993, Delius 1994).

TISSUE DAMAGE IN VIVO

Extracorporeal shock waves are coupled through the skin, propagated in tissue and focused on the kidney- or gallstone by means of an ultrasonic or X-ray localising device. On their way through the body, they can generate tissue damage as a side effect. It consists of focal lesions which are especially evident at blood vessels. Focal destruction of the vascular wall induces bleedings and the formation of blood clots. Veins are most often affected, yet arteries are not spared. Bleedings occur in a high pressure area along the central axis of the shock wave field. They have been observed in all organs which have so far been exposed to shock waves and can be detected in the kidney in a high percentage of patients after kidney stone fragmentation when sensitive diagnostic imaging methods are employed (Baumgartner 1987). Parenchymal cells of an organ, e.g. liver cells in the liver and cells of filtration units in the kidney, are also affected, be it directly by the shock wave action or indirectly as a consequence of a decreased blood supply by clot formation.

CELL DAMAGE IN VITRO

Shock wave application to suspended cells in vitro induces cell lysis and renders part of the cells non viable. Depending on the shock wave dose, the proliferation of the acutely surviving
cells is only slightly or moderately affected. If the cells are exposed as spheroids, i.e. cell balls of 0.5 - 1 mm diameter, the shock wave action is greatly diminished (Bräuner et al. 1989). A similar effect is achieved by immobilization in gelatin (Brümer et al. 1989). The cell membrane is most sensitive to shock waves, the cytoskeleton and nucleus are not as readily damaged (Steinbach et al. 1992).

INCREASED MEMBRANE PERMEABILITY
Shock wave exposure of cells in suspension causes a transient increase in membrane permeability without leading to cell death (Gambhiler et al. 1992). Experiments with large molecules revealed that even those with a mass of 2 million Dalton could enter the cell when present in the medium during shock wave application (Gambhiler et al. 1994). This pointed to the generation of short-lived pores of considerable size in the cell membrane.

TRANSFER OF RIBOSOME INACTIVATING PROTEINS
Via porc formation, shock waves can mediate the transfer of large molecules into the cell which cannot enter naturally. This has been demonstrated for ribosome inactivating proteins (Delius unpublished). They have a molecular weight of 30 kDa and are not toxic because they cannot enter the cell. Once inside the cell, however, a single molecule is sufficient to induce cell death (Eiklid et al. 1980). When cells in suspension were exposed to shock waves in the presence of the ribosome inactivating protein gelonin, its action could be augmented up to 60.000 fold. The principle should also be applicable to other substances. A possible utilization is tumor therapy where exposure of the tumour to shock waves should allow a tumour-selective entry of non-permeating drugs into cancer cells.

GENE TRANSFER
Nucleic acids can also enter cells via pore formation from shock waves. Shock wave exposure of cells to plasmids, which are nucleic acid molecules with a molecular weight of 2-10 MDa, containing either the gene of the enzyme β-galaktosidase or a gene for part of the hepatitis B virus surface protein, led to the synthesis of these proteins by some of the cells (Lauer et al. 1994). The production of β-galaktosidase was demonstrated by staining the cells containing the enzyme blue, and hepatitis protein secretion by determining the protein in the supernatant by immunoassay. Further investigations have to be performed to examine whether extracorporeal shock waves are suited as a new gene transfer system. A major problem of the method is that the efficiency of gene transfer has to be increased for its practical application.

GAS BUBBLE GENERATION AND SHOCK WAVE - GAS BUBBLE INTERACTION
Lithotripters generate cavitation, i.e. the movement of newly formed and preexisting bubbles containing gas or vapor in a fluid (Apfel 1981, Crum 1982). It is well known that cavitation is a powerful mechanism of material damage; it causes surface craters and erosion (Steinberg 1993). Cavitation from a lithotripter is easily demonstrated by the generation of craters at aluminum foils or plaster stones exposed in its waterbath (Coleman et al. 1987).

Two types of cavitation have been observed in lithotripters, the tensile wave generating cavities de novo from invisible cavitation nuclei, and the positive pressure pulse interacting with preexisting gas bubbles. The former are generated along the central axis of the shock wave field. In living tissue, these cavities have been recently visualized indirectly by implantation of a fibre-optic hydrophone (Huber et al. 1994). Their lifetime was shorter than in free fluid. During cavity oscillations, gas is entrapped by rectified diffusion, leading to the generation of long-lived gas bubbles several tens of microns large (Church 1989). In the waterbath, gas bub-
ble sizes up to 40 \( \mu \text{m} \) have been directly observed (Delius 1994). In vivo, long-lived gas bubbles have been indirectly visualized by diagnostic ultrasound (Kuwahara et al. 1989, Delius et al. 1990). When they are generated in blood vessels they are flushed away with the blood flow.

The interaction of a pressure pulse with a gas bubble is a very powerful event (Tomita and Shima 1986, Dear and Field 1988). Shock wave - gas bubble interaction has been directly photographed in a lithotripter after positioning of air bubbles below a plastic foil. Maximal velocities of the bubble wall of 400-800 ms\(^{-1}\) were achieved which are comparable to a rifle bullet (Philipp et al. 1993).

As opposed to direct effects of the high shock wave pressure, cavitation mediated events are expected to be suppressed by overpressure. Exposure of cells in a pressure chamber at 10 MPa static overpressure abolished cell lysis completely (Gambhirer et al. 1990). Not only cell lysis but also the increase in membrane permeability, uptake of ribosome inactivating proteins and gene transfer were abolished by overpressure. Recently, specific evidence was obtained for the importance of shock wave - gas bubble interaction for the generation of cell damage (Delius unpublished). The shock wave effect turned out to be extremely sensitive to low static overpressure. Exposure of red blood cells - which are a good marker of cell lysis because the freed haemoglobin is easily quantified - at an overpressure as little as one atmosphere (!) abolished lysis and liberation of haemoglobin nearly completely. As little as 30 kPa overpressure reduced the free haemoglobin already by half. A similar result was obtained when membrane permeabilisation was assessed. The applied static overpressure in these experiments was negligibly small in relation to the magnitude of the tensile wave of the lithotripter which is in the range of 10-15 MPa (Staudenraus and Eisenmenger 1993). The result suggests that remnant gas bubbles have been brought back into solution by static overpressure, thus decreasing the shock wave - gas bubble interaction.

Evidence for an important role of shock wave - gas bubble interaction in the generation of tissue damage in vivo was obtained by injecting microbubbles into an artery during shock wave application (Prat et al. 1991) and by administering shock waves experimentally at faster sequences, thus facilitating their interaction with gas bubbles (Delius et al. 1988). Both protocols increased tissue damage.

**MECHANISM OF STONE FRAGMENTATION**

The mechanism of stone fragmentation is currently a matter of considerable debate. It was thought for long that the high shock wave pressure acted directly at kidney and gallstones, causing cracks and material removal similar to its known action at other materials (Kolsky 1963). Gallstone fragmentation was, however, suppressed by high static overpressure and by a fluid of high viscosity around the stone (Delius et al. 1988). High speed photographs demonstrated further that fragment removal from gallstones occurs in connection with cavitation bubble collapse (Sass et al. 1991). And it was recently found that fragmentation is also sensitive to low overpressure (Delius unpublished). At 100 kPa overpressure, gallstone fragmentation was already reduced, and at at 1MPa it was nearly abolished. Together these points support the involvement of cavitation in stone fragmentation. What happens is detail is, however, still unknown.
REFERENCES
NONINVASIVE FEEDBACK FOR THERAPEUTIC APPLICATIONS USING HIGHLY-FOCUSED ULTRASONIC FIELDS

Emad S. Ebbini, Ralf Seip
Department of Electrical Engineering and Computer Science,
The University of Michigan, Ann Arbor, MI 48109

SUMMARY

We have developed a signal processing algorithm of diagnostic ultrasound pulse-echo data that can be used for noninvasive measurement of temperature \textit{in-vivo}. Experimental data in tissue mimicking phantoms, \textit{in-vitro}, and \textit{in-vivo} tissue indicate that 0.4 °C resolution and 3 mm spatial resolution is achievable with the new technique. In this paper, we describe the new temperature estimation algorithm and show representative experimental data, both \textit{in-vitro} and \textit{in-vivo}. We outline an algorithm for using the noninvasive temperature measurements in the guidance of HIFU beams with supporting experimental data.

INTRODUCTION

High-intensity focused ultrasound (HIFU), in the frequency range of 500 kHz to 10 MHz, has been used in a wide range of therapeutic applications. For example, HIFU has considerable potential for deep localized hyperthermia, primarily in conjunction with other cancer treatment modalities, e.g., radiation therapy. This is due to the fact that ultrasound can be easily focused (with mm-size heating spots at the focus) at considerable depths (1 cm - 15 cm with focusing) depending on the frequency. The (typically) small heating spot at the focus can be scanned mechanically or electronically to produce the desired heating pattern tailored to the tumor geometry. Clinical hyperthermia systems utilizing HIFU have been used by a number of groups and they continue to demonstrate the promise of this technology. HIFU has become even more attractive with the recent advances in phased-array applicator technology. Phased arrays offer the promise of versatile applicator systems providing several advanced field control features, e.g., high-speed electronic scanning, tissue aberration correction, and applicator/patient motion compensation. Furthermore, optimal pattern synthesis methods have been developed for phased arrays allowing adequate therapeutic heating of tumors while avoiding over heating of normal intervening tissue.

Another application where HIFU was demonstrated to have significant promise is in noninvasive tissue ablation for cancer treatment and in surgery. As early as the 1950s, Fry and coworkers [1] at the University of Illinois have shown that the formation of very precise ablation patterns in cat brains \textit{in-vivo} is feasible using HIFU. A system was developed and numerous experiments were performed further supporting the feasibility of this approach. Similar results were obtained by Lele at MIT [8] and most recently by Hyynen at the University of Arizona and Harvard [9] and Sanghvi et al at ICFAR [10]. It has been demonstrated by these investigators and by our group that thermal coagulation of tissue results from short (1-10 seconds) exposure to intense (1-3 kW/cm²) focused beams. Furthermore, at these exposure levels and time durations, the ablated tissue volumes are largely defined by the specific absorption rate (SAR) of the focused ultrasonic field, i.e., the focal spot. For most practical ultrasound focusing systems, the ablated tissue is cigar-shaped at the focal spot with a diameter on the order of 1 mm and length on the order of 3 to 4 mm.

FIELD PATTERN CONTROL

Recently, phased array applicators were developed for hyperthermia and other HIFU applications. Phased arrays offer significantly improved control features that will be needed for precision lesion formation at depth in the presence of tissue inhomogeneity and patient/applicator movement. This is due to the electronic
focusing capability of these applicators which can be performed dynamically at electronic speed to track the target (e.g., tumor) in real-time for the duration of the treatment. Effects of tissue inhomogeneities can be compensated for simply by the proper choice of the phases of the array elements. A number of optimal phasing schemes were developed [3, 4, 5] to refocus any array in the presence of tissue inhomogeneity based on minimally-invasive acoustic feedback using miniature probes [6]. In addition to finding the proper phase and amplitude distribution for refocusing the array, the algorithms described in [6] identify and deactivate any elements shadowed from the target by the presence of an acoustic obstacle (e.g., bone or air spaces). Therefore, the availability of this kind of feedback will allow for the use of arbitrary array geometries with multiple acoustical windows to maximize the power deposition to the target region.

Despite the obvious advantages of HIFU for minimally invasive therapeutic application and the availability of advanced phased-array applicator systems capable of precision lesion formation at depth, HIFU is not yet a widely accepted modality in the clinic either for normal hyperthermia or for surgery. The reasons behind this lack of progress in clinical utilization of HIFU are as follows:

1. Lack of clinical real-time non-invasive high-resolution temperature feedback.

2. Lack of quantitative non-invasive measurement of tissue response to HIFU fields (especially important for tissue ablation).

In addition to the above mentioned limitations, it would be extremely valuable (indeed, essential) to develop a guidance and visualization mechanism for the therapeutic beams based on a well established imaging modality. This fact has been recognized by a number of investigators with significant clinical support [9, 10]. At this point, MRI [9] and B-scan ultrasound [10] are being used. In the following, we describe a noninvasive temperature estimation algorithm based on signal processing of acoustic pulse-echo measurements.

**NON-INVASIVE ACOUSTIC FEEDBACK**

Pulse-echo ultrasound can be used to non-invasively estimate temperature changes in tissue [11]. Specifically, it is possible to locally measure the value of the speed of sound in tissue and obtain a measure of the average scatterer spacing of the tissue. Both of these parameters are sensitive to changes in the temperature of the tissue. Thus, by measuring them non-invasively, an estimate of the tissue temperature change can be obtained. These measurements are based on the estimation of a fundamental frequency \( f_1 \) and its harmonics \( f_k \), which are defined by:

\[
f_k(T) = \frac{k \varepsilon(T)}{2d(T)} \quad k = 1, 2, \ldots, \infty
\]

where \( \varepsilon \) is the local value of the speed of sound in the tissue, \( d \) is the average scatterer spacing, and \( T \) is the temperature of the tissue. Equation 1 is derived from the two-way travel time of an ultrasound pulse between two scatterers that are separated by a distance \( d \).

The average scatterer spacing as a function of temperature can be approximated by:

\[
d = d_o \left( 1 + \alpha \Delta T \right)
\]

where \( \alpha \) is the linear coefficient of thermal expansion of the tissue. If Equation 2 is used in 1, and Equation 1 is differentiated with respect to \( T \), we obtain the following:

\[
\Delta f_k(T) \approx \frac{k}{2d_o} \left[ \frac{\partial \varepsilon(T)}{\partial T} \right]_{T=T_o} - n \frac{\varepsilon_o}{2d_o} \Delta T
\]

where \( \varepsilon_o \) and \( d_o \) are the values for the speed of sound and the average scatterer spacing in the tissue at the baseline temperature \( T_o \), respectively. Equation 3 thus shows that a change in tissue temperature will manifest itself as a change in \( f_1 \) and its harmonics. Hence, if it is possible to reliably track \( f_1 \) and its harmonics, it becomes possible to track \( \Delta T \). To date, \( f_1 \) and its harmonics have been extracted from the power spectral density (PSD) computed for an acquired A-line using an autoregressive (AR) model. Our experience with the AR PSD, as well as other researcher's experience, has been quite positive [12].

To experimentally verify Equation 3, A-lines were acquired with a focused imaging transducer while a rubber phantom was uniformly in a water bath. This experiment was performed to prove that our observed frequency changes are due to temperature change and are not related the presence of HIFU. Figure 1 shows three A-lines collected during this experiment. The sample temperature at which each of the A-lines was acquired is shown. The effects of the increase in speed of sound in water are exhibited in a pronounced shift in the time (apparent axial distance from transducer assuming \( c_o=1500 \text{ m/s} \)) at which the front end of the sample appears. After aligning the A-line data, the AR PSD is computed for a small data window...
at the same location in each A-line. A sample AR PSD is shown in Figure 2. Some of the peaks in the AR PSD correspond to harmonics of $f_1$. If these harmonics are extracted and scaled as determined by Equation 3, the linear relation between harmonics and measured temperature becomes apparent. Figure 3 clearly demonstrates this phenomenon. Non-invasive temperature change estimation using ultrasound is thus possible. Unlike previous techniques which relied on the estimation of the speed of sound by time-of-flight techniques, this approach avoids the complications associated with tissue and heating field heterogeneities along the path of the ultrasonic pulse. The result derived in Equation 3 has been experimentally verified in phantoms, as well as in an in-vitro and in-vivo setting.

Equation 3 can also be used to non-invasively estimate the spatial distribution of a heating field, as shown in Figure 4. The data window used to compute the AR PSD was moved across the acquired A-lines, which were obtained during a focused high-intensity ultrasound heating and cooling experiment. As expected, the largest temperature change occurs in the focal zone of the therapy transducer, and manifests itself as the largest frequency shift. The spatial resolution of these measurements is better than 3 mm, and the standard deviation of the temperature estimates is on the order of 0.4 °C. This result indicates that it is possible to "visualize" the effect of the therapeutic beam noninvasively from the temperature rise data before ablation occurs. This should allow for the guidance of HIFU beams noninvasively using this form of feedback.

Figure 1: Collection of A-lines acquired as a function of time during a uniform water bath heating experiment.

Figure 2: AR PSD computed for a 3 mm window from 250 A-lines similar to those shown in Figure 1, acquired approximately once every second. The harmonics of $f_1$ are visible as bright bands in the spectrum.

Figure 3: The 5th harmonic of $f_1$ extracted from the spectrum of Figure 2, compared to the measured temperature.

Figure 4: Non-invasively estimated spatial frequency shifts of the 5th harmonic of the AR PSD (°C temperature) for a localized ultrasound heating (and cooling) experiment.

CONCLUSIONS
A new algorithm for noninvasive estimation of tissue temperature change due to external heating fields was presented. The algorithm is based on observed shifts in harmonic frequencies related to the mean scatterer spacing in the heated volume. These frequency shifts were shown to be related to local change in speed of sound and to thermal expansion, both are function to temperature. AR spectral estimation techniques were used and were shown to provide a robust estimate of these frequency shift. Experimental data indicate that a spatial resolution of 3 mm and temperature resolution of 0.4 °C are currently possible with our approach. We have also shown, experimentally and based on theoretical grounds, that our technique is sensitive to temperature variations on the order of 1-2 °C in tissue. Furthermore, we have shown that the approach is appropriate for the localization of the heating field. Therefore, this approach is appropriate for the guidance of HIFU therapeutic beams. Thus it would be possible to envision a HIFU therapeutic device guided by an ultrasound imaging device in a noninvasive surgery application. The imaging device, of course, will have the capability to process RF data on line to obtain spatio-temporal estimates of temperature changes due to the HIFU applicator. Research on the integration of the diagnostic and the therapy device with the guidance capability is currently underway.

Acknowledgements
This work was supported in part by a Grant from the National Science Foundation Young Investigator Award ECS 9356301, and in part by DMRG 2 567 X0070050-25 the Biomedical Research Support Grant Program, National Institutes of Health.

References
ULTRASOUND TRANSDUCERS FOR MEDICAL USE

Helge E. Engan
The Norwegian Institute of Technology, Dep. physical electronics, Trondheim

SUMMARY

The key component in any medical ultrasound system is the transducer. This paper surveys the main development trends and the status within selected areas of ultrasound medical transducers. The main focus is on results achieved within the areas of high power transducers, diagnostic transducers, and transducer arrays.

HIGH POWER TRANSDUCERS

This group of transducers is used for purposes such as lithotripsy and depth heating of tissue for hyperthermia cancer therapy.

Transducers with the highest intensity are used as extracorporeal lithotripters. They typically generate shock waves electrohydraulically which are focused into a small region in the body to create amplitudes too large for the items in question to withstand. These systems are used mainly for pulverizing stones. In a standard system of this type a powerful spark is discharged in a small volume. The generated shock wave is spread radially and refocused by curved (ellipsoidal) reflecting surfaces into the body where the exposed region is located. This type of lithotripter is reported to be able to create shock waves with focal intensities of about 1-10^5 W/cm^2 with a duration of the order of ½ μs when a 20-kV spark is discharged. By replacing ordinary tap water with an electrolyte of 7 Ωcm resistivity, reproducibility was substantially improved. In addition, oscillatory behavior of the electrical feeding circuitry was suppressed. Shock wave maxima corresponding to about 4-10^5 W/cm^2 with a duration of about ½ μs was found.

A scheme where a spherically shaped dome of 10 cm diameter was covered with 25 piezoelectric transducers distributed in a two dimensional array has been published recently. The transducers were made of a composite material and oscillated at 400 kHz. This arrangement created an elliptical zone of focusing with a major axis of 15 mm and a diameter of 5 mm. Shock wave maxima of 5-10^4 W/cm^2 with a duration of about ½ μs were observed. By electronically adjusting the individual delay of each element the focal point could be steered within an ellipsoidal volume of 120 mm in length and 30 mm in diameter. Thus a transducer arrangement was demonstrated that can be incorporated in a system tracking the movements of the stone. This will make such systems considerably more efficient and less prone to overexposure of neighboring tissue.
For extracorporeal hyperthermia therapy which includes more moderate intensities applied over an extended time, phased array systems are often used. Two-dimensional arrays are used to compensate for movements and phase aberrations as well as for producing well specified temperature profiles. In intracavitary ultrasound hyperthermia cancerous tissues are treated which are placed close to body cavities. In such cases one-dimensional arrays are used for focusing. Multiple-beam operation at 500 kHz of 25-element linear arrays suitable for creating specified temperature profiles has been reported. It was demonstrated that the array could create foci at distances of about 20 to 60 mm and over 60 mm along the array.

For deeper lying regions interstitial methods are often used in which the acoustic heating sources are inserted directly into the tumor. The basic scheme of these applicators is a multielement array of tubular piezoceramic radiators transmitting energy radially. Thus the maximum heat dissipation is near the excitation interface. By cooling the applicator surface with water, the maximum temperature is forced away from this region.

DIAGNOSTIC TRANSDUCERS

The active parts of conventional diagnostic transducers consist of piezoelectric ceramics. Such materials have a high piezoelectric coupling coefficient necessary for wideband applications. However they suffer from a high acoustic impedance thus reducing the energy coupling between the transducer and the propagating medium.

The trend over some years has been to replace ceramics with piezoelectric polymers such as PVDF (polyvinylidene difluoride) which have lower acoustic impedance. However they also have a lower piezoelectric coupling and in general a considerably higher electrical impedance. They are therefore not always good substitutes for ceramics.

Recently a strong interest has been taken in piezoelectric composites. These materials are made by combining a piezoelectric ceramic with a passive polymer in various ways so that its properties can be tailored. Thus a low acoustic impedance can be combined with a high piezoelectric coupling and a low electrical impedance. These materials can also be shaped in curved surfaces, and they can be made in all shapes from single transducers to two-dimensional arrays with variable periodicity and variable weighting (shading) for optimum beam pattern. This seems to be a very fruitful approach, although present limitations seem to be cost and high frequency performance.

Several workers have extended the usual single-layer piezoelectric scheme into transducers with multiple active layers. There are two different reasons for this. One problem has been in arrays where each transducer element is so small that the electric impedance is too high for efficient transduction and sensitivity. This problem has been overcome by making the element of several identical layers. These layers are coupled acoustically in series and electrically in parallel, thereby providing a convenient electric impedance while not changing the acoustic properties.

There is a need to increase the transducer bandwidth. This need is caused by spatial resolution requirements or from multisite such as in color flow imaging where typically simultaneous echo amplitude imaging requires a different frequency. To obtain this, several layers with different thickness can be used. By feeding the terminals with electrical signals of varying amplitude and phase, a new type of transducer with more degrees of freedom is obtained. In this way a transmitting transducer structure with substantially increased bandwidth has been constructed. The same technique can be used for reception. A two-layer structure has also been used to decouple the transmit and receive modes. A two-frequency
scheme can be realized by varying the thickness in two steps across a single piezoelectric plate so that two distinct resonance frequencies are defined. The highest frequency transducers are mostly made of PVDF or similar materials. In superficial tissue characterization transmitting frequencies up to 100 MHz and bandwidths of 35 MHz have been reported.

ARRAYS

Arrays are used extensively in medical imaging for electronic beam steering. The annular array concept uses concentric active rings in a Fresnel zone plate arrangement for creating a time variable focus. Two-dimensional imaging has been demonstrated using linear arrays of up to more than 100 transducer elements. By positioning the array elements along a concave circular segment, the transducer beam can be steered and pivoted through a "keyhole" acoustic window near the skin line between the ribs. Two-dimensional arrays are feasible for reduction of B-scan slice thickness by dynamic focusing, correction of phase aberrations by tissue inhomogeneities, as well as phased array steering in three-dimensional imaging. They seem to be limited presently to 16x16 element arrays due to technological problems caused by interconnects rather than the transducer elements. To overcome these problems, it has been proposed that scanning in one transverse dimension can be done by frequency scanning of tapered transducer elements in which the resonance frequency depends on the position within the elements.

A review of specific intraluminal imaging work has been given by Bom et al.

FUTURE DEVELOPMENT

The properties of transducers are very strongly influenced by the materials they are made of. A strong research effort has led to different materials groups such as piezoceramics, piezopolymers, and piezocomposites, with many different materials within each group. Innovations within the materials field will strongly influence the development of new transducers. There is an effort to make high-frequency piezoceramics by reducing the grain size. Newer electrostrictive materials respond to the square of the applied electric field and they have an extremely high permittivity. This makes such materials interesting candidates for arrays where high impedance is a problem, and where the non-linearity can be used for steering of the array elements both in amplitude and phase. Also the forming technology will influence the development of transducers. By a combined photoresist-jet-machining technique very intricate patterns of high resolution can be produced which can be used in arrays or in the manufacturing of piezocomposites.

In addition to the influence from material development as discussed above, the future transducer development depends heavily on a deeper insight in transducer modelling and in the propagation and scattering of the acoustic energy in the tissue.

REFERENCES


MEASUREMENTS OF GAS BUBBLE SIZES USING HIGH FREQUENCY ULTRASOUND IMAGING

Andreas Heimdal, Torgrim Lie, Alf O. Brubakk
Department of Biomedical Engineering, University of Trondheim, Norway

SUMMARY
This paper describes a transform to obtain physical sizes of gas bubbles from the intensities of the bubbles in a 20 MHz ultrasound image. The transform is based on a measurement of the intensity field from the ultrasound transducer, and an estimation of the position of the bubble normal to the scan plane. A decompression experiment was performed to generate nitrogen bubbles in agar gels. The growth of the bubbles was measured visually using a microscope, and by continuous ultrasound imaging of the gel. The visual measurements and the transformed ultrasound measurements compared well.

INTRODUCTION
In recent years there has been a growing emphasis that gas bubble growth and elimination are of central importance to the study of decompression illness and treatment. [Hyldegård and Madsen] have studied the elimination of gas bubbles in both fatty tissue and the spiral cord. These studies have been performed observing the bubbles using a microscope. This method has the disadvantage that bubbles can only be observed on the surface and only one bubble can be observed at a time.

The purpose of this study was to develop and evaluate a method for studying the growth of gas bubbles in agar gels using ultrasound imaging, as a preliminary step to tissue bubble imaging.

THEORY
For spherical air bubbles with radii much larger than the wavelength of an ultrasound wave in water, the scattering cross section will approach $2\pi a^2$, and no resonance modes will appear [Nishi]. Similar to the description by [Kino] and [Medwin], for planar constant amplitude waves, the isotropically scattered intensity $I_s$ at a distance $r$ from a large bubble has the following relation to the incident intensity $I_i$, the bubble scattering cross section, and the steradian area $4\pi r^2$.

$$I_s = \frac{2\pi a^2}{4\pi r^2} I_i$$  (1)

In practical imaging the transducer emits a focused beam rather than a planar constant amplitude wave. An illustration of the beam from a rotating mirror ultrasound catheter and the coordinate system used in the following discussion is shown in Figure 1. The beam is assumed to be circular symmetric around the beam axis.

![Figure 1. The radiated field from the transducer and a rotating focused mirror. A bubble with diameter D is located at (r, w).](image_url)

If the variation of the incident intensity field over the bubble cross section is small, and the bubble is located far away from the focus of the beam, one can assume a planar wave with constant intensity equal the mean beam intensity over the bubble cross section. Then, according to equation (1) the backscattered intensity, the bubble diameter $D = 2a$ and the averaged incident intensity will have the following proportionality:
\[ I_s \propto D^2 \tilde{I}_{r,w,D} \quad (2) \]

\( \tilde{I}_{r,w,D} \) is the one-way intensity at a distance \( r \) from the transducer averaged over a circular cross-section with diameter \( D \) centered at a distance \( w \) from the center of the beam. The denominator in equation (1) is neglected since the distance from the transducer to the bubble either can be maintained relatively constant for each bubble, or the intensity loss can be compensated by the use of time-variable gain.

The intensity measured by the transducer, \( I_m \), will be the product of the scattered intensity from the bubble and the transducer sensitivity function, \( A(r,w) \). Thus, a transform from the bubble diameter to the measured intensity will be:

\[ I_m \propto T^{-1}(D) = A(r,w)D^2 \tilde{I}_{r,w,D} \quad (3) \]

Equal sized bubbles yield different intensities in the image when they are located in different distances \( w \) normal to the scan plane. Given the measured intensity and the measured diameter of a bubble, the distance \( w \) can be estimated using equation (3).

The inverse, \( T^{-1}(w) \), of the transform in equation (3) can then be calculated numerically, and calibrated to give the bubble diameter from the measured backscattered intensity.

MATERIALS AND METHODS

To obtain growing gas bubbles, three decompression experiments were performed. In each experiment, two gels containing 0.75 weight percent agar (1.1045-500, Kebo Lab, Oslo, Norway) in tap water were formed in 1 dl containers with a height approximately 8 mm, and were pressurized with nitrogen to 2 bar in a pressure chamber for 24 hours. After a decompression, lasting 5 seconds, the diameters of the generated bubbles were continuously measured for 30 minutes as will be described.

Both gels were submerged in bubble-free water. In each experiment the first gel was viewed in a microscope as described by [Lie], and the diameters of three distinct bubbles were measured every minute.

The second gel in each experiment was imaged with a modified ultrasound scanner (CVIS Insight, Cardiovascular Imaging Systems, Inc. CA, USA) and a 20 MHz (8F) catheter probe. The modification involved bypassing the logarithmic ampli-

![Figure 2. A 20 MHz ultrasound image of a gel showing several bubbles (three of them marked with arrows).](image)

To be able to find the transform described in equation (3), the radiated field, \( I_{r,\epsilon}(w) \), from the ultrasound transducer is required. This was measured normal to and parallel to the scan plane using a 0.4 mm hydrophone (NT28-4, Force Institutes, Copenhagen, Denmark) and a XYZ-positioning system. The intensity field is assumed to be symmetric around the axis, and the value for \( I_{r,\epsilon}(w) \) was found by averaging the intensities at a distance \( w \) from the peak in the X and Y intensity fields.

Furthermore, the off-axis distance \( w \) for each bubble seen in the ultrasound images needs to be estimated. A simulation of equation (3) with different \( D \) and different \( w \), gave simulated intensities that were calibrated using the measured bubble diameters and intensities. Thus, when the bubble diameter, the measured intensity and the intensity field are known from the previous measurements, an estimate for \( w \) is given.

Now, the transform \( T^{-1}(D) \), the inverse of \( T^{-1}(D) \) from equation (3), could be computed numerically using the measured intensity field for a depth of \( r = 10 \) mm, and the estimated off-axis distance \( w \) for each bub-
The transform was calibrated using the mean of the visual measurements and the ultrasound measurements for each bubble.

Finally, the bubble intensities in the ultrasound images were transformed and the result was compared to the visually measured bubble sizes.

RESULTS

A total of nine bubbles were measured visually in the three experiments. The visual inspection showed that the bubbles were oblate-shaped rather than spherical, and that they had random orientation. Each bubble still had a stable orientation and shape over time. The longest axis was measured and considered to be the bubble diameter. The mean values and the standard deviations of the visually measured bubble diameters are shown in Figure 3.

![Figure 3](image)

Figure 3. Mean value and standard deviation of the visually measured diameter of 9 bubbles as a function of time after decompression.

The peak intensities from a total of 13 bubbles in the three experiments were recorded using ultrasound imaging. The mean and standard deviation of the intensities are shown in Figure 4. The intensity is given as a relative number due to the unknown amplification in the ultrasound receiver. As can be seen in the figure, there was a large variance in the intensities from the different bubbles.

![Figure 4](image)

Figure 4. Mean value and standard deviation of the peak intensity for 13 bubbles in the ultrasound images as a function of time after decompression.

The intensity field at a distance of 10 mm from the transducer was found not to be circular symmetric, as shown in Figure 5. A field circular symmetric around the beam axis was still assumed for the calculation of the transform, and a circular symmetric average was calculated from the measured field.

![Figure 5](image)

Figure 5. Measured intensity fields normal to and parallel to the scan plane in a distance 10 mm from the center of rotation of the mirror. The circular symmetric average field is shown as a dashed line.

The off-axis displacement w was estimated for the 13 bubbles seen in the ultrasound images, and the calculated and calibrated transforms T(L) for the different estimates of w is shown in Figure 6.
DISCUSSION

The bubbles were found to be non-spherical. This means that the theory for spherical bubbles is not directly applicable. Still, the resulting non-isotropic scattering is of little importance since the bubbles have a stable orientation and shape.

Also, from Figure 5 we see that the incident intensity varies significantly over a transverse distance comparable to the bubble diameters. Thus the constant amplitude wave assumption is questionable. The use of the average incident intensity over the bubble cross-section seems the best approximation.

Furthermore, the intensity field used in the experiments and simulations was found to be non-symmetrical around the beam axis. The true intensity field could easily be measured and used, but this would probably have little influence on the results.

Finally, in spite of these discrepancies, there seems to be a good fit between the transformed and visually measured data.

CONCLUSIONS

This study has demonstrated that the growth of gas bubbles in clear gels can be measured continuously using ultrasound imaging. The method requires knowledge of the transducer intensity field, and at least one visual measurement of the bubble sizes for calibration.

REFERENCES


UTILIZATION OF A PANTOGRAPH LOCALIZATION SYSTEM FOR SPECKLE REDUCTION BY SPATIAL COMPOUNDING, IN B-SCAN ECHOGRAPHY

Ange Hernandez, Olivier BAsset, Pierre Chirossel*, Christian Cachard, Gerard Gimenez
Creatis ura CNRS 1216, INSA 502, 69621 Villeurbanne Cedex, France; * Hôpital Cardio-Vasculaire et Pneumologique Louis Pradel, service de radiologie, Bron, France

SUMMARY

A spatial compounding system has been designed in order to reduce the speckle generated in echographic images. It consists in constructing an improved image from the combination of several different images of the same cross-sectional plane. The "final" image is constructed by the registration and the averaging of the "origin" images. For this, the relative position in the space of the origin images has to be known. The use of a localization articulated arm, on which the ultrasonic probe is fixed, makes it possible. The reduction of speckle is shown on images of an agar gel phantom and on images of the kidney of a human being.

INTRODUCTION

Some years ago, an articulated arm was used in order to know the localization of a single transducer element in manual contact B-scan imaging system (Wells 1977). Since then, this system has been abandoned and replaced by mechanical and electronical probes which have permitted the appearance of real-time systems. However, the localization arm is still used in 3D imaging applications, as long as the 2D arrays are not available (Hottier and Billon 1990).

In this study another interest of the scan arm is presented: the spatial compounding for the reduction of the speckle in B-mode echographic images.

The presence of a large amount of scatterers in the biological tissues generates on echographic images a granular texture called speckle (Buckhardt 1978). The scatterers within an elementary resolution cell emit wavelets which interfere with each other. The interferences which are mainly constructive give a high amplitude echo on the image. The interferences which are mainly destructive give a low amplitude echo. Because of the speckle, the information in the image corresponding to large structure (specular reflection) is blurred and the image is low contrasted. The reduction of speckle is a field of research largely explored to improve the image quality and to increase the signal to noise ratio. Several studies conducted on speckle reduction use the technique of the spatial compounding (Shattuck and Von Ramm 1982, Trahey et al. 1986), but to our knowledge, none of these studies uses an articulated arm.
METHODS

Data acquisition

The acquisition of the spatial coordinates of the images is done with an articulated scan arm. It is composed of several segments linked to each other by six articulations. The first segment is fixed on a support which is immobile in the laboratory reference. The other extremity of the articulated arm, on which the transducer probe is fixed, can move in a large volume thanks to the six degrees of freedom given by the articulations. The knowledge of the different segment lengths and of the angular values transmitted to each articulation, allows us to calculate the echographic probe position in a coordinate system relative to the support, after six successive coordinate system changes. The angular values are given by a potentiometer placed in each articulation. It returns an electrical voltage proportional to the rotation angle of the articulation. The electrical signals provided by the six articulations are converted in digital signals with a Digital Analog Converter (DAC). Then, they feed a computer (PC 386 - 33MHz) via an acquisition card (Nautilus LOGIC-40). The localization of the probe can be calculated and consequently the spatial position of the corresponding image.

For the echographic image acquisition, an image is frozen on the scan screen simultaneously with the corresponding coordinate acquisition of the probe. The image is then digitalized and stored in the Personal Computer. The ultrasonic scanner is a RADIUS (General Electric), equipped with two sectorial phased arrays of 3,5 MHz (HRA) and 7,5 MHz (MRA). The images are digitized with a Matrox PIP 1024 card, in 512 x 512 pixels, with 256 grey levels. The acquisition set-up of the images and their coordinates is presented in Fig.1.

![Connection of the different equipments allowing the acquisition of images and their spatial coordinates.](image_url)

The various images used for spatial compounding must be acquired in the same cross-sectional plane.

Image reconstruction phase

The first step of the image reconstruction is the registration of the images obtained under different angles of view during the acquisition phase. To clarify the understanding of this paper,
the registration technique is described from only two original images. Image 1 stays fixed in the screen plane. Image 2 undergoes a rotation and a translation to be registered with regard to image 1. The rotation and the translation parameters are calculated from the coordinates of each image stored in the computer.

The second step of the reconstruction phase is the average, pixels with pixels, of the N registered images which gives the reconstructed image. The greater the number of images, the more the noise is attenuated.

RESULTS

The reconstruction method has been validated on an agar phantom. It consists in a cylinder of agar mixed with talc, 2.3 cm in diameter, included in a larger cylinder only made of agar. The talc is added in order to modify slightly the acoustic properties in the small diameter cylinder. Sixteen cross-section images are acquired around this phantom with a slightly different angle of view. One of these images is presented in Fig.2A. Then, these images are registered and averaged. The reconstructed image is presented in Fig.2B.

![Fig.2. Validation of the reconstruction method for acoustic speckle reduction. Image B is reconstructed from the registration and the averaging of 16 independent images, acquired on the same plane of an agar phantom. Image A shows one of these images.](image)

The visual observation of this figure shows an obvious reduction of speckle. The contour of the disk is preserved, artefacts are eliminated and the two regions (the disk and the surrounding medium) present a smooth texture.

An in vivo speckle reduction has been realized from cross-section images of a normal subject abdomen. These images show a partial view of the kidney and the liver of this patient. Figure 3A presents one of the eleven images which were acquired. The origin images are all recorded at a same instant of the breathing activity, in order to limit abdominal organ movements and then reconstruction errors. The Fig.3B shows the reconstructed image, after the registration and the
average of only six images among the eleven which were acquired. In point of fact, N=6 images seem to be in this case the best compromise between the speckle reduction and the preservation of tissue details.

Fig 3. Speckle reduction on a partial view of the kidney and the liver of a normal subject. A presents one of the eleven images acquired on a same cross-section plane of the kidney and the liver. B shows the speckle reduction after the registration and the average of six of the acquired images.

The necessary condition to reduce efficiently the speckle by spatial compounding is the decorrelation of the ultrasonic signals of each origin image. For this, the origin images are acquired with different incidences. Buckhardt (1978) calculated theoretically that the signal amplitude of two adjoining exploration lines are "almost" decorrelated if the transducer is moved by half its width. In our study, in order to decorrelate the acquired images, the echographic probe has undergone a rotation with an angle of nearly 4 degrees and a translation of nearly 5 mm.

CONCLUSION

A spatial compounding system has been developed in order to reduce the speckle, in echographic B-mode images. The registration of the images is made by an articulated scan arm. The average error has been measured at less than 1mm in the reconstructed images. The increase of the SNR as a function of the number of registered images follows approximately a square root function.

REFERENCES

Hottier, F.; Collet Billon, A. 3D echography: status and perspective. In: 3D imaging in medicine, ed. Höhne, K.H.
Berlin: Springer-Verlag; 1990:21-41.
MEDICAL ULTRASOUND TRANSDUCERS AND BEAMFORMING

Sverre Holm
Dep. Informatics, University of Oslo, P.O. Box 1080, Blindern, Oslo, Norway

SUMMARY

The development of piezocomposites has been the main recent source for improved performance of transducers. This has given improved nearfield, higher bandwidths, and is about to give 1.5D probes with two-dimensional focusing, and eventually 2D arrays for steering and focusing of beams in 3D data volumes.

Due to the advent of digital beamformers implemented with custom VLSI chips, there has also been an improvement in image quality due to a more precise control over the beam, as well as provision for individual correction of transducer imperfections. In addition the reception of several beams in parallel is possible. New hybrid image formats such as steering of linear and curved arrays, and scanning of phased arrays in combination with steering, now emerge. Adaptive correction for phase aberrations due to tissue inhomogeneities may also become reality.

INTRODUCTION

Beamforming in ultrasound instruments for medical imaging has traditionally been implemented using analog delay lines. Typically 1D arrays with between 32 and 192 elements are used such as shown in Fig. 1, upper right-hand panel. The signal from each individual element is delayed in order to steer the beam in the desired direction. In addition focusing of the beam is performed. In the receive beamformer this gives rise to the concept of dynamic focusing. For each pulse which is transmitted from the array, the receive beamformer tracks the depth and focuses the receive beam as the depth increases. It is also important to let the receive aperture increase with depth. This gives a lateral resolution which is constant with depth, and decreases the sensitivity to aberrations in the imaged medium. This gives a requirement for dynamic control of the number of elements that are used. Since often a weighting function (apodization) is used for sidelobe reduction, the element weights also have to be dynamically updated with depth.

Optimization of image quality in medical ultrasound instruments requires a system's perspective on transducers and beamforming as there is an interaction between new developments in both fields. In this paper developments in transducers and beamformers and their relationship will be discussed.

COMPOSITE ARRAYS

The advent of piezocomposites has been the main recent development in transducer technology. A piezocomposite is a combination of a piezoelectric ceramic and a polymer which forms a new material with different piezoelectric properties. Piezocomposites have improved the performance of
commonly used arrays such as the mechanically scanned annular array and the linear phased array of Fig. 1 upper panels, in the following ways [2]:

1. Acoustic impedance is reduced giving a better impedance match with tissue. This results in a reduction in reverberation level in the near field as the transducer surface to a less extent reflects back incident energy.
2. The composite materials make the radiators closer to the ideal of a vibrating piston. Primarily this is due to the suppression of unwanted surface waves propagating laterally over the transducer.
3. Piezocomposites make it possible to vary the electromechanical coupling constant, thus allowing more control over the trade-off between sensitivity and bandwidth. In general there is a trend towards transducers with large bandwidth, typically 60-80% relative bandwidth (-6 dB).

![Array types: Annular array, rectangular array, 1.5D array, and 2D array](image)

Fig. 1. Array types: Annular array, rectangular array, 1.5D array, and 2D array

New transducers that open up new possibilities are now possible:

1. Transducers that consist of two overlaid orthogonal linear phased arrays, resulting in a biplane transducer. The element patterns for the two arrays are defined on opposite sides of the ceramic. Grounding of one set of electrodes and scanning on the other gives an image in one plane, reversing the roles gives the other plane.
2. One of the advantages an annular phased array has over a linear phased array is variable focus in both dimensions. The 1.5D array shown in Fig. 1, lower left-hand panel, is designed to give variable focus in the short-axis direction also. The 5 elements shown are patterned according to a Fresnel lens. Typically such an array will require a doubling of beamformer channels. The figure shows a 32 element array where 22 + 10 elements are added on each side. Due to sym-

- 340 -
metry the side elements can be connected in parallel, resulting in 64 channels. A realistic array would have two to three times more elements than this [12].

3. Finally the composite technology makes it possible to consider true 2D arrays with square elements. This enables beam steering in 3D space [6]. Such arrays are a prerequisite for real time 3D imaging with transducers without moving parts. One of the challenges of such designs is the small size of each element, resulting in a high electrical impedance. One approach to overcome this is a multilayer composite where a layered structure of ceramic are connected electrically in parallel, but acoustically in series. Placing some of the scanner electronics in close vicinity to the transducer also helps overcome the problem of poor impedance match [10].

DIGITAL BEAMFORMING

Digital beamforming is now about to become feasible in beamformers for medical ultrasound. The concept has long been known, but availability of high-speed analog to digital converters, and VLSI technology improvements have now made digital beamformers feasible [4].

Digital beamforming first of all gives better control over time delay quantization errors. In analog beamformers, delay accuracy is typically in the order of 20 ns. For operation at frequencies at or even above 10 MHz, quantization noise will manifest itself as an increase in sidelobe level and thus reduce contrast resolution [7], [8]. In digital beamformers the delay accuracy can be greatly improved, thus allowing higher frequency operation.

A second advantage of digital beamformers is that it is possible to implement true time delay beamforming. This assures close-to-ideal operation over a wide bandwidth and is a necessity as transducer bandwidths have increased. Many analog beamformers have been implemented using a mixture between wide-band time-delay techniques and narrow-band phase-compensation techniques and will not allow true wideband operation.

The improved near-field of modern transducers also has an impact on design of beamformers. With analog beamformers the number of focal zones is related to size and cost, and therefore the number is kept as low as possible. A simple calculation of focal zone size in the near-field will give an indication of what is required to really take advantage of the clearer near-field of modern composite transducers. The depth of field is given by $6\lambda FN^2$ for an allowed 2 dB of defocusing loss [5]. In this expression $\lambda$ is the wavelength, and $FN$ is the f-number, i.e. the ratio of focal depth and aperture. A high frequency transducer at 10 MHz, operating at an f-number of 1.5 will give a depth of focus of about 2 mm. This necessitates continuous focus beamforming [9].

Even though transducers have been improved over the years, they are still not ideal. Sources of degraded image quality are found mainly in a time delay variation and in a variation in sensitivity from element to element. The time delay variation is caused by the surface not being exactly plane or curved. The variations can be measured and stored in the probe, and the beamformer can use the information to improve the image quality.

A more challenging effect to compensate is phase aberrations caused by for instance fat layers in the tissue. This requires adaptive compensation of time delays. So far only experimental systems have been demonstrated using this technique [3], but it may become reality in the future.

In addition to obtaining a more accurate realization of a beamformer, digital beamforming also opens up for new possibilities such as several parallel receive beams [1]. Although this has been around for a long time for instance in sonars, it is quite recent in medical ultrasound. This is primarily due to the high frequencies and high bandwidths involved. Front-ends with two to four parallel beams are now being introduced [9]. This can give a corresponding increase in frame rate, something which is of great importance in certain modes in cardiology, especially pediatric imaging and in color flow imaging. A 3D imaging system requires parallel beams over an area instead of along a line and thus may benefit from using four to sixteen parallel beams. This comes in addition to a vast increase in elements, typically from 64 for a phased linear array to $\pi 64^2/4 = 3200$
for a circular 2D array. Techniques for reducing this number, such as sparse array methods [11], are now being investigated in order to reduce the number of front-end channels.

Analog beamformers have traditionally been of two types: phased and sequentially scanned. The phased beamformers have enough delay to handle linear phased arrays and are used for sector formats such as required in cardiology. The less expensive sequentially scanned beamformers are used for focusing and slight steering of linear and curved arrays (i.e. about 20% of the delay of the phased beamformer). These probes are used in abdominal and obstetric applications, and are characterized by the same delay pattern sequentially repeated over the surface of the transducer. As beamformers become less expensive a larger proportion of the ultrasound scanners will contain a phased beamformer. This makes new hybrid image formats possible. Formats such as steering of linear and curved arrays that up till now only have been scanned, and scanning of linear phased arrays in combination with steering now emerge. Thus there is a tendency for the distinct transducer array categories to merge.

CONCLUSION

Piezocomposites have improved performance of transducers in common use today as well as made 2D arrays a possibility. On the other hand the development of the digital beamformer has meant that there is a shift in cost: analog channels including transmitters, preamplifiers, connectors, and cables are still expensive while digital capabilities such as increased accuracy and parallel beams will become less costly. This development gives increased image quality of today's scanners as well as opens the way for real-time 3D ultrasound systems in the near future.

REFERENCES

DEVELOPMENT OF ULTRASOUND TRANSDUCERS FOR MEDICAL IMAGING

Hans T. Kaarmann, reinhard Lerch
Siemens medical systems, Inc., ultrasound group, Issaquah, WA, USA
* Institut für elektrische messtechnik, University of Linz, Austria

SUMMARY
The latest trends in the development of medical ultrasound transducers are reported. The optimization of the main transducer performance parameters can be obtained by using new transducer designs, materials, fabrication techniques as well as sophisticated computer modeling tools.

INTRODUCTION
Medical imaging utilizing ultrasound waves made significant progress from its first serious attempts about 30 years ago. The ultrasound transducers are still the crucial elements in the signal processing chain which determine the achievable image quality. The first imaging systems provided a black-and-white image constructed by mechanically moving a single-crystal transducer across the surface of a human body. Today mostly high-density transducer arrays are used (with element counts of up to 512 and densities of more than 50 elements per inch) and the real-time signal processing provides gray-scale structural and color-coded flow information from inside the human body.

TRANSDUCER DEVELOPMENT
On the one hand the transducer development is driven by requirements from the medical application. Those requirements are size and shape (for utilizing all extracorporeal acoustic windows, all possible human cavities, and even interventional access like intra-operative or intraluminal), field of views, and resolution in order to provide an answer to a given medical diagnosis problem.

On the other hand those requirements flow together with the results from technological developments in the area of materials for ultrasound transducers and advanced fabrication methods, both heavily supported by simulations in order to reduce the otherwise huge experimental workload.

TRANSDUCER PERFORMANCE PARAMETERS
The most important performance parameters are those which affect directly the resolution requirements. However, due to the high attenuation of the acoustic waves in human tissue (about 1 dB/(MHz*cm)) that requires a high cleanliness in the frequency and temporal response (down to -80 dB levels) of an ultrasound transducer in order to achieve an acceptable axial resolution. That requires in general extensive experimental and simulation efforts.
In addition the lateral resolution in lateral direction needs also attention at levels below -50 dB. That makes up for the uniformity requirements for sensitivity and time-of-flight (focusing) which are still challenging for some fabrication techniques. The high attenuation in the human body together with the limits for the transmitted power (regulatory requirements) also asks for high sensitive transducer materials and for a perfect matching of the transducer to the input circuitry of an ultrasound imaging system.

DEVELOPMENT TRENDS FOR ULTRASOUND TRANSDUCERS

Transducers which are flexible to use in several different applications are today the high-density phased array transducers. As a standard they come with element counts of 256 in the frequency range from 2 to 10 MHz. They support all common imaging formats like parallel, trapezoidal, and sector images. Whereas today mostly one-dimensional arrays are in use, further improvements of the "slice-thickness" of the cross-sectional images ask for capabilities to control the aperture, to focus, and to steer in the elevation direction of a scanning ultrasound array. That is realized in 1.5-dimensional and 2-dimensional arrays, respectively. Special applications for transducers like endo-cavity applications require small size transducers with convenient shapes in order to introduce them into the human body. On the other hand such applications ask for wide field-of-views, which is accomplished with microconvex transducer arrays. Today such transducer arrays provide up to 360 degree field-of-views at a diameter of less than about 12 mm with sufficient sensitivity and resolution. For even more sophisticated applications like intraluminar imaging even far smaller arrays at significant higher frequencies are possible.

MATERIAL DEVELOPMENTS

Supporting developments continue in the field of both, the active materials as well as the passive materials. Today's mostly used active materials are the well known piezo-polymers (like PVDF), piezo-ceramic materials, and composite materials made out of those [1]. Passive materials are used for backing and matching the active materials in order to tailor the desired performance parameters by. This can be done by development of materials with adjustable acoustic impedance in the range from 3 to 12 MRaysls.

FABRICATION TECHNIQUES

The development of fabrication techniques is challenged by the mechanical, electrical, acoustical, and chemical (application environment including disinfection and sterilization) requirements. In the area of interconnect technologies it asks for a high-density interconnect to high-element micro-coax cables, or for an integration of at least part of the input electronic circuitry into the scanhead. The other fabrication techniques try to strengthen the mechanical and chemical stability of the array assemblies.

MODELING

Even today the most commonly used models to simulate the performance of an ultrasound transducer are the Mason or KLM models. However, due to the one-dimensional nature of those models they are - despite some sophisticated adaptations - not the appropriate tool to describe the behavior of complex transducer arrays. The more advanced tools here are simulations based on finite element modeling including piezoelectric and acoustic elements [2].
RESULTS
The following should only give examples of some new developments where the tools mentioned applied.
Fig. 1 shows the time response of an element in an 10 MHz ultrasound array built out of an 1-3-composite-ceramic material. The pulse shows a very clean shape without phase reversals even at very low ring-down levels. That is due to the absence of lateral vibration modes.
Fig. 2 shows a comparison of measured and simulated time and frequency response of an element out of an 5 MHz phased array transducer, where the thickness of one of the matching layers is varied by 20 microns. The exceptional precise prediction of the simulation can be seen.

CONCLUSIONS
Today's ultrasound transducer for medical imaging are driven in their development by new medical applications with their requirements as well as by the development of materials and fabrication techniques. Given those increasing numbers of degrees of freedom it is necessary to shorten the development time to reasonable durations by making use of advanced simulation tools.

REFERENCES

Fig. 1: Time response of 10 MHz transducer element built out of an 1-3-ceramic composite material
Finite Element simulation

Standard matching layer (nominal thickness)

Modified matching layer (nominal thickness + 20 micron)

Measurement

Modified matching layer (nominal thickness + 20 micron)

Echo voltage (a.u.)

Magnitude (dB)

Frequency [MHz]

Fig. 2: Pressure response of 5-MHz array transducer element
EXTENDED AUTOCORRELATION METHOD FOR COLOR FLOW IMAGING

Xiaoming Lai, Hans Torp, Kjell Kristoffersen*
Department of Biomedical Engineering, University of Trondheim, Norway
* Vingmed Sound, Horten, Norway

SUMMARY

The conventional autocorrelation method for color flow imaging (CFI) is based on the phase estimation of the autocorrelation function with temporal lags. A new method for velocity estimation based on the autocorrelation function with lags both in temporal and axial direction, is presented. The new algorithm shows better performance than the conventional autocorrelation technique in the estimation variance and the capability of resolving velocity ambiguity. This can be explained by the axial information added in the estimation. The performance of this new algorithm is compared to the RF cross-correlation technique in the estimation variance and the ability to estimate the maximum velocity up to three times the Nyquist limit. The estimation variance is calculated by simulation using a theoretical signal model with different pulse bandwidth and signal-to-noise ratio. The improvement of this new algorithm is demonstrated by digitized ultrasound RF-data from a jet flow in a water tank model.

INTRODUCTION

The conventional autocorrelation method for CFI is based on the phase estimation of the autocorrelation function. Parameters such as velocity, velocity spread and signal power are calculated from the autocorrelation function of the complex demodulated pulsed Doppler signal from each range gate on the beam axis. They are coded into a color image which overlay with the standard gray level tissue image. This method was firstly developed for weather radar applications and applied to ultrasound blood velocity measurement in 1983 by a Japanese group[1]. There are mainly two limitations in this method. One is the velocity ambiguity problem caused by the sampled nature of pulsed Doppler. The other is the large variance of the velocity estimation. Because high range resolution need wideband transducer which leads to the increase of variance in the velocity estimation.

An alternative algorithm called crosscorrelation method for blood velocity estimation has been presented by Bonnefous[2]. The crosscorrelation algorithm is based on estimation of the time delays of the received echoes from the pulse-to-pulse from the crosscorrelation function of the RF signal. The two limitations of the conventional autocorrelation can be overcome by the crosscorrelation method[2]. Because the time delays is found by detecting the peak amplitude of the crosscorrelation function, the wrong peak detection can happen in some cases because of the estimation variance of the crosscorrelation function. In this paper, the velocity estimation error caused by wrong peak detection of the correlation function is called aliasing estimate error.

In this work, a new extended autocorrelation (EAM) is proposed. It uses both of the phase and the amplitude of the autocorrelation function to estimate the velocity. It uses firstly the phase of the autocorrelation function to give a limited number of candidates in time delays which is proportional to the velocity. Then it selects the candidate which corresponds to the maximum amplitude of the autocorrelation function. The true velocity is determined by the selected candidate, consequently. The similarity between the EAM and crosscorrelation method is analyzed theoretically, and a quantitative comparison for different velocity estimators is performed with computer simulations.

ALGORITHM DESCRIPTION

Assuming $z(t, k)$ is the received complex demodulated signal. The parameter $t$ is the elapsed time after pulse transmission, which corresponds to a certain range distance from the transducer, and $k$ is pulse number. If there is a frequency dependent attenuation, the signal spectrum of $z(t, k)$ is not symmetrical around zero frequency but around a frequency of $\Delta \omega$. This $\Delta \omega$ can be estimated by the phase of the autocorrelation function $\hat{R}_x(1, 0)$. If $x'(t, k)$ is the frequency shifted complex demodulated signal of $z(t, k)$ and assume the spectrum of $x'(t, k)$ is symmetrical,
around zero frequency. Then we have: \( x(t, k) = e^{-j\omega t} \mathcal{R}_x(t, k) \). The relation between the autocorrelation functions is

\[
\mathcal{R}_x(t, m) = e^{-j\omega t} \mathcal{R}_x(t, m)
\]

(1)

where \( \tau \) is the arbitrary lag in range, and \( m \) in time (or pulse number). The approximate relation between the correlation function of the RF signal \( \mathcal{R}_x(t, m) \) and autocorrelation function \( \mathcal{R}_x(t, m) \) was derived in a previous paper [3] as:

\[
\mathcal{R}_x(t, m) = \frac{1}{2} Re \left\{ e^{j\omega t} \mathcal{R}_x(t, m) \right\}
\]

(2)

where \( \omega_0 \) is the center frequency in the transmitted signal. From (1) and (2), we have:

\[
\mathcal{R}_x(t, m) = \frac{1}{2} Re \left\{ e^{j(\omega_0 + \Delta\omega) t} \mathcal{R}_x(t, m) \right\}
\]

(3)

If the envelope of \( \mathcal{R}_x(t, 1) \) is sufficiently smooth, then the peak in the correlation function \( \mathcal{R}_x \) occurs when \( \tau = \tau_{\text{max}} \) i.e.

\[
\text{phase} \left\{ e^{j(\omega_0 + \Delta\omega) t} \mathcal{R}_x(t, 1) \right\} = 0
\]

(4)

So the following relation between the autocorrelation phase angle estimate and the 'peak crosscorrelation estimate' can be found as:

\[
(\omega_0 + \Delta\omega) \tau_{\text{max}} = 2\pi n - \text{phase} \left\{ \mathcal{R}_x(t, \tau_{\text{max}}, 1) \right\}
\]

(5)

In [4], it is shown that the phase of \( \mathcal{R}_x(t, 1) \) is independent of \( \tau \) when there is only one velocity component inside the sample volume. That means we can use the phase of \( \mathcal{R}_x(t, 1) \) for any \( \tau \) instead of the phase of \( \mathcal{R}_x(t, \tau_{\text{max}}, 1) \).

From (1) and (5), we have:

\[
(\omega_0 + \Delta\omega) \tau_{\text{max}} = 2\pi n - \text{phase} \left\{ \mathcal{R}_x(t, \tau_{\text{max}}, 1) \right\} + \Delta\omega \tau_{\text{max}}
\]

(6)

The new method is firstly to estimate the phase of \( (\mathcal{R}_x(t, \tau_{\text{max}}, 1)) \) and find a number of the candidates which are:

\[
\tau_n = \frac{1}{2(\omega_0 + \Delta\omega)} \left( 2\pi n - \text{phase} \left\{ \mathcal{R}_x(t, \tau_{\text{max}}, 1) \right\} + \Delta\omega \tau_{\text{max}} \right), \quad (n = 0, \pm 1 \ldots \pm K)
\]

(7)

where \( \tau_{\text{max}} \) is the rough estimate of \( \tau_{\text{max}} \), because true \( \tau_{\text{max}} \) is unknown in this step and \( \tau_{\text{max}} \) is found by maximizing the amplitude of \( \mathcal{R}_x(t, 1) \). \( K \) is the number of the Nyquist repetition. Then, it uses the amplitude of \( \mathcal{R}_x(t, 1) \) to determine which \( \tau_n \) corresponds to the maximum amplitude of \( \mathcal{R}_x(t, 1) \). The velocity is determined by the selected \( \tau_n \).

In a practical situation, the autocorrelation function is sampled in the radial coordinate \( \tau \). The amplitude of \( \mathcal{R}_x(t, 1) \) for any \( \tau \) can be found using some kind of interpolation technique.

SIMULATION RESULTS

To evaluate the velocity estimators, some simulation experiments were performed. In paper [4], a parametric model for the 2D signal from blood flow with constant, rectilinear velocity field is described. The signal is completely described by the single scatterer response, where the transmitted pulse waveform, beamwidth and velocity magnitude and direction can be selected. Two-dimensional blood flow signals were generated by convolution of a matrix of independent Gaussian random variables, with the ultrasonic system single scatterer response[3]. The simulation parameters in this paper were given as follows.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transducer center frequency</td>
<td>2.5MHz</td>
</tr>
<tr>
<td>Pulse repetition frequency</td>
<td>6564Hz</td>
</tr>
<tr>
<td>Speed of sound</td>
<td>1540m/s</td>
</tr>
<tr>
<td>Nyquist velocity</td>
<td>1.0265m/s</td>
</tr>
<tr>
<td>RF sampling rate</td>
<td>10MHz</td>
</tr>
<tr>
<td>Ultrasonic measurement angle</td>
<td>30degree</td>
</tr>
<tr>
<td>Temporal averaging</td>
<td>1.8ms</td>
</tr>
<tr>
<td>Radial averaging</td>
<td>3 * 2.4µs or 3 * 0.8µs</td>
</tr>
</tbody>
</table>
The velocity estimators had been applied to simulated signals with constant velocity \( v = 0.2 \text{ m/s}, v = 0.5 \text{ m/s} \), \( s = 1.2 \text{ m/s}, v = 2.2 \text{ m/s} \) and \( v = 3.2 \text{ m/s} \) from 50 independent simulations. The values in the table are standard deviation or the probability of aliasing estimate error of velocity estimators. The probability is denoted by "P" and it is defined as

\[
P = \frac{\text{the number of aliasing estimate}}{\text{the total estimate number}} \times 100\%
\]

The probability of aliasing estimate error will be displayed in the table instead of standard deviation when \( P > 0 \).

The result shows that the EAM has less velocity estimate variance than the conventional autocorrelation method both for low bandwidth and high bandwidth. This is because the radial information has been added to the EAM. There is no significant difference between the EAM and the crosscorrelation method. Both of them have the ability to estimate the velocity which is beyond Nyquist limit and give the similar variance. The aliasing estimate errors have been observed in the case of low pulse bandwidth and low signal to noise ratio. This is because the correlation function of low pulse bandwidth or under poor SNR condition is flat compared to high pulse bandwidth and under high SNR condition, respectively. The estimation variance of the correlation function has heavy influence on the detection of the true peak in the correlation function and the aliasing error can occur.

### Table 1: The variance of the velocity estimators

<table>
<thead>
<tr>
<th>pulse length</th>
<th>velocity</th>
<th>Autocorr. SNR=∞</th>
<th>EAM SNR=∞</th>
<th>Cross-corr. SNR=∞</th>
<th>Autocorr. SNR=0</th>
<th>EAM SNR=0</th>
<th>Cross-corr. SNR=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4 μs</td>
<td>0.2 m/s</td>
<td>0.0241</td>
<td>0.0120</td>
<td>0.0122</td>
<td>0.0573</td>
<td>10%(P)</td>
<td>10%(P)</td>
</tr>
<tr>
<td>2.4 μs</td>
<td>0.5 m/s</td>
<td>0.0344</td>
<td>0.0163</td>
<td>0.0164</td>
<td>0.1105</td>
<td>4%(P)</td>
<td>0.0192</td>
</tr>
<tr>
<td>2.4 μs</td>
<td>1.2 m/s</td>
<td>______</td>
<td>0.0177</td>
<td>0.0179</td>
<td>______</td>
<td>4%(P)</td>
<td>4%(P)</td>
</tr>
<tr>
<td>2.4 μs</td>
<td>2.2 m/s</td>
<td>______</td>
<td>0.0206</td>
<td>0.0209</td>
<td>______</td>
<td>8%(P)</td>
<td>10%(P)</td>
</tr>
<tr>
<td>2.4 μs</td>
<td>3.2 m/s</td>
<td>______</td>
<td>0.0237</td>
<td>2%(P)</td>
<td>______</td>
<td>6%(P)</td>
<td>4%(P)</td>
</tr>
<tr>
<td>0.8 μs</td>
<td>0.2 m/s</td>
<td>0.0317</td>
<td>0.0095</td>
<td>0.0094</td>
<td>0.1696</td>
<td>0.0136</td>
<td>0.0136</td>
</tr>
<tr>
<td>0.8 μs</td>
<td>0.5 m/s</td>
<td>0.0609</td>
<td>0.0165</td>
<td>0.0162</td>
<td>0.1149</td>
<td>0.0189</td>
<td>0.0194</td>
</tr>
<tr>
<td>0.8 μs</td>
<td>1.2 m/s</td>
<td>______</td>
<td>0.0178</td>
<td>0.0170</td>
<td>______</td>
<td>0.0207</td>
<td>0.0190</td>
</tr>
<tr>
<td>0.8 μs</td>
<td>2.2 m/s</td>
<td>______</td>
<td>0.0228</td>
<td>0.0221</td>
<td>______</td>
<td>0.0267</td>
<td>0.0258</td>
</tr>
<tr>
<td>0.8 μs</td>
<td>3.2 m/s</td>
<td>______</td>
<td>0.0196</td>
<td>0.0195</td>
<td>______</td>
<td>0.0241</td>
<td>0.0249</td>
</tr>
</tbody>
</table>

### EXPERIMENTAL RESULTS

a. Water-tank model

Figure 1 shows a schematic diagram of the water-tank. It consists of an upper reservoir tank and a flow tank which has two rooms. There is a small jet aperture between two rooms. The fluid flow from the upper reservoir tank to the left room of the flow tank by a tube with a valve. This valve is used to control the water pressure in the left room which determines the jet velocity. A pump in the right room is controlled by an adaptive water level regulator. The fluid is pumped back to the upper reservoir tank. Figure 2 is an illustration of the jet stream.

b. The parameters in this experiment

- Transducer center frequency: 2.5MHz
- Pulse repetition frequency: 5670Hz
- Temporal averaging: 2.1ms
- Pulselength: 0.8 μs
- RF sampling rate: 10MHz
- Acoustic velocity: 1540m/s
- Nyquist velocity: 0.8732m/s
- Radial averaging: 1.28 μs

### c. The gray flow imaging of the jet stream

Figure 3 is the velocity image estimated by the extended autocorrelation method.

Figure 4 is the velocity image estimated by the conventional autocorrelation method.

Figure 4 illustrates that the velocity aliasing occurs and only low velocities which are below the Nyquist limit are estimated correctly. Figure 3 shows that the EAM can estimate velocities up to three times the Nyquist limit. But the aliasing estimate error occurs at some places in Figure 3.
CONCLUSIONS
The extended autocorrelation method (EAM) of velocity estimation for color flow imaging has been presented. The performance of EAM, crosscorrelation and autocorrelation velocity estimators has been assessed by simulation using a theoretical signal model. The results show that the performance of EAM is better than the conventional autocorrelation method. The EAM and crosscorrelation method give a similar performance in the estimation variance and capability of resolving velocity ambiguity. These two methods can work well for high pulse bandwidth. Therefore, the range resolution and accuracy of the estimated velocity are improved. For low pulse bandwidth and under the situation of poor signal to noise ratio, significant aliasing error occurred for both two methods.

REFERENCES


Figure 1 Schematic diagram of the water tank

Figure 2 An illustration of jet stream

Figure 3 Velocity image of the jet stream using the EAM method

Figure 4 Velocity image of the jet stream using the autocorrelation method
ULTRASONOGRAPHIC STENT-IMAGING ARTIFACTS

Torgrim Lie, Aage Gronningsaeter, Klaus-Dieter Bolz*, Andreas Heimdal
Department of Biomedical Engineering, University of Trondheim, Norway
* Radiology, University Hospital of Trondheim, University of Trondheim, Norway

I. INTRODUCTION

Intravascular ultrasoundography (IVUS) is now widely used as an adjunct to angiography during and after intravascular stent implantation [1-4]. It provides cross-sectional images of the stent, the vessel wall and the composition of the atherosclerotic lesion. Thus the method offers information which is important to evaluate whether proper placement of the stent has been obtained.

However, the IVUS technique is subject to different kind of artifacts and noise sources [5-9]. This paper addresses artifacts related to ultrasound imaging of metallic stents. The presence of metal structures in biological media results in wave-propagation effects that causes the image of a stent to deviate from the true physical shape of the stent. The artifacts obscure the IVUS image and may inhibit the investigation of structural changes in the stented vessel segment as well as the detection of small cavities which may exist between the stent and the vessel wall due to sub-optimal stent placement. It is important that the operator understands these effects in order to avoid misinterpretation.

II. MATERIALS AND METHODS

Stents: Four different stents were included in the study and noted Stent1-Stent4. The types and dimensions are listed in Table 1. Stent1 and Stent2 are balloon-expandable stents, manufactured by cutting slots in a stainless steel tube by laser technique. Stent3 and Stent4 are made of biocompatible stainless steel monofilaments, braided in a tubular mesh configuration.

Instrumentation: An intravascular scanner with a 20 MHz (8F) rotating mirror catheter was used to obtain data in vitro experimental data. (CVIS Insight, Cardiovascular Imaging Systems, Inc. CA, USA.)

Phantom: A vessel-wall phantom was made in order to illustrate the situation where a stent is improperly adapted to the vessel wall. A 1mm thick membrane was made of silicon and folded to form a tube of diameter approximately 9mm and with slightly irregular shape. The IVUS speckle pattern from this material is very similar to that from a vessel wall, however the penetration at 20 MHz is less, typically 2mm.

Table 1 The four stent types used in the experiments. The symbols D and T are referring to stent diameter (expanded) and filament thickness respectively.

<table>
<thead>
<tr>
<th>#</th>
<th>Type</th>
<th>D (3)</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Palmaz-Schatz (PS153)</td>
<td>5mm</td>
<td>64μm</td>
</tr>
<tr>
<td>2</td>
<td>Palmaz (PS30)</td>
<td>7mm</td>
<td>127μm</td>
</tr>
<tr>
<td>3</td>
<td>Wallsten (2)</td>
<td>10mm</td>
<td>152.5±4μm (3)</td>
</tr>
<tr>
<td>4</td>
<td>Wallsten (2)</td>
<td>9mm</td>
<td>131.4±4μm (3)</td>
</tr>
</tbody>
</table>

(1) Johnson & Johnson, Interventional Systems Co, Warren, New Jersey, USA.
(2) Schneider (Europe) AG, Pfizer Hospital Products Group, Zürich, Switzerland.
(3) Measured quantities.

III. THEORY

The three most important wave propagation effects related to IVUS-imaging of intravascular stents are:
i) Shadowing: The acoustic impedance in biological media and metal differ significantly, almost no acoustic power penetrates through a metal structure. As a result of this, a shadow region will occur behind the metal.

ii) Transducer-stent reverberations: As the transmitted acoustic pulse hits a stent, a strong echo is reflected back to the transducer. This is the primary echo from the stent. However, this echo is reflected at the transducer surface causing a secondary pulse which propagates into the medium generating a secondary image of the stent and the surrounding tissue structures. Due to two times longer propagation path, this echo will be superimposed on the primary echo from tissue structures at twice the depth of the stent.

iii) Stent-filament reverberations: When the ultrasound pulse hits a stent filament, several possible wave-propagation effects may occur in the vicinity of the hitting point: I) Energy will be transmitted into the metal and undergo multiple internal reflections. II) The metal filament may be excited and vibrate. III) Energy may be converted to surface waves along the stent surface. These wave-propagation effects may result in re-emitted signals which return back to the transducer after a certain delay, thus resulting in a tail on the primary (and multiple) echo(s).

We postulate that the dominating effect in intravascular stent imaging is the first one: internal reflections in the metal. This is motivated from theoretical and experimental findings: The transducer/blood/metal/tissue-configuration in Figure 1 was modeled theoretically and arranged experimentally. There was a close correspondence between the numerical calculations and the measured results.

IV. RESULTS

A. Echo from thin metal plate

Numerical calculation: A numerical calculation was performed by inserting a digitized pulse measured by a hydrophone as the incident wave in Figure 1. This simulation is described more thoroughly in [10]. The result is shown in Figure 2, where the horizontal axis is range (primary echo to the left, the tail aims to the right) and the vertical axis is metal thickness.

Figure 2 illustrates that the echo tail depends on the plate thickness T. The following observations can be made with reference to the figure:

i) Constructive interference occurs between pulses when $2T = \lambda_k M$ (M=1,2,3,..., and $\lambda_k$ is the wavelength in steel). This results in a bright and long tail occurring for the first time for thickness $T = 142\mu$m. Destructive interference occurs when $2T = \lambda_k/2(2K+1)$ and $K=0,1,2,...$, for the second time for thickness $T = 213\mu$m (K=1).

ii) Energy loss causes the signal (tail) to diminish as a function of range. Our model accounts for no other energy drop than power transmitted out of the plate at each surface. The intensity will drop a certain fraction at each point of reflection. Thus, the tail diminish more rapidly in a thin plate (due to many reflections in a short time interval) than in a thick plate.

iii) When the plate thickness equals half the pulse length in steel, $T = L_p/2$, then multiple pulses will "touch" each other without overlapping. The result is a periodic pattern of resolvable pulses. A periodic pattern also occurs when $T < L_p/2$, i.e., the pulses overlap. The reason is constructive and destructive interference between the overlapping fractions of the pulses.
iv) The primary echo will interfere with the internal reflections and cause constructive or destructive interference depending on $T$. This is shown as a periodic intensity variation along the vertical line in depth 4 mm.

Experiments: The theoretical results presented in Figure 2 were verified experimentally for three different stainless steel plates of dimension (10mm x 100mm) and thicknesses $T_1=150\mu$m, $T_2=200\mu$m and $T_3=700\mu$m. There is a close fit between the experimental results presented in Figure 3 and the theoretical result in Figure 2.

B. In vitro stent images

Figure 4 (left) shows an image of Stent2. The primary echo, the echo tail and the first order transducer-stent reverberation are clearly shown.

The length of the tails is approximately 4-5 mm, but a stent with 127$\mu$m wall thickness is supposed to induce a tail of approximately 1.5 mm length according to Figure 2. This may indicate that other effects than multiple reflections between the filaments surfaces account for the stent-filament reverberations as well.

Figure 4 (right) shows an image of Stent4. The ultrasound beam is wide compared to the mesh-size, thus several filaments contribute to the echo in each beam direction. The result is a quite random speckle pattern, not very different from images obtained from vessel walls.

C. In vitro image of stent in vessel-wall phantom

An experiment was conducted in order to illustrate that filament shadowing and stent-filament reverberations can make interpretation of ultrasound-guided stent-adaptation images difficult. The silicone vessel-wall phantom was fixed over a 20 MHz catheter in a water tank, and Stent2 was inserted between the catheter and the phantom.

The phantom was slightly larger than the stent yielding two water-filled cavities at either side of the phantom. The result is shown in Figure 5 (left) where the stent is easily recognized
between 4 and 10 o'clock, but difficult to distinguish from the phantom in other locations. The lower left cavity is recognized from 5 to 9 o'clock, but the upper right is hard to locate precisely. A reference image is shown in Figure 5 (right) where the stent has been removed, and the white circle indicates the location of the stent in (left).

![Figure 5 Left: Silicone vessel-wall phantom imaged through Stent2 in a water tank. Shadowing and stent-filament reverberations make it difficult to discriminate the stent from the phantom as well as determining the water filled cavities. Right: Same image without the stent. The white circle indicates the location of the stent (left).](image)

### V. CONCLUSION

Early endothelialization of the stent filaments seems advantageous to obtain long-term patency of a stented artery. Re-endothelialization is supposed to occur by lateral growth of endothelial patches between the struts [11]. Therefore documentation of optimal stent adaptation to the arterial wall is of special interest. Proper stent adaptation is also of utmost importance for endovascular treatment of aortic aneurysms. Intravascular ultrasonography can provide such information during the procedure, but our findings demonstrate that mis-interpretation is possible due to wave-propagation artifacts.

The shadowing effect as well as the reverberations inhibit investigation of structural changes in the vessel wall after stent implantation. In addition, small dissections or cavities which may exist between the stent and the vessel wall due to sub-optimal stent adaptation can be difficult or impossible to detect with this method.

### REFERENCES

EXPERIMENTAL OBSERVATION OF SUBHARMONIC OSCILLATIONS IN INFOSON BUBBLES

O. Lotsberg, J. M. Hovem, B. Aksum
Division of Telecommunications, The Norwegian Institute of Technology, N-7034 Trondheim, Norway

SUMMARY
This paper reports the result of an experimental study of nonlinear emission from gas filled bubbles with particular emphasis on the subharmonic emission. The gas bubbles are Infoson, a contrast agent for use in echocardiography consisting of small gas filled microspheres with mean diameter of approximately 4 μm. Pulsed signal at frequencies 3.5 and 4 MHz was transmitted through a cloud of bubbles and the level of the subharmonic component was measured as function of level of the exciting signal. No sharp threshold as expected from theory is found, but a relative rapid increase in subharmonic level occurs when the driving signal exceeds approximately 50 - 100 kPa.

INTRODUCTION
Contrast agents for use in medical ultrasound imaging are now becoming available on the market. The main idea behind these agents is to enhance the acoustic backscatter from blood, and thereby increase the information in the images and improve their diagnostic value. One such contrast agent is Infoson (Nycomed Imaging AS, Oslo, Norway, or in America: Albunex, a registered trade mark of Molecular Biosystems, San Diego, USA). Infoson consists of microspheres containing gas encapsulated in a shell of human serum albumin. The mean diameter of the microspheres is 4 μm, the shell thickness is between 20 and 25 nm. The microspheres are small enough and have sufficient stability to pass the lung capillary system and reach the left side of the heart after intravenous injection, being useful for left ventricular imaging.
Infoson microspheres increases the acoustic backscatter of the blood dramatically, especially when the frequency of the incident ultrasound matches the resonance frequency of the microspheres. The contrast agent is intended for use in echocardiography, with source and receiver frequencies in the same frequency band, typically 2 to 5 MHz.
Oscillations of gas-filled bubbles in a sound field is a very non-linear phenomena. When a gas filled bubble is excited with a sinusoidal acoustic signal, the scattered signal may contain both higher and lower harmonic frequency components. A major difference (at least according to existing theory) is that subharmonic generation occur only when the exciting signal exceeds a certain threshold value. The generation of higher harmonics is,

Lotsberg et al.: Subharmonic oscillations

- 355 -
on the other hand, a continuous process and occurs at various degree for all levels of excitation.

It is relevant to ask if nonlinear effects can be used for imaging and this question was the motivation for the present study. Since generation of higher harmonics is much better known both theoretically and experimentally, emphasis was given to subharmonic generation and in particular to determine the threshold conditions for subharmonic emission from Infason. The measured values are discussed and compared with predictions from existing theory.

**THEORY**

Several theories for non-linear emission of bubbles have been developed and they are summarized in the recent book by Leighton [2]. In particular, the theory developed by Eller and Flynn [3] seems relevant for the present study, and therefore a brief account of their approach is described in the following.

Eller and Flynn start the analysis with the well known equation of a spherical bubble in an incompressible liquid driven by a sinusoidal pressure field.

\[
R \ddot{R} + \frac{3}{2} \dot{R}^2 + \frac{P_A}{\rho} \left[ 1 - \left( \frac{R}{R_e} \right)^n \right] + \frac{P_A \cos \omega t}{\rho} = 0
\]  \hspace{1cm} (1)

Here, \( R \) is the bubble radius, and \( R_e \) is the equilibrium value of \( R \). The quantity \( \rho \) is the density of the liquid, \( P_0 \) is the hydrostatic pressure, and \( P_A \cos \omega t \) represents the applied pressure field of amplitude \( P_A \). In this equation, surface tension and damping are neglected, and the gas within the bubble is assumed to behave as an ideal gas under adiabatic conditions. The ratio of specific heat is \( \gamma \).

First, a solution of equation 1 is found that does not contain subharmonic components, only components with frequencies that are integral multiples of the driving frequency. This solutions may, however, become unstable and a subharmonic component at half the driving frequency is generated when the values of the acoustic pressure and frequency are within certain ranges. Without damping, Eller and Flynn determine the unstable region by

\[
\beta = 4 \pm \frac{2}{3} \rho + \left[ (-87\gamma^2 + 96\gamma + 16) / 648\gamma^2 \right] \rho^3
\]  \hspace{1cm} (2)

where

\[
p = \frac{P_A}{P_0}
\]  \hspace{1cm} (3)

\[
\beta = \frac{\omega}{\omega_0}
\]  \hspace{1cm} (4)

Loxberg et al.: Subharmonic oscillations
\[ \omega_0 = \frac{3gP_0}{\rho R_s^2} \]  

(5)

\( \beta \) is the ratio of the angular frequency to the angular resonance frequency for small-amplitude pulsation.

Damping opposes the generation of sub harmonic and tends to counteract the effect of the non-linearity. Approximately the threshold of acoustic pressure for subharmonic generation is by Eller and Flynn found as

\[ p^2 = \left( \frac{3}{2} \beta^2 - 6 \right)^2 + \left( 6\Lambda / \pi \right) \]

(6)

where \( \Lambda \) is the logarithmic decrement, representing the damping of the pulsation. This result (fig 1) shows that the effect of the damping is to increase the threshold pressure and to give it a non-zero value at \( \beta = 2 \).

**EXPERIMENTAL PROCEDURE**

The bubble cloud was held by a cylinder made of thin rubber. The density of Inhoson was 30 \( \mu l/100 \) ml water. Air bubbles in water tend to float up to the surface, so the measurements had to be done within a few seconds after pouring the mixture into the cylinder.

![Fig. 1 Region of subharmonic generation](image)

The detailed measurements of the sub harmonic generation used a 2.25 MHz receiving transducer. In this experiment the transmitted frequency was 4 MHz. The transmitted signal was a pulsed sinusoidal signal with 200 periods.

After reception the signals were sampled and digitized at a rate of 100 MHz and transferred to a computer for analysis.

**RESULTS**

From the sequence of measurements, one can produce a graph showing the level of the subharmonic component at half the driving frequency as function of the level of the transmitted signal. This graph is shown in figure 2.

Loosberg et al.: Subharmonic oscillations

- 357 -
The theory of Eller and Flynn predicts a very sharp increase in subharmonic generation when the level of excitation exceeds a certain frequency dependent threshold. Such a sharp threshold is difficult to find from the present measurements, but in the region from 50 to 100 kPa the level of the subharmonic component increases with more 20 dB.

Since Infoson contain bubbles with a wide range of diameters (fig 3) one should not expect the threshold to be as sharp as if the bubbles were of one size. It is difficult to compare this result (threshold between 50 and 100 kPa) with the theoretical predictions, because one does not know the damping of the pulsations for the bubbles. It is, however, obvious that if the damping coefficient is between 0.2 and 0.4 (fig 1) a threshold value between 50 and 100 kPa is close to the theoretical value. de Jong et al. [1] has found that, at resonance, the bubbles have a damping coefficient of approximately 0.3.

CONCLUSIONS
In summary, the pulsation of the spherical bubbles in the contrast agent Infoson will contain a subharmonic component of order one-half. This component is strongly dependent on the driving acoustic pressure. It should be possible to use this frequency component as a tool in ultrasound imaging if the driving pressure is strong (higher than approximately 100 kPa).

ACKNOWLEDGMENT
We would like to thank for the support from Nycomed, and especially to Lars Hoff for providing the contrast agent and his assistance with the experiments.

REFERENCES
REDUCTION OF STATIONARY REVERBERATIONS IN CARDIAC ULTRASOUND IMAGING

Michael Nickel, Aage Gronningsaeter*, Bjørn Angelsen
Department of Cybernetics, the Norwegian Institute of Technology
* Department of Biomedical Engineering, University of Trondheim, Norway

SUMMARY

This paper presents a method to reduce reverberations (multiples) obscuring the tissue structure of the heart in cardiac ultrasound imaging. Reverberations constitute signal depend noise with characteristics very similar to those of the primary echo signal. Consequently, they are difficult to remove by simple filtering. However, in cardiac ultrasound imaging reverberations occur mainly near to the transducer due to multiple reflections between subcutaneous stationary tissue layers and the transducer surface. The heart on the other side is moving and this opens up a possibility to discriminate between the signals from the heart tissue and the overlaying stationary reverberations.

In our approach, we examined the use of a simple temporal high pass filter to reduce reverberations. Such a filter should be implemented in a real time system, thus we were interested in a simple design. FIR and IIR filters were investigated.

It was difficult to formulate a reliable error criteria thus the results were assessed visually. Our evaluation gave that a 3 tap FIR filter is an appropriate choice to reduce reverberations while no degradation of the heart tissue was apparent.

INTRODUCTION

Ultrasound imaging is nowadays a wieldly used tool in medical diagnostics. However, the image quality can be still assessed to be mediocre and medical personal has to be trained to be able to interpret ultrasound images right. This is due to noise and degradation processes which are inherent to ultrasound imaging[1]. Echo signals add coherently introducing speckle patterns[2], phase aberrations degrade the ultrasound beam form and lead thus to blurring[3][4]. Reverberations or multiples will show up at locations where there is actually no target or overlay on target structures and thus influence their appearance. In this paper we look especially at cardiac ultrasound imaging where reverberations manifest themselves as a stationary fog overlaying on parts of the moving heart near to the transducer. The reverberations are due to the strong reflecting subcutaneous tissue layers reverberating either against themselves or against the transducer surface. This artifact is not clearly visible in still images, but annoying when studying animated image sequences. Our objective is to reduce this artifact.
The outline of the rest of the paper as follows. First, the methods used to reduce reverberations are given. We then present the results of the filter approach and finally sum up with conclusions.

METHODS

We acquired several ultrasound image sequences of the heart on a commercially available imaging system. The sequences were generated using a $3.25 \, MHz$ probe with a bandwidth of $\approx 1.5 \, MHz$. The ultrasound beam was scanned mechanically over a sector. The frame rate was $r_{frame} = 39$ frames per second varying slightly with the number of beams per frame. The scanner system was modified at our lab to get access to linear analog radio-frequency (RF) signals which were then digitized by a 12bit analog to digital converter operating at $10 \, MHz$. The sequences differed in view direction (long axis, short axis) and were of varying quality i.e. more or less disturbed by reverberations and other acoustic noise.

Inspecting the data we saw in several occasions a foggy stationary reverberation signal overlaid on moving parts of the heart. Here, it should be possible to discriminate between these two signal components by temporal (along frames) filtering. A simple high pass filter will suppress the stationary component while retaining the fluctuating signal from the moving heart. To determine the cutoff frequency of such a filter we analysed the frequency spectra of the different signal types. Typical examples are shown figure 1. Plot (a) is the spectrum of a stationary reverberation pixel whereas plot (b) stems from a pixel lying most of the time within the ventricle but which is passed by the moving heart wall. The magnitude of the two spectra are normalized to $0 \, dB$ and frequency at the abscissa is normalized to the frame rate, $39 Hz$.

\begin{figure}
\centering
\includegraphics{figure1.png}
\caption{Frequency spectra from a stationary pixel signal (a) and from a pixel passed by a moving target (b).}
\end{figure}

Note that there is a stationary component in spectrum (b) which is supposed to come from a reverberation, too.
We see from the spectra that a cutoff frequency of $f_c = 0.05$ seems to be an appropriate choice for the purpose to reject stationary reverberations. However, to determine a proper transition bandwidth is more involved. A steep transition band is desired in order to affect the higher frequency components in the signals as little as possible. This will result in a long impulse response which again can lead to an undesired artifact in the case of wide band events e.g. if we have a pixel where the moving wall enters just for one time sample. In this case the impulse response of the temporal filter will be imaged. The RF-signals are commonly amplitude range compressed by a logarithmic amplifier which increases temporal blurring, especially, if the impulse response is long. Thus, using a filter with an impulse response as short as possible suggests itself. Further, apart from the frequency response specifications, we are interested in a simple design (few coefficients) in order to implement the filter in real time at low costs.

To get an appropriate trade-off between all these constraints, we decided to assess the filtering results visually rather than to formulate a criteria which could be optimized mathematically.

RESULTS

We compared equiripple FIR filters and Butterworth filters having a relative cutoff frequency, $f_c = 0.05$ and a low filter order. The IIR filters performed comparable to the FIR filters in terms of stationary signal suppression but suffered somewhat from blurring due to their long impulse response. Further, for the FIR filters, we did not see any advantage of using a filter order higher than 2, although we get a very smooth transition band. Figure 2 shows a part of a frame with (right) and without (left) temporal filtering with a very simple 3-tap FIR filter. Even though the filter effect is more apparent when studying an animated sequence, we can see the reduced reverberation signal level in the center of the image to the right.

Figure 2: Transthoracal image of the apex of the heart before (left) and after (right) reverberation reduction.
The high pass filter was applied on all pixels without any differentiation where they are located in the image. Consequently, the stationary signals from the subcutaneous tissue layers (at the top of the image) are also suppressed, although they are not reverberations. This is however assessed to be no drawback in cardiac imaging because the strong signals in the near field of the transducer tend to distract the observer from the actual object, the beating heart.

In some cases high frequency noise was observed in the lateral direction within a frame in region of the stationary tissue layers. We assume that this is due to saturation effects in the analog signal.

CONCLUSIONS

In this paper we addressed the problem of stationary reverberations overlaying on the first order echoes from the moving heart. The disturbing artifact was reduced by high pass filtering in the temporal direction. From our experiments it seems that a 3-tap FIR filter is an appropriate choice. It turns out to be a good trade-off between the constraints of low-frequency (stationary reverberation) rejection, a short impulse response (little blurring) and simple implementation (low costs). However, due to its rather broad transition band there is the risk that slow moving targets will be suppressed, as well. So, it might be worthwhile to try an adaptive scheme where pixels are classified whether they stem from reverberations or first order echoes. This will be a topic of future research.

References


ACCURACY, LIMITS AND TRADEOFFS IN ULTRASOUND DISPLACEMENT ESTIMATES AND THEIR APPLICATION TO TISSUE ELASTICITY ASSESSMENT

William D. O'Brien, Jr., Eric J. Chien, W. Kenneth Jenkins, Jan Novakofski*
Dep. Elect. and Computer Eng., Univ. Illinois, Urbana, IL, USA
* Dep. Animal Sciences, Univ. of Illinois, Urbana, IL, USA

SUMMARY

Reliable ultrasound elasticity assessments rely on the accurate and precise estimates of tissue displacement. The accuracy of displacement data directly influences the accuracy of reconstructed elasticity fields, since elasticity data are usually constructed by processing tissue displacement data. The accuracy of displacement estimates, using conventional speckle tracking, is limited by the Cramer Rao Bound and, therefore, depends on many factors such as deformation size and type, window/kernel sizes and transducer frequency. Selection of specific values can result in tradeoffs between displacement accuracy parameters and fundamental elasticity imaging parameters such as strain resolution and strain SNR. In this paper, we discuss some of the important parameters and relationships between accuracy, limits and tradeoffs in ultrasound displacement and elasticity determination.

INTRODUCTION

In many medical and nonmedical ultrasound imaging applications, data are computed from estimates of tissue displacement. An often overlooked aspect of the process is the importance for the user to understand to what degree the data are reliable. Evaluations are only as good as the data on which they are based. Information derived from tissue displacement estimations can be used to assess tissue elastic properties which have tremendous potential applications. In food science, objective measurements of food mechanical properties are of interest in maintaining proper food quality control and in assessing the viability of new foodstuffs. Medical applications include the early detection of breast and prostate cancers and liver cirrhosis; diseases which are believed to significantly alter tissue elastic properties. The potential of imaging tissue elastic properties is enormous considering the prospective size of consumer and patient populations, the importance placed on quality and cost effective health care, and the enormous size of food industry and health care markets. Therefore, the need for a full understanding of accuracy issues and limitations in tissue displacement estimates has implications for many potential applications.

The accuracy of ultrasound displacement estimates depends on many factors including deformation size and type, window size, transducer frequency and tissue type. In elasticity imaging, the accuracy of displacement estimates also affects the accuracy of reconstructed elasticity fields since elasticity fields are directly computed from displacement data. The highest accuracy that can be achieved in displacement measurements using conventional speckle tracking is limited by the Cramer Rao bound. In addition, the digital sampling frequency limits the time resolution in a backscattered ultrasound echo and ultimately the smallest detectable displacement. The sampling frequency and tissue element size also limit the smallest detectable strain in elasticity imaging. Selection of deformation size, window/kernel sizes, transducer frequency/bandwidth and sampling frequency in displacement measurements can result in tradeoffs between displacement accuracy and fundamental elasticity imaging parameters such as strain resolution, strain dynamic range and strain SNR. In this paper, we first discuss the influence of several important imaging parameters on the
accuracy of displacement estimates. Next, the effects of window sizes, digital sampling frequency and strain level on limiting displacement accuracy, strain resolution and strain SNR are discussed. Finally, several possible tradeoffs in ultrasound elasticity estimates are discussed and related to accuracy issues.

ULTRASOUND DISPLACEMENT ESTIMATES

Accuracy. The ability to estimate accurately and precisely ultrasound-derived displacement depends on several important factors. (1) Deformation (magnitude). Deformations may consist of tissue translations, rotations or compressions (Bertrand et al., 1989). Displacement estimates have been reported to deteriorate with all three types of increasing deformations. Displacement accuracies were much worse for the case of rotations and compressions compared to translations using conventional speckle tracking (Chen et al., 1991). (2) Deformation (direction). Ramamurthy and Trahey (1991) have reported an improved accuracy for measurements taken in the axial direction over the lateral direction. (3) Window/kernel size. Ultrasonic displacement estimates deteriorate with decreasing region of interest (ROI) size in two-dimensional tissue tracking (Chen et al., 1991; Ramamurthy and Trahey, 1991) as has been reported for one-dimensional blood flow measurements (Foster et al., 1990). (4) Tissue type. The accuracy of ultrasonic displacement estimates was found to be better in porcine muscle samples than in porcine liver samples wherein muscle provided more structure in the ultrasound image than did liver (Chen et al., 1994b; Chen et al., 1995a). (5) Transducer frequency. Displacement accuracy will improve with increasing ultrasonic frequency of interrogation (Foster et al., 1990). This result stems from the increase in spatial resolution which is traded off against a decrease in depth of penetration. (6) Sampling frequency. Simulation results suggested that by sampling at four to five times the fundamental signal frequency, noise becomes the limiting factor with regard to accuracy (Parilla et al., 1991). (7) Signal type. There is substantial improvement in estimating displacement both axially and laterally when using RF data instead of envelope detected signals (Ramamurthy and Trahey, 1991).

Limits. Conventional displacement estimates using speckle tracking usually consist of applying cross-correlation or other window matching methods (Fourier methods have also been reported) to compute displacements (Chen et al., 1995a). The precision of cross-correlation time shift estimation can be characterized by the mean square error of the time shift estimate. The precision of displacement estimates is generally inversely proportional to the ultrasound signal bandwidth, the window size and the signal-to-noise ratio (SNR) of the echo signal (Bendat and Pierson, 1986). Ultimate limits on the accuracy of displacement estimates have been investigated using the Cramer Rao Lower Bound (CRLB) under a variety of conditions (Ophir et al., 1994; Walker and Trahey, 1995).

The digital sampling frequency ultimately limits the time resolution of ultrasound echo data. Sampling at or above the Nyquist rate can make SNR the limiting factor (Parilla et al., 1991). The accuracy of displacement estimates determined from mean square error can also be limited by low bit quantization (Trahey et al., 1988). In tissue elasticity measurements, a very high time resolution may be required in order to achieve reasonable strain estimates (Ophir et al., 1994). Sample interpolation methods have been applied to obtain time shifts less than the sample period. The interpolation method (parabolic, sinc, quadratic, cubic) can contribute to time shift errors (Jong et al., 1991).

In elasticity estimates, determination of the appropriate strain level is crucial. High strain levels may be required in order to achieve sufficient strain SNR (O'Donnell et al., 1994). However, the strain level directly affects the degree of signal distortion between pre-compression and post-compression signals. Elasticity estimates based on correlation methods are accurate only to the extent that segments of the pre- and post-compression signals can be considered as linear translations of one another. As previously discussed, high strains can result in significant signal decorrelation and limit the accuracy of displacement measurements (Ophir et al., 1994; Chen et al., 1995b).
**Tradeoffs.** The ability to make accurate and precise ultrasound displacement and elasticity estimates can result in several tradeoffs. (1) **Time shift estimation error vs. window/kernel dimensions** and (2) **Time shift estimation error vs. bandwidth vs. SNR.** The CRLB indicates there is a tradeoff between time shift estimated error in inversely proportional to bandwidth and window size. (3) **Strain resolution vs. window/kernel dimensions.** Ophir *et al.* (1991) have noted a classic tradeoff between window size and resolution in elastography measurements. Larger window sizes can improve displacement accuracy but there is a resulting loss in spatial resolution (Skovoroda *et al.*, 1994; Chen *et al.*, 1994a). (4) **Strain dynamic range vs. minimum detectable time shift.** The elastography type measurements the strain dynamic range is directly dependent on the minimum time shift estimate. The strain range will typically be in increments of \((c/2)(t_{\text{max}}/L)\) where \(t_{\text{max}}\) and \(L\) represent the minimum time shift and pre-compression length of a tissue element, respectively, and \(c\) is the tissue speed of sound. (5) **Multiple incremental compressions vs. single large compression.** Since large strains can improve strain SNR, the use of multiple incremental compressions to achieve a large net strain has been employed in elasticity measurements of gel based phantoms and muscle tissue (Skovoroda *et al.*, 1994; Chen *et al.*, 1994c; Chen *et al.*, 1995b). Multiple incremental compressions attempt to avoid problems with signal distortions and decorrelation that can result from a single large compression. We have previously reported that displacement measurements made over small deformations (translations) and large number of increments showed significant improvement in accuracy vs. single large deformation (Chen *et al.*, 1994b). (6) **Depth dependent stress vs. compressor size.** Cespedes and Ophir (1992) have reported stress field non-uniformities which are strongly dependent on compressor radius. Stress non-uniformities can be largely eliminated in sample measurements if the compressor size is larger than the sample surface dimensions (Ophir *et al.*, 1991; Chen *et al.*, 1994c). However, the depth dependent stress problem is complicated since accurate one-dimensional elasticity measurements may require a minimum height-width aspect ratio in order to meet the requirement of a bar sample for true one-dimensional mechanics measurements. In addition, the stress field estimations from simple (circular) compressors (Cespedes and Ophir, 1992) assume a homogeneous media. At present, there are no convenient methods for estimating internal stresses, although Ophir *et al.* (1991) have introduced the concept of the stress meter, where the stress in a layer of tissue of known elastic modulus is computed and used to reconstruct internal stress fields. However, the accuracy of this method still needs to be tested. (7) **Linear vs. non-linear elastic behavior.** Current static and quasi-static elasticity measurements have been largely based on the assumption of linear elastic behavior of tissues. In much of the present work, deformations have been limited to low strains in order to maintain linear behavior. As previously discussed, high strains may be required to obtain reasonable SNR. Using two independent elasticity measures Chen *et al.* (1994c) have observed a strain hardening effect in soft tissue, with non-linear behavior and permanent plastic deformation of tissue at high strains. It should be understood that for high strains, elasticity measurements are representative of only tissue pseudo-elastic properties (elasticity at a specific strain level) and not true tissue material properties. Current elasticity measurements can also be limited by problems due to stress relaxation, creep and hysteresis effects in repeated compression measurements (Chen *et al.*, 1995b).

**CONCLUSIONS**

Reliable ultrasound elasticity assessments rely on the accurate and precise estimates of tissue displacement. Accuracy, limits and tradeoff concerns must all be taken into account.

**ACKNOWLEDGMENTS**

The authors acknowledge the funding: US National Institutes of Health grants CA09067 and HD21687, the US Department of Agriculture and the National Live Stock and Meat Board.
REFERENCES


EVALUATION OF PHASE ABERRATION CORRECTION DUE TO DIFFRACTION OF THE ABERRATED WAVEFRONT; A SIMULATION STUDY

Lars Odegaard
Department of Biomedical Engineering, University of Trondheim, Norway
Dep. Mathematical Sciences, the Norwegian Institute of Technology, Trondheim

SUMMARY
When a wavefront propagate through human tissue phase and amplitude aberrations are generated. This is a problem in medical ultrasound imaging. The distortions caused by aberrations typical for the abdominal wall are demonstrated by simulations. The simulations indicate that correlated aberrations give larger distortions than uncorrelated aberrations. Simulations to show how a rough wavefront change by propagation is presented. It is observed that a phase front is not smoothed by propagation, but the shape is changed. It is also demonstrated that phase aberrations generate amplitude aberrations that increase by propagation and by increased frequencies. These observations are used to discuss correction methods presented in the literature. It is demonstrated that amplitude aberrations are a limiting factor for phase aberration correction, especially when the phase aberrations are generated far away from the transducer and for high frequencies. The simulations show that amplitude information from receive can be used to reduce both phase and amplitude aberrations even when the aberrations are generated far away from the transducer.

INTRODUCTION
The last years scientists have studied size, consequences and possible methods to correct phase aberrations in medical ultrasound. Recently amplitude aberrations are also studied. In this paper, different correction methods are evaluated due to a rough boundary between media with different speed of sound, for example fat and muscle. The distortions caused by propagation through a rough boundary can be modelled by adding a phase screen to the original wavefront. The case to study is then a transducer in a homogeneous medium and a phase screen some distance away from the transducer.

METHOD
Aberrations are simulated by an AR1 process:

\[ x(k) = a x(k-1) + w(k), w(k) \sim N(0, \sigma^2) \]  

(1)

The phase, \( x(k) \), in position \( k \) is given as a weighted sum of the phase in position \( k-1 \) and a random term \( w(k) \). The correlation structure in an AR1 model is described by the autocorrelation function \( \rho(m) \).

\[ \rho(m) = \frac{R(m)}{R(0)} = a^m, R(m) = \frac{a^m}{1 - a^2} \sigma^2 \]  

(2)

The correlation length is defined as the distance where \( \rho(m) > 0.5 \). To simulate a phase screen with specified correlation length we set the parameter \( a \) to:

\[ a = \left( \frac{1}{2} \right)^{\frac{\text{corr. length}}{\text{discretization step}}} \]

(3)

The standard deviation (std) of the total phase front is \( \sqrt{R(0)} \). The parameter \( \sigma \) is then given by:

\[ \sigma = \sqrt{\text{tot. standard deviation}} \sqrt{1 - a^2} \]

(4)

Figure 1: Phase and amplitude aberrations typical for human abdominal wall simulated with an AR1 model.

To simulate realistic phase aberrations for the human abdominal wall the parameters in the AR1 model is based on measurements from Hinkelman et al. [1]. The correlation length and rms value for typical phase aberrations were 7.9 mm and 43.0 ns. The
values for typical amplitude aberrations were 2.3 mm and 3.3 dB.

To simulate how a phase aberrated wavefront change by propagation the Rayleigh integral is used [2]. The phase front is divided into points where each point generate spherical waves with amplitude and phase given by the wavefront.

RESULTS

First we simulate how aberrations typical for the abdominal wall will influence on a beamprofile. The simulations are done for a 128 elements 50 mm linear transducer electronic focused to 100 mm. The pulse has 3.5 MHz center frequency and is three wave lengths long. Typical phase and amplitude aberrations, fig. 1, are introduced on the transducer. The calculated beamprofile 100 mm away from the transducer is presented in fig. 2. The phase and amplitude aberrations make the mainlobe broader and increase the sidelobe level. If we have amplitude aberrations only, i.e. do perfect phase aberration correction, the beamprofile is improved, but it is not as good as in the undisturbed case.

![Figure 2: Distortions in the beamprofile due to aberrations typical for the abdominal wall.](image)

Simulations with uncorrelated aberrations is presented in fig. 3. Note that the mainlobe is less disturbed and the sidelobes are constant. How uncorrelated amplitude aberrations influence on the sidelobe level is also studied by Zhu and Steinberg [3].

Phase aberration correction methods seems to work best when the aberrating layer is close to the transducer. To study why, phase and amplitude of an aberrated wavefront is simulated for increasing distances to the phase screen. The change in the phase front is shown in fig. 4. Note that the standard deviation and the correlation length of the phasefront are unchanged by propagation, i.e. the phasefront is not smoothed by propagation. The shape is changed.

![Figure 3: Distortions in the beamprofile due to aberrations with size typical for the abdominal wall but with correlation length=0.](image)

The amplitude distortions of the wavefront are shown in fig. 5. We observe that size and correlation length of the aberrations increase by propagation away from the phase screen. Other simulations show that amplitude aberrations increase if we use higher frequency.

![Figure 4: Phase aberrations [as] in the wavefront when a plane wave propagate through a phase screen at z=0. The figure shows the phasefront at increasing distances to the phase screen, f=3.5 MHz, CW.](image)

To study different correction methods it is used a 500 elements 100 mm linear transducer. The signal is a 3.5MHz continuous wave. The goal is to compensate for a rough boundary 50 mm from the transducer. The phase screen to model the effect of the rough boundary are shown in fig. 6. The first simula-
tion is for a plane wave crossing the phase screen and propagate 50 mm before arriving the linear transducer. The phasefront received on the transducer had $\text{std} = 33 \text{ns}$ and $\text{corr.length} = 7.9 \text{mm}$. The amplitude of the received wavefront had $\text{std/mean} = 0.33$ and $\text{corr.length} = 1.7 \text{mm}$. The effect of similar aberrations are demonstrated in fig. 2.

- human abdominal wall aberrations, — no aberrations
  
  $z=0 \text{mm}, \text{std/mean}=0, \text{corr}=\text{inf}$
  
  $z=5 \text{mm}, \text{std/mean}=0.1519, \text{corr}=0.9 \text{mm}$
  
  $z=10 \text{mm}, \text{std/mean}=0.2195, \text{corr}=0.9 \text{mm}$
  
  $z=15 \text{mm}, \text{std/mean}=0.2907, \text{corr}=1.3 \text{mm}$
  
  $z=20 \text{mm}, \text{std/mean}=0.3666, \text{corr}=1.3 \text{mm}$
  
  $z=25 \text{mm}, \text{std/mean}=0.3913, \text{corr}=1.7 \text{mm}$
  
  $z=30 \text{mm}, \text{std/mean}=0.3811, \text{corr}=2.5 \text{mm}$
  
  The wavefront from $-50 \text{ mm}$ to $50 \text{ mm}$

![Figure 6: Phase screen 50 mm from the transducer](image)

We want to use the wave observed on the transducer to do modifications on the next transmit pulse. We want the transmitted wavefront to be plane after crossing the phase screen. The simplest idea are to use electronic delays equal but with opposite sign of the phase screen. This will work only in the nearfield because an aberrated wavefront change its shape by propagation. Another possibility is to use the opposite of the aberrated phase front observed on the transducer. Fig. 7 show the wavefront after propagating from the transducer through a 50 mm homogeneous medium, and after crossing the phase screen. For a perfect correction we will expect no phase or amplitude aberrations in this wavefront. Unfortunately the standard deviation of the phase aberrations is still as high as 13.5 ns and the $\text{std/mean}$ for the amplitude aberrations as high as 0.21. This method, can be viewed as optimal phase aberration correction, is not optimal.

A third possibility is to use the amplitude information in the received signal. Phase correction is still the opposite of the aberrated phasefront observed at receive, but we also use the amplitude in the received signal as an apodization at transmit. The observed wavefront observed directly after crossing the phase

![Figure 7: The opposite of the aberrated phasefront observed on receive is used as phase correction for next transmit pulse. It is no apodization or amp. correction. The figure shows the wavefront after travelling from the transducer at $z=0$, through a 50 mm homogeneous medium, and after crossing the phase screen at $z=50 \text{ mm}$](image)

We want to use the wave observed on the transducer to do modifications on the next transmit pulse. We want the transmitted wavefront to be plane after crossing the phase screen. The simplest idea are to use electronic delays equal but with opposite sign of the phase screen. This will work only in the nearfield because an aberrated wavefront change its shape by propagation. Another possibility is to use the opposite of the aberrated phase front observed on the transducer. Fig. 7 show the wavefront after propagating from the transducer through a 50 mm homogeneous medium, and after crossing the phase screen. For a perfect correction we will expect no phase or amplitude aberrations in this wavefront. Unfortunately the standard deviation of the phase aberrations is still as high as 13.5 ns and the $\text{std/mean}$ for the amplitude aberrations as high as 0.21. This method, can be viewed as optimal phase aberration correction, is not optimal.

A third possibility is to use the amplitude information in the received signal. Phase correction is still the opposite of the aberrated phasefront observed at receive, but we also use the amplitude in the received signal as an apodization at transmit. The observed wavefront observed directly after crossing the phase

![Figure 8: The opposite of the aberrated phasefront observed on receive is used as phase correction for next transmit pulse. The aberrated amplitude observed at receive is used as apodization. The figure shows the wavefront after travelling from the transducer at $z=0$, through a 50 mm homogeneous medium, and after crossing the phase screen at $z=50 \text{ mm}$](image)
screen is shown in fig. 8. The standard deviation of the phase aberrations is now reduced to 5.8 ns and the std/mean of the amplitude aberrations to 0.11. This method can be viewed as the time reversal mirror [7].

**Discussion/Conclusion**

The simulations show that both phase and amplitude aberrations disturb the beamprofile of the transducer. It is observed larger distortion for correlated aberrations than for uncorrelated.

The simulations show that the phase of an aberrated wavefront is not smoothed by propagation, but the phase is changed. It is then possible to do phase aberration correction at receive. Methods as maximizing the correlation between the signal in the different elements may be used [4]. It is also possible to add delay to the different elements until the intensity of an area of interest in the image is maximized [6].

A phase screen make only phase aberrations, but amplitude distortions are developed as the aberrated phasefront continues to propagate. The simulations show increased amplitude distortions by propagation. This factor limits the effect of phase aberration correction for large distances between the aberrating boundary and the transducer. Other simulations show increased amplitude aberrations by using higher frequencies. Amplitude distortions at receive can be reduced by backpropagation [6]. The wavefront is backpropagated to the phase screen where the phase corrections are done. The amplitude aberrations will then be reduced to the size in fig. 8. Note also that the max speckle intensity criterion [5] may be used for amplitude corrections too, because the criterion optimize the beamprofile.

To do optimal corrections, we have to correct at transmit too. One possibility is to use the opposite of the phase aberrations observed at receive as a delay correction for the next transmit pulse. This give better results than no corrections for a phase screen 50 mm from the transducer. This phase correction only technique work best for small distances between an aberrating boundary and the transducer, because the amplitude aberrations increase by propagation. An improvement of this phase correction method is to include the amplitude observed at receive as apodisation for the next transmit pulse. This give good results even when the phase aberrations are generated far away from the transducer. A plane wave is studied here, but for a point reflector the last method is the same as Fink's time reversal method [7]. Since the simulations is done with continuous waves it is not necessary to take care of the waveform. A sum of continuous waves with different phase and amplitude but equal frequency, will have the same frequency but a modified phase and amplitude. So for Fink's method used on continuous waves it is enough to modify the delay and apodization.

Backpropagation and time reversal mirror is based on the same idea. The size of the transducer is a limitation (100 mm). For perfect time reversal mirror we have to return all signals, not only they hitting the transducer. This explain why the wavefront is better corrected close to the center axis to the transducer than close to the edges of the transducer. Another limitations is that we need many small elements when the phasefront is that rough as hier. We must also avoid minimize the size between the elements so we can time reverse as much as possible of the received signal.

**References**


ULTRASOUND DOPPLER BLOOD FLOW MEASUREMENTS: SIGNAL MODELS, AND SIGNAL PROCESSING METHODS

Hans Torp
Dep. of Biomedical Engineering, University of Trondheim, Norway

SUMMARY
This is a review of medical ultrasound Doppler and color flow imaging, covering basic principles, signal models, and some of the recently developed signal processing methods, which has been made possible due to the latest developments in A/D- and DSP-technology. A stochastic model for the signal from blood is presented, and algorithms for resolving velocity ambiguity due to frequency aliasing are described. An example with ultrasound Doppler data from a human artery is shown, comparing a new algorithm for velocity spectrum analysis with the conventional fft-method.

INTRODUCTION
There are two methods used for ultrasound blood flow measurements. One is utilizing the transit-time of the ultrasound signal through the vessel [1], but this requires direct access to the vessel, and is therefore only possible during surgery. An alternative way of retrieving blood flow information non-invasively was developed during the 1960-70's, utilizing the Doppler shift of the backscattered signal. By coherent demodulation of the backscattered signal, the weak signal components from blood could be extracted by high-pass filtering, and quantitative velocity information estimated, using the Doppler equation. The first use of continuous wave Doppler (CW) to measure blood flow velocities was reported as early as 1957 [2], and pulsed wave Doppler (PW), which provided range resolution, was introduced ten years later [3].

By sampling the scattered signal at high sampling rate, blood velocity information from different spatial positions along the ultrasonic beam can be retrieved simultaneously. The movement of the blood cells causes a shift in the delay of the received echoes from pulse to pulse which is detectable both in the phase and the amplitude of the received signal. The ultrasonic beam is swept over a sector either by mechanical motion, or electronically, by varying the time delay between the elements in the transducer array, to obtain a two-dimensional image.

It is convenient to perform quadrature demodulation of the signal before range gating (sampling). In this way, the phase information in each range gate is preserved, independent of the range gate density. At this level, the strong signal components from tissue structures appear as a low frequency signal in each channel. These components are suppressed in the wall motion filter, which in its simplest form is a bank of high-pass filters, acting separately on each range gate. From the signal samples, velocity information in each spatial point is retrieved, and presented in a 2D display. This process is repeated with a frame-rate about 10 to 30 times per second, which make real-time monitoring of the dynamic velocity field in the cardio-vascular system possible.

- 371 -
SIGNAL MODEL FOR SCATTERING FROM BLOOD

Human blood consists mainly of red blood cells in a volume concentration of 40-50 %, and blood plasma. The red cells have a mean diameter of 7 µm, which is much smaller than the wavelength used in medical ultrasound (λ = 0.1 - 1.0 mm). Even though red cells tend to aggregate into larger particles, the conditions for Raleigh scattering is satisfied in most cases. Due to the high volume concentration, it is convenient to use a random continuum model for the scattering from blood[4]. The spatial fluctuations in mass density and compressibility, which determine the incoherent part of the scattering, are assumed to be proportional to the spatial fluctuation in the red blood cell concentration \( n_c(t) \). The fluctuation function is defined as the difference between the actual cell concentration and its local average. Thus \( n_c(t) \) is a zero mean random process in space and time. Assuming a short correlation in space for a fixed time \( t \), and neglecting diffusion, each small blood volume can be regarded as independent scatterers. The received signal is modeled as a Gaussian random process, due to the large number of independent scatterers that contribute to the signal. The power spectrum is obtained by adding the contribution from each small blood volume passing through the ultrasonic beam.

![Diagram of ultrasonic beam and signal](image)

**Fig. 1** Single scatterer echo response in depth range and time.

In fig. 1 the received signal from the small sub volume is indicated. The delay of the echo increases from pulse to pulse caused by the movement of the scatterer. By sampling the signal at a fixed delay, a Doppler shifted signal is obtained, with frequency proportional to the axial velocity component. By 2D Fourier-transform, the power spectrum of the signal is obtained. This is illustrated in fig. 2, showing how the different frequency components in the transmitted signal (horizontal axis) give a Doppler shift frequency (vertical axis) proportional to the transmitted
frequency, and the velocity of the scatterers. The 2D spectrum characterizes the signal as a 2D Gaussian process [5].

![Doppler shift](image)

**Fig. 2.** Two dimensional power spectrum of the signal from blood with uniform velocity field

**SIGNAL PROCESSING ALGORITHMS**

Full digital processing of ultrasound Doppler blood flow signals has recently been made possible, due to the development of high speed and precision analog-to-digital converters. With programmable digital signal processing components it is possible to implement a wide range of algorithms operating directly on the ultrasound RF-data in real time. Simultaneous processing of the data from several samples along the ultrasonic beam makes it possible to track the movement of the blood scatterers when the velocity is higher than the Nyquist limit. This can be utilized to resolve velocity ambiguity for both 2D color Doppler and spectral Doppler. A number of different methods applicable for 2D color Doppler have been proposed, based on direct measurement of time-shift [6], as well as Doppler shift with several frequencies [7, 8].

A related method can be applied to pulsed wave Doppler to suppress frequency aliasing [9]. By simultaneous processing of several signal samples along the ultrasonic beam, the 2D frequency information can be utilized to suppress frequency aliasing. The resulting velocity/time spectral display shows a more clearly defined spectral envelope of the maximum velocity than with conventional methods based on the discrete Fourier transform. This makes it possible to delineate velocity wave-forms with peak velocity up to several times the Nyquist limit. Fig. 3 shows an example from human subclavian artery, where the new method is compared to conventional spectrum analysis.

**CONCLUSIONS**

The signal from moving blood can be modeled as a two-dimensional Gaussian process. Based on a priori knowledge about the form of the two-dimensional power spectrum, algorithms can be
deduced, which is capable of resolving velocity ambiguity caused by frequency aliasing. This is important for the measurement of high velocities occurring in arterial stenosis, and heart valve leakage.

Fig. 7. Velocity spectrum analysis based on 2D data (right) compared to conventional spectrum analysis (left) applied to subclavian artery blood flow with peak velocity at 4.5 times the Nyquist limit. Velocity labels are in [cm/sec].

REFERENCES

APPLICATION OF COMMON FAN ULTRASONIC DIAGNOSTIC APPARATUS IN OCULAR DISEASES

Zhang Yu Xia, Zhou jian Hua
Air Force Hospital of Urumqi, China

SUMMARY This article describes the ophthalmopathy diagnosis for 8870 cases of patients (7750 eyes) from September 1980 to August 1984 utilizing general purpose cardio/vascular ultrasound imaging system instead of proprietary imaging system designed exclusively for eye ailments. Clinical application examples for various eye diseases are provided to illustrate the effectiveness and feasibility of this practice. Accuracy and precision of diagnosis and measurements are evaluated and compared with those of general purpose cardio vascular sonography instrument. Pros and cons of this attempt are analyzed objectively. The conclusion is that ophthalmopathy diagnosis with general purpose imaging system is viable in those hospitals which still don't have installed proprietary imaging system. The work described in this article actually laid a foundation for future application of general purpose imaging system for ophthalmopathy therapy.

In China's hospitals, application specific high frequency ultrasound imaging system with linear array transducer designed exclusively for ophthalmopathy diagnosis are being utilized. However, from September 1980 to August 1984, colleagues in China Air Force Hospital examined 5750 cases of eye ailments (total number of eyes was 7750) utilizing general purpose cardio/vascular ultrasound imaging system instead and got satisfactory results. With this report, we would like to share our experience with others.

Instrument and Methodology
Instrument: A HP sonos 100 cardio/vascular imaging system was used for diagnostic purpose. The transducer, which transmitted and received signals with frequency of 7.3 MHz, was equipped with specially designed water seal adapter. System features included PW/CW Doppler and capabilities such as frozen frame, image processing, split screen display, image zooming, data storage and retrieve, automation measurement and imaging photographing etc.

Method: During examination, let the patient be lying flat on his desk with eyes closed. Apply an appropriate amount of coupling on the eyelids. Firstly, scanned horizontally and vertically for eye-ball axis tomography, then rotated the transducer left and right for multi-orientation scanning for global examination. Instructed the patient to practice ocular ography to facilitate observation of the relationship between pathology and eye-ball tissues. Simultaneously, imaging system processed the received echo images, performed measurements and recorded. Doctors settled for diagnosis with the reference of patient's clinical information.

Information and Results
The age of the patients who had undergone ultrasound diagnosis ranged from 2 months to 84. Total eyes examined were 7750, with the help of optical examination, medication treatment, CT, pathological study and surgery, we have been able to confirm that the overall accuracy of ultrasound diagnosis for ophthalmopathy was 83.3%. Even better results would be expected from specific diagnosis of.

Retinal detachment—97.4%

- 375 -
Summary:
Retinocytoma——98.3%
Turbid vitreum——99%
Dislocation of lens——98%

Discussion:
A. we characterized the main features of ultrasound echo signals for various eye diseases' diagnosis made use of general purpose cardiovascular imaging system as follows:
(1) Lens crystalline disease:
1. Cataract:
Compared to those of normal crystalline, the echo image of front and back sac of crystalline showed significantly intensified and thickened. However, the signal intensity varied and displayed as a short light band. When come to maturity, the image of nucleus part shaped as a white sphere. Most of the senile change manifested as intensified nucleus image and back sac thickening, while traumatic cataract as front sac thickening.
2. Lens dislocation:
Ringlet light band or disk-shaped low echo signal appeared in the vitreous image dark area with clear boundaries, moved along with the eyebulb's movement. For the case of half-detachment, only shift of ringlet echo signal and disk-shaped signal from normal location and thickening were displayed.
3. No lens crystalline:
Echo signal from crystalline disappeared, or just the image of sac wall was displayed.
4. Artificial crystalline:
There were attendant echo signals which followed crystalline's echoes.
5. Retinocytal fibropia:
There were crystalline echoes followed by horn-shaped or mitral echo signals with medium intensity.
(2) Vitreous humor disease:
1. Acute hemorrhage:
There were plenty of floating point echoes in the vitreous image dark area, most of these signals precipitated on the eyebulb image area and shifted along with the body movement.
2. Turbid vitreum, hemorrhage organization:
Dispersed weak light points existed in the vitreous image dark area with flickered light blocks. When tunica were created as a result of hemorrhage, demarcated thin and bent light band appeared, the back movement showed positive. The light band was felt stiffer after organization, meanwhile, the back movement weakened.
(3) Retinal disease:
1. Detachment:
"Y", "Y", "Δ", or "-"-shaped echo light band existed in the vitreous image dark area. The band trembled along with the movement of eyebulb. The middle part of the band splited with the image of optical papilla. In the case of time-worn detachment, the light band got thickened and intensified. Most of the patient constricted disease of turbid vitreum. Secondary detachment was often followed by the image change of primary fooi.
2. Retinocytoma:
There were irregular image blocks with crisp boundaries in the vitreous image dark area. Signal intensity was medium to strong, and we were able to witness dispersed spots of calcification. back movement showed positive. Infants are most vulnerable to such disease.
(4) Grape tunica disease:
1. Choroid detachment:
A hemispherical light band of gibbous vitreous chamber appeared in the vitreous image dark area. Its front end spliced with the ciliary muscle image, and the back end with the image of eyebulb wall's equator part.

2. Choroid melanoma:
A semicircular mushroomed, solid and protruded light block with smooth surface appeared in the vitreous image dark area, on the base of which displayed an echoless dark area. Patients who contract this disease are vulnerable to secondary detachment.

(5) Orbit diseases: A variety of pathologies might be categorized to this kind of disease. Following is a list of most ones and their signal characteristics:

1. Chronic dacryocystitis.
   It is an intraorbital inflammation. The main characteristics of echo were expansion of lacrimal gland pit image, with clear boundary and moderately intensified signal amplitude and with fissures around its peripher. Typically, this disease occurred with attendant dacryorhea, swelling and sore. It is possible to examine satisfactorily with ultrasound imaging.

2. Cavernous angiosa.
   It might develop at any part of the orbit. The clinical symptoms were eyebulb protrusion and eyebulb movement blockage. Ultrasound image characterized circular or irregular low intensity light block, with clear boundary and smooth surface. Echoes from internal tissue appeared homogeneous.

3. Lacrimal gland mixed tumor.
   It is essentially a lacrimal gland adenomyoepithelium. The ultrasound echo characterized an elliptical, solid light block with clear boundary on top of the outer orbit. Echoes from internal tissue was moderate, evenly dispersed, non-compressed. The eyebulb was enforced to deform.

4. Orbit cystitis.
   This disease’s symptom categorized as simple cysts or hydatid disease. Their echo image featured apparent tumoricysticofuscular dark area in orbit or on the eyebulb's back end. you might have a feeling of compression. For hydatid disease, the patients typically developed thicker cystic wall. Occasionally, there were partitions in cystic or subcystic. This category of patients usually had the history of some sort of epidemic disease.

(6) Trauma: It is usually accompanied by turbid dioptral interstitial tissue. Therefore, ultrasonography played an important part in diagnosis and treatment of eye trauma. Typical trauma was corpus alienum within eyes, images often displayed strong echo, isolated light points or blocks accompanied by lateral or back cast tail. Through dynamic observation, you could confirm the relationship between the alienum and eyebulb's wall:

1. Eyebulb rupture:
   It is the rupture on the front and back end of eyebulb wall typically results from heavy blow on the eyebulb. you might witness the destruction of normal echo, associated with healthy eyebulb, otherwise smooth and clean light are echo signals were disrupted, echoless or low echo dark area around it appeared. There were irregular weak fascia in the vitreous. Even stronger echo, echoless or weak echo are with irregular, shapes and blurred boundaries dropped up around eyebulb.

2. Eyebulb contraction
   It is the eyebulb degradation caused by destruction of eye's tissue structure due to serious trauma or other eye disorders. On the display, you might aware of the change of eyebulb diameter to less
than normal with random shape, also of rough echo from eyeball wall, thickening of echo light arc, contraction of vitreous dark area and uneven intensity of echo signal from internal structures etc.

B. Pros and Cons

It is evident that general purpose ultrasound imaging system operating in B mode with 7.5 MHz transducer could play same important part as the specially designed ophthalmopathy imaging system in clinical applications. In the practice, it was found that thanks to dynamic focusing features provided by the 7.5 MHz transducer, high resolution image and panoramic view of eyeball and orbit were able to be displayed. Transducer tip soft membrane of wart zero adapter which was applied with oculoplast allowed application flexibility and non-invasive eyeball probing.

We could make use of system's B mode imaging to measure automatically eye axis length, depth of the anterior chamber and thickness of crystalline. Capabilities such as zooming in on the local lesions of eyes, Doppler detection and image processing were readily available. There were no apparent disparities between the outcome of clinical applications utilizing general purpose or proprietary imaging system. No difference existed either between the diagnosis results of eye diseases which could not be detected with ordinary optical instruments such as turbid dioptical interstitial tissue with both type of imaging system. However, there were certain disadvantages which should not be ignored on the part of general purpose imaging system employed in eye diseases' diagnosis. Among them, the most prominent was the poor diagnosis accuracy in measurement of eye axis length. It was found that the error could be as high as 100% and will negatively affect the accuracy of eye surgery. Nevertheless, general purpose imaging system is a viable replacement where no proprietary ophthalmopathy imaging system is available.

REFERENCES

COMPUTED SYNTHETIC APERTURE ULTRASOUND IMAGING: 
APPLICATION TO MEDICAL DIAGNOSIS 

Juha Ylitalo*, Torgrim Lie, Hans Torp 
Department of Biomedical Engineering, University of Trondheim, Norway 
* on leave from Dept. of Electrical Engineering, University of Oulu, Oulu Finland 

SUMMARY 
Ultrasonic imaging applying the synthetic aperture approach is widely used in NDE, in which the target 
stability can be guaranteed. In medical ultrasonics real-time scanning is required. Therefore, in modern 
clinical ultrasonic scanners beam focusing and steering has been achieved by an expensive custom-design 
hardware. Today, the rapid progress in microprocessor technology allows also the development of a low-
cost real-time ultrasonic scanner in which the image focusing is performed numerically. 

In this paper a fast ultrasonic synthetic aperture imaging method suitable for handheld scanners is 
described. The method employs one-dimensional holography in the spatial frequency domain. Due to the 
efficient algorithm numerical focusing of a 128x128 pixel image takes only about 1.5 s in an ordinary 
microcomputer (PC 80486/66). Image quality was demonstrated by imaging a standard tissue-mimicking 
phantom. In experiments 16 elements of a 3.2 MHz array were used as a single transducer which 
resulted in a -12 dB ultrasonic beam width of 17° in the scanning direction. Data were recorded by 
mechanical scanning in a monostatic manner which allows real-time data acquisition if a multielement 
probe is used. Numerically focused images of the phantom show that both the lateral and the range 
resolution were about 1 mm in a depth range of 10 mm-160 mm. In addition, dynamic resolution was 
adequate to resolve small cysts (diameter 4 mm) in a depth of 130 mm in this phantom. In-vivo 
experiment and the application of the current method for curved apertures are also discussed. 

INTRODUCTION 
Acoustical holography was intensively studied in 1960’s and 1970’s. A practical problem with the early 
methods was the requirement for two-dimensional scanning and optical image reconstruction. Even if 
numerical image reconstruction of detected ultrasound plane holograms was introduced soon the 
technique of ultrasonic holography was not suitable for medical imaging since it required two-
dimensional data acquisition [1,2]. From clinical point of view such methods that employ one-
dimensional scanning of the transducer and display a cross section orthogonal to the accessible surface 
are more attractive. Single-line acoustic holography can be applied in this way but it has, however, the 
disadvantage that the longitudinal resolution is even ten times worse than the lateral resolution [3]. 
This drawback can be overcome by combining the pulse-echo technique (B-scan) and the principle of 
holography [4,5]. 

This paper describes a practical synthetic aperture ultrasound imaging method in which one-dimensional 
holography is combined with the ordinary pulse-echo technique. Since only one-dimensional processing 
is needed and it takes place in the spatial frequency domain the numerical focusing of the hologram can 
be performed efficiently in a microcomputer. Moreover, the method allows a simple and fast data 
acquisition scheme in which a target is scanned with an unfocused wide-angle beam. If a multielement 
probe is used the method is feasible for a low-cost real-time clinical scanner.
Fig. 1. Principle of the computed synthetic aperture imaging method.

METHODS

Fig. 1 shows the principle of the present method. In data acquisition a probe with a wide-angle beam is scanned mechanically over the target area. The probe is used in a pulse-echo mode in which an ultrasound burst is transmitted and received in the same probe position on the synthetic aperture (monostatic approach). Wavefronts reflected from the target are detected as holograms so that both the amplitude and phase angle information are preserved. The length of ultrasound burst has to be short (typically 4 wavelengths) in order to achieve a good depth resolution. Due to linear scanning wavefronts from a point-source form a hyperbolic trace in the hologram data. This means that it is not possible to obtain the entire one-dimensional hologram from one depth level. The task of curvature compensation is to correct for the hyperbolic traces and rearrange the hologram data table. The result is that the hologram corresponding to a single point-source becomes a straight line in the hologram data table. Curve compensation can be most efficiently applied in the spatial frequency domain since correction is needed only for each depth level and not for each image pixel as is the case in time domain focusing. Compensation is based on the fact that the spatial frequencies (sin α / λ) represent incident angles of the wavefronts and thus also the correct depth levels where the echofronts originate [5]. After correction for the curvature distortion the numerical focusing follows the ordinary wavefront backward propagation method [2]. Since one-dimensional processing is applied in the spatial Fourier domain the reconstruction algorithm allows an extensive use of FFT-routines and also parallel processing. Therefore, the algorithm is feasible for a real-time system.

It is interesting to note that the monostatic approach in data acquisition improves the lateral resolution by a factor of two compared to systems that employ the whole imaging aperture in a multistatic manner [6]. In practice, the angular aperture size of the probe sets a limit for the lateral resolution. Accordingly, the limit is λ/4 tan(α/2), where λ is the wavelength, and α is the angular aperture size of the ultrasonic beam. The beamwidth of 15° gives, approximately, the lateral resolution of 2λ [7]. In addition to improved lateral resolution, a simple scanning scheme is obtained which makes the design of a simple data acquisition unit possible.

In this work, a standard tissue-mimicking phantom was used as an object. The phantom is made of medium which has an ultrasonic velocity of 1450 m/s and a mean attenuation of 0.5 dB/cm/MHz. Accordingly, the wavelength for the 3.2 MHz ultrasound frequency that was used in this study is 0.45 mm. Imaging apparatus consisted of a microcomputer, a 10-bit digital oscilloscope (Philips PM3323), an ultrasound transmitter/receiver unit, a stage motor unit and a 3.2 MHz linear array (Siemens). The probe was formed by connecting 16 transducer elements in parallel. The element width was 0.2 mm and so the
effective transducer aperture was about 3 mm. The elements were weighed according to the Blackman-Harris window which resulted in a beam width of $17^\circ$ in the image plane. The side-lobe level of the beam was about -50 dB. In the elevation plane, the ultrasonic beam was focused by a lens. At each position on the synthetic aperture the probe was excited by 2 cycles of 3.2 MHz, 60 V square wave. Measured rf-signal was sampled with 10 bits at the rate of 4 samples/wavelength and 512 samples were recorded for each probe position. From these samples, 128 complex values were detected. When imaging deep targets, 16 recordings were averaged at each measurement position to obtain acceptable signal-to-noise ratio. The probe was moved through 256 positions which lead to complex data of 64 kb (256*128).

RESULTS

The advantage of the computed synthetic aperture method is that uniform spatial resolution can be achieved in a large depth range. This is demonstrated in the following experiment where a tissue-mimicking phantom was as a target. Fig. 2 shows the image of 10 wires in the near-field ('dead-zone', 0-30 mm) of the phantom. The distance between the point targets is 5 mm in a lateral direction and 10 mm in a range direction. All of them are well resolved with good lateral and longitudinal resolution. The ringing of the transducer is seen on the top of the image. Fig. 3 depicts the spatial resolution in a depth range of 20-80 mm. In this case the target area has a resolution test target in which the axial spacings between the 6 wires are 5, 4, 3, 2, and 1 mm. Fig. 4 demonstrates that practically the same spatial resolution could be obtained also in the far-field (depth range 134-164 mm). In Fig. 4 side-lobes are present at the depth of the lateral row of point-targets which is due to interference. The side-lobe level could be reduced by increasing the sampling rate in range (time) direction. From Figs. 2, 3, and 4 it can be concluded that the spatial resolution of about 1 mm could be achieved independently of the depth range.

Fig. 2. Point-targets at depth of 0-30 mm.

Fig. 3. Point-targets at depth of 20-80 mm.

Fig. 4. Point-targets at depth of 134-164 mm.

Fig. 5. Cysts (φ4, φ6, φ8 mm) at depth of 130 mm.
Fig. 5 demonstrates the contrast resolution of the present imaging method. Three cysts with diameters of 4, 6, and 8 mm were imaged at a depth of 130 mm in the phantom. Clearly, all the cysts could be resolved. This result is encouraging since it is expected that in real-time scanning the contrast between the cysts and the background medium would be even better due to time-domain averaging and optimal positioning of the probe.

CONCLUSIONS

The work reported here demonstrates the feasibility of a computed synthetic aperture ultrasound imaging method for medical diagnosis. Good and practically uniform spatial resolution was achieved in a very large depth range (10-160 mm). Moreover, contrast resolution was adequate to resolve small cysts at a depth of 130 mm in a tissue-mimicking phantom. It is concluded that the method allows a simple and cost-effective implementation and is fast enough for real-time applications.

Next, it is important to carry out in-vivo experiments. These can be made with a modern digital scanner which allows a programmable control of the beamform and the scanning sequence of the linear array. Real-time data can be acquired, for example, by scanning the array aperture with a group of four transducer elements in parallel and using no focusing of the beam. If data are recorded as rf-waveforms image reconstruction can be performed off-line in a microcomputer.

The present method can be applied also for curved (concave) or enclosing circular apertures. It has been shown that good image quality can be achieved also for curved apertures with no significant increase in image acquisition time compared to the linear case [8]. Recent studies indicate that the method is also applicable for convex apertures. Therefore, potential areas of application include also cardiac and intravascular imaging.

ACKNOWLEDGEMENTS

The financial support from the Academy of Finland and the Nordic Academy for Advanced Study (NORFA) is acknowledged.

REFERENCES

MUSICAL ACOUSTICS
BRASS SOUND SIMULATION USING A TWO-DIMENSIONAL LIP VIBRATION MODEL

Seji Adachi, Masa-aki Sato
ATR Human Information Processing Research Laboratories, 2-2, Hikaridai,
Seika-cho, Soraku-gun, Kyoto 619-02 Japan.

SUMMARY
Sound production system of the brass instrument is developed. The player's lip is modeled as a two-dimensional harmonic oscillator having one mass, two sets of a linear spring and a damper along the directions both parallel and perpendicular to the air flow. Lip collision is also modeled. Time-domain simulation indicates that self-oscillations in various air-column resonance modes and at various sound levels are generated by changing the lip eigenfrequency and the blowing pressure, respectively. Waveform and spectrum of mouthpiece pressure varies according to the fundamental frequency and sound level, as is the case in those observed in the the actual instrument.

INTRODUCTION
On the study of sound production in brass instruments, behavior of the player's lips is the central question. The measurements provide us valuable information about the lip vibration but, only with this approach, our understanding of the total system would be incomplete. Simulation, as more than a complement to the experimental means, has an important role in examining the performance of the modeled lip when incorporated into the sound production system.

This paper examines a two-dimensional lip vibration model by carrying out time-domain simulation. Mouthpiece pressure waveforms and spectra are compared with those in the literature observed experimentally[1].

This lip vibration model is based on the experiments that reveal the vibration parallel to the air flow as well as the perpendicular one[2][3]. The previous simulation[4] shows that this model replicates the transition of the lip vibration states observed in the simultaneous measurement of mouthpiece pressure and lip vibration[3].

SOUND PRODUCTION SYSTEM
2-D Lip Vibration Model
For simplicity, the upper and lower lips are assumed to have symmetric motion by the axis of air flow. The dynamics of only the upper lip are thus considered. The lip is approximated as a simple mechanical oscillator composed of one mass, stiffness and damper. Fig. 1 depicts the schematic diagram of the 2-D lip model. The lip is represented by a parallelogram fixed to the mouth through the joint. The tip of the lip moves in directions both parallel (x-axis) and perpendicular (y-axis) to the air flow. The restoring force \( F_{restore} \), is supplied by two springs with the same stiffness \( k \) along the \( x \) and \( y \) directions. The player's mouth
Fig. 1 (left): 2-D lip vibration model
The lip joint is fixed 4.0 mm apart from the axis of air flow. Lip thickness \( d \) is 2.0 mm.
Lip quality factor \( Q \) is set to 3.0 except during lip closure. Lip breadth \( b \), lip mass \( m \) and
stiffness \( k \) vary with the lip eigenfrequency \( f_{\text{lip}} \) (Hz) as follows:
\[
\begin{align*}
  b &= 14.0 - 0.01f_{\text{lip}} \text{ mm}, \\
  m &= 1.5/(4\pi^2f_{\text{lip}}) \text{ kg}, \\
  k &= 1.5f_{\text{lip}} \text{ N/mm}.
\end{align*}
\]

pressure \( p_0 \), mouthpiece pressure \( p \) and lip opening pressure \( p_{\text{lip}} \) cause the external force
\( \vec{F}_{\text{ext}} \) acting on the lip surface.

The equation of lip motion is given by
\[
\frac{1}{2}m \frac{d\vec{\xi}}{dt^2} = -\frac{1}{2} \frac{mk}{Q} \frac{d\vec{\xi}}{dt} + \vec{F}_{\text{restore}} + \vec{F}_{\text{ext}},
\]
with lip mass \( m \) and quality factor \( Q \). Forces, \( \vec{F}_{\text{restore}} \) and \( F_{\text{ext}} \) can be written as follows:
\[
\begin{align*}
  \vec{F}_{\text{restore}} &= \frac{1}{2}k(\vec{\xi} - \vec{\xi}_{\text{equil}}), \\
  \vec{F}_{\text{ext}} &= b(p_0 - p)(\vec{\xi} - \vec{\xi}_{\text{joint}}) + bdp_{\text{lip}}\vec{e}_y.
\end{align*}
\]

where \( b \) is lip breadth, \( d \) lip thickness, \( \vec{\xi}_{\text{joint}} \) position of the lip joint and \( \vec{e}_y \) the unit vector
along the \( y \)-axis. Let the origin of the \( x-y \) coordinate system at the center of the plane of
the mouthpiece entryway. The lip opening area is thus given by \( S_{\text{lip}} = \max \{2k\xi_y, 0\} \). Air flow
is generated in part by the lip movement. On the lip closure condition \( \xi_y < 0 \), additional
restoring force proportional to \( \xi_y \) is supplied, and also the quality factor \( Q \) decreases less
than unity.

Total Sound Production System
The oscillation is described by mouthpiece pressure \( p(t) \), pressure in the lip opening \( p_{\text{lip}}(t) \),
air volume velocity \( U(t) \), and the lip opening area \( S_{\text{lip}}(t) \). They satisfy the following
equations: Eq. (1), two equations governing the air flow dynamics near the lip opening
derived from energy and momentum considerations[6], and a feedback equation representing
acoustical response from the instrument[7].

SIMULATION RESULTS
Resonance Mode Variation
Changing the lip eigenfrequency and the blowing pressure with the adjustment of \( \xi_y \), we
obtain typical self-sustained oscillations in air-column resonance modes II to VI. Blowing
pressures and the sound frequencies are listed in Table 1. Let us examine the mouthpiece
pressure waveforms for the steady state portion in Fig. 2. Waveforms with plateaus and
downward spikes in the lower modes gradually become sinusoidal ones at the higher modes.
This is in good agreement with the observed waveforms of the trombone and the trumpet
in e.g. Elliot and Bowsher[1].
Spectra of these pressure waveforms are shown in Fig. 3. The same tendency as is found in the waveforms that higher harmonics attenuate more rapidly in the higher modes than in the lower ones is illustrated, although the linear slope of the decline in higher harmonics, which was predicted and observed\cite{0}, is not definite.

**Sound Level Variation**

Change in the pressure waveform with sound level for the oscillation in mode II is shown in Fig. 4. Blowing pressures for p, mp, mf and f levels and the obtained sound frequencies are listed in Table 2. The change is very similar to that predicted by the theoretical consideration to the nonlinearity in the sound production system\cite{0}, and is also in good agreement with the observed waveforms of the trombone in different sound levels\cite{0}.

Spectra of the mouthpiece pressures in various sound levels are shown in Fig. 5. Higher harmonics attenuate more rapidly in the lower sound levels than in the higher ones. This is in good agreement with the simulation results with the 'swing-sliding door' model of the lip vibration\cite{0} and with the observed spectra for trombone playing\cite{0}.

**CONCLUSIONS**

With a 2-D lip vibration model, which is based on the observed 2-D lip vibration and capable to replicate the transition of the lip vibration states experimentally observed, brass sound simulation is carried out. Mouthpiece pressures of self sustained oscillations in various air-column resonance modes are in good agreement with those observed in literature.
Table 2: Blowing pressure $p_0$ and sound frequency $f$ for sound level variation.

<table>
<thead>
<tr>
<th>Level</th>
<th>$p$ (kPa)</th>
<th>$mp$</th>
<th>$mf$</th>
<th>$f$ (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_0$</td>
<td>1.0</td>
<td>1.5</td>
<td>2.5</td>
<td>3.0</td>
</tr>
<tr>
<td>$f$</td>
<td>225</td>
<td>229</td>
<td>231</td>
<td>233</td>
</tr>
</tbody>
</table>

Fig. 4 (left): Waveforms for $p$, $mp$, $mf$, $f$ levels
Fig. 5 (above): Spectra for $p$, $mp$, $mf$, $f$ levels

Change of mouthpiece pressure with sound level is also obtained and appears realistic.

REFERENCES


STABLE AND UNSTABLE BEHAVIOURS IN WOODWINDS

Ana Barjau Condomines, *Vincent Gibiat
Dep. d’Enginyeria Mecànica, Univ. Politècnica de Catalunya, (Barcelona).

SUMMARY

Woodwind models involve a set of basic non-linear equations which can be written in the form of more or less intricate non-linear differential equations. The resulting flow allows the study of the stability, instability and bifurcations of those acoustical systems with the usual tools of Non Linear Dynamical Systems. Classical descriptions lead to an infinite number of equations (which can eventually be truncated) while more recent ones reduce the flow to a few non-linear differential equations eventually containing delayed variables. We will discuss the conditions of stability of a simple geometry of the bore and will give a more realistic example.

INTRODUCTION

Traditionally, the time-domain study of woodwinds has been based on algebraic or differential equations for the non-linear part of the system (the reed and more generally the excitation system) and a convolution equation coupling the reed to the linear part (the bore). This formulation can be very suitable for simulations but does not give any clue to find the values of the parameters leading to all the possible behaviors of the system. Particularly it is not easy to predict under which conditions stable an unstable oscillations will take place.

The study of the stability of a Non Linear Dynamical System can be done with the classical tools and concepts of this domain provided that the system description is made entirely through differential equations. In order to transform the convolution equation into a differential one, it is necessary to have an analytical expression for the kernel (in this case the impulse response \( h(t) \) of the bore).

For very simple geometries an ‘ad hoc’ formulation is possible. For the case of the cylinder and the cone these formulation leads to Non Linear Delayed Differential Equations [Grand 1994(a)]. For the general case, the differential description can be obtained through a modal analysis of the bore [Agullo 1986].

The aim of this paper is to develop a systematic study of the stability at the threshold of oscillation for woodwind-like systems through the analysis of the corresponding set of NLDDDEs. Some simulations are presented to assess the validity of the theoretical predictions.

DIFFERENTIAL MODELS FOR WOODWINDS

For the simplest geometry, the cylinder, various models can be found in the literature, [Maganza 1985, McIntyre 1983, Polack 1987]. We proposed a formulation based on simple exponential functions which retains the shape of the exact solution [Grand 1994(a)]:

\[
h_{\text{mod}}(t) = \frac{1}{2} \frac{Z_0}{S} \left[ \delta(t) - \beta \delta(t - T) e^{-\sigma(t-T)} \right], \quad (1)
\]

where \( h_{\text{mod}}(t) \) is defined as in ref. [Agullo 1988], \( \delta(t) \) and \( \varepsilon(t) \) are the Dirac and the Heavyside distributions respectively, \( Z_0 \) is the air characteristic impedance, \( S \) the pipe entrance section, \( T \) is the time of the round trip in the bore, \( \sigma \) is related to the losses, and \( \beta \) is a parameter properly adjusted to fit the experimental data.

The formulation of eq. (1) allows the transformation of the convolution into a differential equation:

\[
\frac{dp(t)}{dt} + \sigma p(t) + \beta p(t - T) = \frac{Z_0}{S} \left[ \frac{du(t)}{dt} + \sigma u(t) - \beta u(t - T) \right], \quad (2)
\]

where \( p(t) \) and \( u(t) \) are the acoustic pressure and air flow.
For arbitrary geometries a modal analysis of the bore makes possible the decomposition of $h(t)$ as a sum of elementary responses $h_n(t)$ [Darja 1990].

The Laplace transform of the convolution $p(t) = \sum p_n(t) = (1/S) \sum h_n(t) u(t)$ leads then to a set of ordinary differential equations:

$$m_n \frac{d^2 p_n(t)}{dt^2} + c_n \frac{dp_n(t)}{dt} + k_n p_n(t) = \frac{m_n}{S} \frac{d^2 u(t)}{dt^2} + k_n \frac{du(t)}{dt}.$$  \hspace{1cm} (3)

The modal parameters $(k_n, m_n, c_n)$ are deduced for an experimental impedance curve [Agullo 1986].

As a preliminary study we have chosen the simplest non-linear dependence representing the coupling between pressure and flow:

$$u(t) = \alpha \left( p(t) - p_f \right) \left[ P_0 - p(t) \right] \left[ p(t) - p_f \right],$$

where $P_0$ is the blowing pressure, $p_f$ the pressure under which no flow is introduced in the system and $\alpha$ is a properly adjusted coefficient [McIntyre 1983]. The lack of description for the reed dynamics does not invalidate the model since any linear system can be included in the modal description.

THE CASE OF THE PERFECT CYLINDER

From all the possible system behaviours (fixed points, limit cycles, N-torus and strange attractors), we will focus our attention on the fixed points since they are the only kind of attractors which can be studied in a systematical analytical way. From an equilibrium position (fixed point) an infinitesimal perturbation is introduced and its evolution is studied through a linearized description depending upon control parameters. In our study $P_0$ will be the only control parameter.

The delayed differential equation for the cylinder (2) presents a single fixed point: $p = 0$.

The linear approximation leads to:

$$\left[ 1 + \alpha(P_0 + p_f)(S_0 / S) \right] \frac{dp(t)}{dt} + \alpha p(t) = \beta \left[ \alpha(P_0 + p_f)(S_0 / S) - 1 \right] p(t - T).$$  \hspace{1cm} (4)

This equation suggests a solution of the form: $p(t) = p(0)e^{(\alpha + \beta) t}$. The possible solutions, called 'modes', can have various behaviours depending on the parameters of the problem. We will talk of oscillatory modes whenever $b \neq 0$, of stable and unstable modes when $a < 0$ and $a > 0$ respectively, and of neutral modes when $b = 0$ and $a = 0$ (which define a bifurcation).

Eq. (4) becomes:

$$a + \sigma = - (\beta / \lambda)e^{-\sigma t} \cos(bT), \hspace{0.5cm} b = (\beta / \lambda)e^{-\sigma t} \sin(bT),$$

where $\lambda = \frac{2}{1 + \alpha(P_0 + p_f)(S_0 / S) - 1}$ is the only parameter depending on $P_0$.

It is possible to eliminate a variable and write a single equation for $b$:

$$b^2(1 + \tan^2 b) = f(\lambda)e^{2kmb},$$

where $f(\lambda) = (\beta T / \lambda)^2 e^{2\lambda T}$. Not much can be said about this equation for the general case. For non-oscillatory and neutral modes, however, the resulting transcendental equations are simple enough as to allow the calculation of threshold values.

MODAL ANALYSIS AND NON LINEAR DIFFERENTIAL EQUATIONS

For a N-mode system, equations (3) show that the only equilibrium position is $p_n = 0$. The linear approach in the vicinity of this position leads to a set of N second order differential equations which can be rewritten as 2N first order ones:

$$\begin{bmatrix} \dot{q}(t) \\ \ddot{q}(t) \end{bmatrix} = \begin{bmatrix} M^{-1}C & M^{-1}K \\ I & 0 \end{bmatrix} \begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix}, \hspace{0.5cm} \{q(t)\}^T = \{p_1(t), ..., p_n(t)\},$$

where $M_y = \delta_y - c_y(\alpha / S)(P_0 + p_f)$, $C_y = -\delta_y(\alpha / S)(P_0 + p_f)$, $K_y = -\delta_y(k_y / m_y)$.

$\delta_y$ is the Kronecker symbol and $I$ is the identity matrix.

The study of the eigenvalues of this matrix gives all the information about the stability of the fixed point as a function of $P_0$. 

- 390 -
THE EXAMPLE OF A TENORA AS A MULTIMODAL SYSTEM

The modal approach has been used to study the behaviour of a 'tenora', a conical Catalan folk woodwind. The bore has been represented by seven modes whose parameters have been derived from previous measurements [Agullo 1986]. Fig. 1 gives the value of $P_0$ for which the imaginary part of the eigenvalues becomes zero (threshold of oscillation). Non oscillatory behaviours can only arise for very high values of $P_0$ (over 19000 Pa). Fig. 2 shows the $P_0$ values for which the modes are neutral (zero damping or bifurcation). The lower value from which an unstable behaviour can take place is 1482 Pa, though such a behaviour is not guaranteed due to the negative damping associate to the other modes. Nevertheless over 2200 Pa the oscillation is necessarily unstable.

Several time domain simulations have been carried out in order to verify the previous conclusions. The particular form of $A_0(t)$ allows the use of the 'fast-convolution' algorithm [Barjau 1990]. We present a selection of these simulations for different values of $P_0$. It is clear from fig. 3 that at the threshold (1482 Pa) the neutral mode succeeds in establishing the oscillation. For higher and lower pressures the other figures show the dependence of the transient duration on the deviation from the threshold.
CONCLUSIONS

A differential formulation for woodwinds has been presented which allows the study of stability of the system through a linear approach. The predictions at the threshold of oscillation for the particular case of a Tenora have proved to be correct. For a more complete exploration of such systems it is necessary to consider the non-linear terms of the whole model [Grand 1994(b)]. However the linear threshold study is always useful as a first step to the effective dynamics description [Manneville 1991] once the transient has been extinguished.

REFERENCES

Agullo, J.; Barjau, A. (1986) 'The reflection function \( r(t) \): a matrix approach versus FFT.' J. Sound Vib., 106(2), 193-201

Agullo, J.; Barjau, A.; Martinez, J. (1988) 'Alternatives to the impulse response \( h(t) \) to describe the acoustical behavior of conical ducts.' J. Acoust. Soc. Am., 84(5), 1606-1612


Grand, N.; Barjau, A.; Gibiat, V.; Derode, A. (1994 a) 'Instruments à vent: formulation intégrale ou différentielle?' J. de Physique IV, CS, 621-624


ASPECTS OF TRUMPET PLAYING

Matthias Bertsch
Institut für Wiener Klangstil, Musikhochschule, Wien.

SUMMARY
The acoustics of brass instruments is a broad field. Some elements, where the "human-factor" is not dominant have already been investigated. As one model, the author shows the influence of several trumpet mutes on the timbre, intonation and responsiveness of trumpets. The actual tone generation (embouchure) is very complex. The factors influencing lip action are so numerous that no quantitative theory can be formulated without further experiments.

INTRODUCTION
It is very obvious that a note with a defined intensity sounds different on various musical instruments. One trumpet also doesn't sound like another trumpet. That is also convincing, but why doesn't one and the same trumpet doesn't always sound the same? There can be several reasons for this. The most obvious one is the use of different mutes, which changes the tone color of the sound intentionally. These effects of trumpet-mutes will be discussed first in this article. Even without mutes one tone sounds different on the same trumpet if different players produce it. This depends on the different tone generation and individual embouchure set-up. Finally analysis shows that even one player produces dissimilar timbres on the same trumpet for all his efforts. This is caused by the complexity of embouchure. The second part of this article deals with parameters which have an effect on this complex.

I. TRUMPET MUTES
1. types: brochures reveal many types of mutes. Surveys of the author result in a ranking list of types which are in use. 1. Cup (93 % of using); 2. Straight (92%); 3. Harmon (75%); 4. plunger or anything like (44%); 5. Wah-Wah (40%); 6. Velvet or Bucket (22%); 7. Whisper (8%); 8. Hat or Derby (6%); 9. Mega-Clear-Tone (3%); 10. Buzz WOW (3%); 11. Mel-O-Wah (2%); 12. Pixie or Snubtone (2%). (The Practicemute (47%) is out of ranking, because different types are used as Practicemute). The six most-employed mutes have been subject matter of an acoustical investigation.

2. dynamik: The dynamic range of the trumpet without mute depends on the register: about 30 phone in the lower and about 13 phone in the upper register (Meyer/1980). Measurements of a crescendo-tone in the anechoic chamber in the IWK reveal the following dynamic range in the lower-register (c1). The reference-amplitude 0 dB corresponds with the ppp (as soft as possible) on
the trumpet without mute. The range for the trumpet without mute and with the plunger almost opened is about 30 dB. The Cup, Wah-Wah, Straight and Velvet Mutes have reduced ranges of about 24 dB. The Plunger has 21 dB in the almost-closed position and the Harmon has even only 17 dB range in the crescendo. The ability to play softer with a mute is true for the Cup, Wah-Wah, Straight and Velvet and Harmon mute. The ppp (as soft as possible) sounds -5/-8 dB lower than with without mute. The chance to play ff (as loud as possible) is most reduced with the Harmon. The ff is 20 dB weaker than without mute. This explains why the Harmon mute is usually amplified when it is in use. The ff played with Cup, Wah-Wah, Straight or Velvet mute is 12 dB softer then without. The dynamic maximum of the Plunger depends very much on the gap size. Almost closed (1cm gap) the ff is about 6dB weaker than without mute.

3. timbre: The sounds produced by some mutes are very characteristic, others sounds similar. The physical reason for a certain timbre are changes in the spectrum. Mutes cause typical formants and above all antiformants. The FFT-Spectrum of the Trumpet without mute is shown in the graph (tone c1, blown fortissimo). The formant area is around 1.2-1.5 kHz. The intensity of the higher partials diminish gradually. The FFT of the Cup mute indicates antiformants at around 2.5 and 5 kHz. Also to be seen are the weakened partials over 10 kHz. The Cup prevents the radiation of wavelengths shorter than the dimension of the mute. Very characteristic is the "Donald Duck" sound of the Wah-Wah mute. The FFT shows the alternating formants and antiformants. The foundamental is very faint. The strong partials around 1.5 kHz entail the nasal timbre. Some more examples for particular characteristics of the other types: The "classical" Straight mute has weak low partials, a formant around 2 kHz and an antiformant at 4 kHz. The Velvet has no antiformant or formant. It darkens the sound by attenuating the high frequencies. (The small wavelengths disappear in the cotton wool bucket).

The Formants of some mutes correspond with vocal formants. This is why the Harmon sounds like "ee" (it nickname is bee) and the Plunger sounds in the closed position like "oo" (doo-wah discribes the closed-open omnopoeetically).

4. response: Impedance-measurements display the influence of the mutes on acoustical behaviour. All investigated mutes except the Velvet add a additional resonance peak to the curve. This peak causes a shift-effect on the other resonance peaks. The dimension of the shift depends on the positon
and magnitude of this additional peak. Good specimens of the Cup-, Straight-, Harmon-, Wah-Wah- mute push the peak below the playing range of the trumpet, and the unintended shift is to a less degree. Bad specimens shift, and even suppress the resonances in the lower register considerable. The additional peak of the Plunger (closed position) lies in between the playing range and actual prevents the sound generation of a "correct" musical pitch. That doesn't matter, because the Plunger is normally used for special "growl-technique".

5. intonation:
The described shifts of the resonance-peaks affect the intonation. The graph shows one sample. In the lower register the trumpet with Wah-Wah mute is much sharper than the trumpet without mute (28 Cent above equal temperature pitch instead of 10 Cent below).

II. ON EMBOUCHURES

The embouchure is the interface between the musician and the brass instrument. The term embouchure is used in two different ways. On one hand, it implies in a narrower sense the on-set of the mouthpiece on the lips and the actual tone generation of the lips. On the other hand there is the meaning of the word in the wider sense. Phrases like "I have no good embouchure today" or "Soft-drinks are not beneficial for your embouchure" indicate two aspects of parameters which affect the player. There are quite a lot of complex parameters influencing the "human-part" of the linked system "player-instrument". Scientific approaches on this subject have been done from pedagogical side and from the instrumental-acoustic side. The bridge is missing. In fact, the tone generation is determined by the air flow and the lip action. The principle have been known for many years. What makes the differences between the same note, played with the same instrument (and even the same player)?
Brass players are no determinate machines who can repeat the same MIDI sample every moment. Recent investigation in Vienna - using new tools to work out more detailed informations about the embouchure will try to explain the differences. Here are some conditions (and examples) which influence the tone generation of a brass player:

**General conditions**
- **Ability**
  - more talented / less talented
- **Age**
- **Education Level**
  - pupil / student / professional
  - school / teacher
  - Vienna / German .......
- **Teeth - Constellation**

**More specific conditions**
- **Time (moment)**
  - day / night
  - morning / evening
  - hungry / not hungry
- **Room**
  - echoic chamber / church
- **Lung Capacity**
- **Surrounding**
  - alone / in front of orchestra
- **Cognitive Processes**
- **Temperature**
  - cold / warm / hot
- **Intention (Music-context)**
  - classic / jazz / else
- **Auditory-system**
  - hearing-ability

**Specific conditions**
- **Motivation**
  - sympathy / antipathy
- **Lips-constitution (and mucous membranes)**
  - "warmed-up" / not "warmed-up"
  - wet / dry
- **Endurance**
  - more pressure / less pressure
- **Muscles**
  - relaxed / forced
- **Breathing (air-flow)**
- **Dynamic**
- **Register (upper / lower)**
- **Instrument**
  - response, intonation
  - quality, characteristic
- **Feeling (used to mouthpiece and instrument?)**

**References:**
TRENDS AND CHALLENGES IN PHYSICAL MODELING OF MUSICAL INSTRUMENTS

Antoine Chaigne
Ecole Nationale Supérieure des Télécommunications, Signal Department, CNRS URA 820, 46 rue Barrault, 75634 Paris cedex 13, France.

SUMMARY
In this paper, the various sources of errors which generally affect the numerical modeling of musical instruments, and the strategies for reducing them to acceptable proportions within the context of audio signals, are presented. These considerations are illustrated by some recently obtained results, with regard to the modeling of exciters, damping and radiation for stringed and percussive instruments.

INTRODUCTION
Numerical simulations of musical instruments start from the definition of a so-called physical model, which is, in essence, an approximation of the reality. In most cases, this model takes the form of a set of continuous differential and partial differential equations. The complexity of such a system of equations imposes the use of numerical methods to solve them. Within the context of music, one leading idea is to keep in the model the terms which are perceptually relevant, and leave aside the complicating factors which may not have audible consequences. The main problem follows, here, from the fact that these consequences are not known until the solutions are obtained. Thus, a systematic approach with step-by-step refinement of the physics seems to be the right way to define the model. Another advantage of such a method is to allow a perceptual evaluation for each improvement of the physical description, and relative comparisons between different degrees of refinement.

A second class of discrepancies between the model and the reality is due to the numerical treatment of the continuous equations. From the point of view of the analysis, one can say that putting the equations into a discrete form simultaneously affects the magnitude, phase and frequency of the components which are present in the solution. These errors can be reduced by selecting appropriate schemes and a sufficient number of spatial points on the numerical grid. This resolution should be selected in accordance with the averaged ear's sensitivity with regard to sound amplitude (roughly 1 dB) and frequency (about 0.3%).

The third class of errors in the results is due to insufficient accuracy in the estimation of the physical parameters appearing in the equations. This mainly refers to the estimation of geometrical parameters of the structures, as well as to the elastic properties of the materials involved in the motion of the instrument under simulation. Slight differences in the damping constants, for example, yield dramatic changes in the results. Therefore, it turns out that a simulation program is necessary linked with series of experiments carried out on real instruments. Even if the ultimate goal is not only restricted to the mimicry of existing instruments, comparisons with such sound sources are necessary, at least during the validation protocol of the simulation program.
Considering the physical, numerical and experimental causes of errors presented above, one has now to imagine a series of validation procedures in order to test a given model. First, in some restricted cases where a comparison with simple analytical results can be made, one has to evaluate the ability of the numerical solutions to behave similar to analytical ones. The degree of resemblance can be then assessed by a mathematical criterion of distance. In other situations, one should content themselves with a test of convergence for the algorithm, which means that the result tends to an asymptotic solution as the number of spatial points increases.

The next step consists in making systematic comparisons between numerical and experimental waveforms and spectra. If the previous test has been carried out correctly, one can be confident in the performance of the numerical algorithm, and thus the remaining errors are either due to the limits of the physical model itself, or to insufficient precision in the determination of the physical constants. A particular attention must be paid to the measurements of damping factors, and a relatively high degree of sophistication for the experiments is usually required. In order to measure the variations of damping factors with frequency for a thin vibrating plate (or a kettledrum), for example, we conducted non contact measurements with acoustic excitation, and optical detection of the velocity, for each mode of vibration of the structure in an anechoic room [Lambourg 95]. The time constant of the free decay, after suppression of the excitation, is then obtained through application of parametric methods [Laroche 93]. Finally, a perceptual evaluation of the results must be conducted, since the ear is sensitive to details which may not be visually detectable on waveforms and spectra. From series of simulated sounds based on systematic exploration of parameters, one can estimate the range in which the numerical algorithm reproduces the most significant features of the modeled instruments. This is obviously a key factor if the purpose is to take advantage of such simulations in order to help the design of real instruments.

In order to illustrate some of the basic ideas presented above, for different families of instruments, a number of recently obtained results will be now briefly described. Attention is paid on the modeling of the following points: the exciters, the damping of materials, and the radiated sound pressure. For percussive instruments, including the piano, it has been shown that the introduction of energy into the the vibrating system can be modeled by a force density, resulting from a complex interaction between an exciter (hammer, mallet,...) and the propagation medium [Ramdane 92], [Chaigne 94], [Doutaut 94]. This interaction imposes the magnitude and duration of the force impulse, which determines both the loudness and spectral width of the sound during impact and decay. Despite significant differences in geometry, material and elastic constants between them, the interaction force for three typical classes of instruments can be described in a unified manner by mean of a power dependance between force and compression (see Table 1). The waveforms shown

<table>
<thead>
<tr>
<th>Instruments</th>
<th>Piano</th>
<th>Tympani</th>
<th>Xylophone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exciter</td>
<td>Hammer</td>
<td>Soft Mallet</td>
<td>Hard Mallet</td>
</tr>
<tr>
<td>Propagation Medium</td>
<td>String</td>
<td>Membrane</td>
<td>Bar</td>
</tr>
<tr>
<td>Interaction Force</td>
<td>$F = K\delta p$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponent (obtention)</td>
<td>$2.3 &lt; p &lt; 3.0$</td>
<td>$1.7 &lt; p &lt; 3.2$</td>
<td>$p=1.5$</td>
</tr>
<tr>
<td>Experiments</td>
<td>Experiments</td>
<td>Theory (Hertz's law)</td>
<td></td>
</tr>
<tr>
<td>Complicating factors</td>
<td>1 Hammer for 2(resp. 3) strings</td>
<td>Hysteresis</td>
<td>Flexibility of the mallet</td>
</tr>
</tbody>
</table>

Table 1: Excitation of various percussive instruments

in Figure 1 confirm, for example, the good agreement obtained between measured and simulated sound pressure, when this model of excitation is used for a xylophone bar.
In stringed and percussive instruments, the major causes of damping are due to internal losses in the material, losses at the boundaries, radiation, and coupling with a resonator. One main difficulty of the measurements on real instruments is to separate these causes of damping, and thus it is often necessary to use a global model for the damping, although each component results from a specific physical mechanism. This method has been successfully applied to the modeling of damping for thin vibrating plates, as a preliminary test for accurate modeling of stringed and percussive instruments. The degree of accuracy of the fit between real and simulated damping factors can be seen in Figure 2 at four successive steps of the modeling. The fluid damping model is represented by a horizontal line in the frequency domain. Adding a viscoelastic term yields a parabolic dependence vs. frequency which is a better representation of the reality, especially in the upper frequency range. Another improvement includes a relaxation term which accounts much better for the losses in the low-frequency range. This third model is shown as a continuous curve at the bottom of Figure 2. Finally, the best fit between the model and the experimental data is obtained when the damping constants are depending on the modal geometry. This model is represented as white circles (o) at the bottom of Figure 2. These successive changes in the modeling of damping are clearly audible in the simulated sounds. Here, the damping factors globally account for both the internal losses and radiation.

Comparisons with real sounds finally impose to simulate the sound pressure radiated by musical instruments. In the case of a xylophone, the following procedure has been used: it is assumed that the action of the fluid on the bar is negligible, so that the pressure can be directly computed from the velocity profile of the oscillating bar. The bar is considered as a 1-D antenna of dipoles, and the total pressure at the given position in the surrounding fluid is obtained by summing the delayed contributions of each dipole, in the far-field. The validity of this model is attested by comparisons between real and simulated waveforms, and by sound examples. This model is presently extended to the modeling of the interaction between the bar and the tubular resonator.

CONCLUSION
In this paper, emphasis has been put on a number of numerical methods used for modeling musical instruments in the time-domain. The results of the simulations may be used for musical composition, and for helping the design of musical instruments. Within the context of musical applications, on the one hand, the question of reducing as much as possible the number of control parameters by the player will probably be one of the main challenges in the upcoming years. The usefulness
Figure 2: Damping factors (in s⁻¹) for thin vibrating plates. Experiments (+). Four successive modeling: fluid damping (top: x); fluid damping and viscoelastic term (top: o); fluid damping and constant relaxation (bottom: x); fluid damping and relaxation depending on the modal geometry (bottom: o).

of such simulations for computer-aided-design of musical instruments, on the other hand, will be established under the condition that a sufficient number of perceptually relevant parameters and variables are taken into account in the model. It is hoped that future design based on physical modeling will not only improve the quality of the instruments, but will also facilitate the emergence of scientific questions relative to the physics, the numerical analysis and the experimental methods in musical acoustics.


CONSTRUCTION AND FREQUENCY CHARACTERISTICS OF CHINESE BOWED STRING INSTRUMENT

Tong Chen
Institute of Acoustics, Academia Sinica, Beijing, China.

The bowed string instrument in China was developed thousand years ago. It is developed into the present instruments of different kinds, such as "Er-hu", "Jing-hu" and "Ban-hu", etc. "Er-hu" is the most widely prevalent instrument in China, fit for solo, ensemble and accompaniment. "Jing-hu" is the special instrument primarily used for the accompaniment in the Beijing Opera. The constructions of other Chinese bowed string instruments are similar to the two instruments, except the "Ban-hu". In this paper, only the former instruments are discussed.

CONSTRUCTION

Fig.1 Components of "Erhu"

1-wood screw;  
2-qianjin;  
3-rod;  
4-strings;  
5-bow;  
6-membrane;  
7-resonator;  
8-bridge;  
9-bottom.

The essential parts of an "Er-hu" are shown in Fig.1. It has two strings called inner and outer strings, respectively. Usually, they are tuned in fifth. The effective length of string measures from the bridge to the "qianjing", which is a double hock nut. The resonator of "Er-hu" consists of a membrane and a section of wood tube. The membrane is made of snake skin and fixed under tension on the resonator tube. The bridge is a small wedge made of bamboo or hard wood. The construction of "Jing-hu" is similar to that of "Er-hu", but it is smaller in size. For most "Jing-hu", the resonator has a bamboo cylinder covered with stretched snake skin at one end.

The specialities of the Chinese bowed string instruments in comparison with the western instruments are:

1. the bow hair touches to the wall of resonator during
bowing, so the bowing point is almost fixed;
2. the bowing direction is nearly perpendicular to the
vibrating surface of the membrane on resonator;
3. the bridge is very small and light.

**RESONATOR**

![Fig. 2 Electrical analog of resonator](image)

![Fig. 3 Waveform and spectrum](image)

Theoretically, the snake skin covered on the resonator can be
considered as a flexible membrane stretched at one end of the
resonator tube. The reactions of air medium in front of membrane
and the sound field in the resonator tube should be included. In
the low frequency range at frequencies well below the first
resonant frequency of the membrane itself, the electrical analog
of resonator can be shown in Fig. 2(1). The membrane with radius a
has an area density \( s \) and is stretched under tension \( T \). In Fig.2,
the following symbols are used:

\[
\begin{align*}
N &= 4.7a^2 = \text{equivalent mass of membrane} \\
C &= \frac{1}{27.4T} = \text{equivalent compliance of membrane} \\
R &= \text{internal loss of membrane} \\
m_a &= 2 \rho a^2 = \text{equivalent mass due to radiation} \\
r_a &= 0.8 \rho \omega a^4 / c = \text{equivalent radiation resistance} \\
m &= 0.35 \rho a^3 = \text{mass load due to air in resonator} \\
Z &= \text{input impedance of resonator tube at the end of membrane}
\end{align*}
\]

where \( \omega \) is the angular frequency; \( \rho \) is air density and \( c \) is the
sound velocity in air.

It can be seen from Fig.2 that the resonator has a main resonance
due to the cancellation of reactances of membrane compliance with
the total inertance of the system. At frequency higher than that
of the main resonance, the resonances due to different
vibrational modes of the membrane exist.

**SOUND SPECTRUM**

The sound spectra of instruments have been obtained by playing
musical notes on inner and outer strings. Fig.3 shows an example
for "Er-hu". The waveform of sound in a period duration is also
shown in the figure. We put seven or more note sound spectra
overlapped together to obtain the spectrum envelopes for inner
and outer strings. Fig. 4 shows the spectrum envelopes obtained for a "Er-hu", in which the relative sound level for inner and outer strings are arbitrary. The spectrum envelopes obtained for individual strings may be useful for understanding the frequency characteristics of sound of bowed instruments.

Fig. 4 Sound spectrum of Er-hu

Fig. 5 Peak sound pressure spectrum

To verify the method used for obtaining the spectrum envelope, comparison with the ordinary technique was made. The 1/3 octave peak-sound level spectrum for another "Er-hu" measured by playing Chinese ancient music in the laboratory of Qinghua University is shown in Fig. 5. The spectra shown in Fig. 4 give the main points of Fig. 5, if the sound level of the spectrum envelope for outer string given in Fig. 4 is adjusted as the real case.

DIMENSIONS AND FREQUENCY CHARACTERISTICS

The sound spectra shown in Figs. 4 and 5 are characterized by some peaks. The lowest frequency peak is due to the main resonance of the resonator. Its resonant frequency can be calculated according to the equivalent circuit shown in Fig. 2. As an estimation, we neglect the inertances and obtain the condition for resonance as follows:

$$\rho c \pi a^2 \tan (kL_e) = 27.4 \frac{T}{\omega}$$

where $k=\omega/c$ is the wavenumber and $L_e$ is the equivalent length of the resonator tube. Then, we have the equation for solving the resonant frequency:

$$(kL_e) \tan (kL_e) = (8.7 \frac{T}{\rho c^2})(L_e/a^2)$$

If the value of $T(L_e/a^2)$ is large enough as in the case of "Jin-hu", then $(kL_e)$ approaches to $\pi/2$, which is the 1/4 wavelength resonance of the resonator tube. In this case, the resonant frequency is mainly determined by $L_e$. $(kL_e)$ decreases with the decrease of $(T(L_e/a^2))$. For a "Jin-hu", $a=2.3$ cm, $L_e=13.3$ cm and $T=3.6 \times 10^4$ N/m. The calculated main resonant frequency of the resonator is 628 Hz, which is a little higher than the measured value 625 Hz due to the omitting of inerance.

A lower main resonant frequency is expected for "Er-hu", since its resonator is larger than that of "Jing-hu". For an "Er-hu".
a=4.5cm and $L_e = 16cm$. If tension $T$ is kept unchanged, the calculated main resonant frequency of resonator for "Er-hu" is about 500 Hz due to the decrease of the value $L_e/s^2$. The experimental value is 462 Hz, since the tension cannot be the same as that for "Jing-hu".

The next two peaks on the sound spectrum are due to the oscillations of that part of string between bow point and bridge for the inner and outer strings respectively, which has been discussed by Lawergren(2). The frequency of such an oscillation depends on the relative bow point along the string, and can be expressed as

$$f = (L/D) F$$

where $F$ is the fundamental frequency of string with length $L$ and $D$ is the distance between bow point and bridge. $D$ depends on the size of the resonator. Since $F$ is inversely proportional to $L$, the frequency of the oscillation between the bow point and the bridge, and hence the frequency of the peak on the sound spectrum, is constant for a constant $D$.

The two peaks of sound spectra shown in Fig.4 for the "Er-hu" are at frequencies about 1400 Hz and 2100 Hz for the inner and outer strings, respectively. The resonances of the resonator with the resonant frequencies around the frequency of the oscillation between the bow point and the bridge make the peak broad and have a little influence on the frequency of the peak. "Jing-hu" is smaller in size than "Er-hu". The sound spectrum of "Jing-hu" has two peaks at frequencies of 2300 Hz and 3800 Hz.

At frequency higher than that of oscillation between the bow point and bridge, both spectrum of Helmholtz motion and that of the oscillation decrease with ripples. Thus, the trend of the sound spectrum shown in Figs.4 and 5 falls after the peak due to the oscillation between the bow point and bridge. However, the resonator has more significant influence on the sound spectrum. In Fig.4, two small peaks at high frequencies are common for both spectra of the inner and outer strings. It is the evidence of the influence of the resonator resonance. In some cases, harmonics of the oscillation between the bow point and bridge may occur as discussed in the literature(2). The spectrum has sharp peaks near the frequencies of these harmonics. It is only present in the sounds of some notes and not included in the spectra given in Fig.4.

REFERENCES


TONAL ARCHITECTURE, SOUND CHARACTER SPACES
Francesc Daumal i Domènech and Diana Möller Parera

SUMMARY

we propose to use The Tonal Sounds in Architecture as a first/last method of Acoustical Design Tools about sound energy according the -indoors and outdoors- space Character.

"FOSSAR DE LA PEDRERA",
Beth Gali, Arq.
Barcelona
THE ARTISTIC ACOUSTICS

The acoustic technology and artistic feeling unify both in the artistic acoustics. This can be a reference to perceive intentions through different spaces.

In the architectural acoustic discipline there is a hole to fill: "The Acoustical Poetry". The architecture has been making poetics for the "homo videns" but not yet for the "homo audiens", in order to be more complete.

Does the acoustical poetry exist?. We designed some examples in our country, but they are others in the world, which represent the top of a great iceberg. The knowledge of the state of the art in acoustical poetics remains still scarce.

However my teach and work office at the Architectural School of Barcelona demonstrate the architects' and students' interest for those unknown subjects. It's necessary to find this examples to study their poetic language and acoustical foundation.

"CENTRO DE ARTE SANTA MONICA"
Albert Viaplana - Helio Piñon, Arqtos.
Barcelona
TONAL ARCHITECTURE

With this state of the art, we can make a complete study of sound tonality of new examples. This research claims to synthesize these knowledges for an intelligent application of space designs.

The tonal acoustic character is based in the study of shapes, proportions, structures and break materials of the halls and indoors and outdoors spaces.

The hall's proportion was always a subject of special attention in almost all the treatises on architecture as the "gold rules" show.

For halls it's interesting to determinate the relations between the proportions called "musical", the best "acoustics" proportions proposed by the experts, and the "gold numbers" of distinguished architectural writers.

The halls proportions have been related to the musical strain/tone of three notes.

It was necessary to redetermine the fields of application in the different formulations, analized for this tonal purpose.
BIBLIOGRAFÍA


BARDYN, Jean-Luc (1993), "L’appel du port", Cresson, Grenoble.


DAUMAL, Francesc (1993), "La Tonalidad como herramienta de diseño acústico-arquitectónico", Comunicaciones de las Jornadas Nacionales de Acústica – Tecniaónica’93–, Sociedad Española de Acústica.


COMPUTATION AND MODELISATION OF THE SOUND
RADIATION OF AN UPRIGHT PIANO USING MODAL
FORMALISM AND INTEGRAL EQUATORS

Philippe Dérogis, René Caussé*
IRCAM, 1, Place Igor-Stravinsky 75004 Paris, France.

SUMMARY
In all string instruments, the sound is radiated as a result of the coupling of the string with a plate. In the case of the piano the energy given by the hammer to the string, flows through the bridge to the soundboard which dissipates it into internal losses and sound power. Thus the soundboard acts like a mecano-acoustic transducer. In this study a modal analysis is used in order to investigate the vibration characteristics of an upright piano soundboard. The results of this analysis are used for computing the radiation characteristics of the soundboard. Finally a way to improve the radiation efficiency of the soundboard in the low frequencies is proposed.

MODAL ANALYSIS

The use of modal analysis is very helpful for understanding the vibrational behavior of a structure in the low frequencies (where the modal density is not too high). In order to perform a modal analysis on a continuous structure we must consider it as a system with finite degrees of freedom (vector $x$) which verify the equation

$$M \ddot{x} + C \dot{x} + K x = p$$

(1)

Without any hypothesis on the damping matrix (the matrix $C$) this equation can not be formulated as an eigen value problem. A way to deal with this is to write (1) in the form

$$\begin{bmatrix} C & M \\ M & 0 \end{bmatrix} \begin{bmatrix} \ddot{x} \\ \dot{x} \end{bmatrix} + \begin{bmatrix} K & 0 \\ 0 & -M \end{bmatrix} \begin{bmatrix} \dot{x} \\ x \end{bmatrix} = \begin{bmatrix} p \\ 0 \end{bmatrix}$$

or

$$A x + B \dot{x} = s$$

(2)

Which is a classical eigen value problem. The eigen values $\lambda_k$ and the eigen vectors $y_k$ of equation (2) appear in complex conjugate pairs and the eigen vectors can be written in the form

$$y_k = \begin{bmatrix} z_k \\ \lambda_k z_k^* \end{bmatrix}$$

(3)

where $z_k$ are the eigen vectors of equation (1). The response $x(t)$ of the system to an harmonic excitation applied at points $s$ is then given by

$$x(t) = \sum_{k=1}^{n} \left( \frac{1}{j \omega - \lambda_k} \frac{z_k z_k^*}{\rho_k} + \frac{1}{j \omega - \lambda_k} \frac{\overline{z_k z_k^*}}{\rho_k} \right) se^{j \omega t}$$

where

$$\rho_k = Z_k^* C z_k + 2 \lambda_k z_k^* M z_k$$

(4)
From equation (4) we can derive the response \(a_{rs}(w)\) at the point \(r\) to an excitation at the point \(s\). The frequency response function is given by

\[
a_{rs}(w) = \sum_{k=1}^{n} \left( \frac{A_{rs}(k)}{jw - \lambda_k} + \frac{\overline{A_{rs}(k)}}{jw - \overline{\lambda_k}} \right) \quad \text{with} \quad A_{rs}(k) = \frac{z_r(k)z_s(k)}{\rho_k}
\]  

\(5\)

**MEASUREMENTS AND RESULTS**

We have performed our measurements on the soundboard (1.3mx0.9m) of an upright piano with strings mounted. The plate was discretised into 13x13 meshes each having a surface of 0.11mx0.075m. The transfer functions between the point (6,6) of the mesh and the 144 other points were then measured with an impact hammer and an accelerometer. The spectrum of the excitation was almost flat in the frequency range 0-600 Hz, yielding a good estimation of the responses in this range. The measured frequency response functions didn’t exhibit sharp peaks and the eigen frequencies were found to be close. We therefore used the matrix pencil method [1], which is especially appropriate for highly damped signals with close frequency components. A statistical treatment of the results allowed us to choose the same eigen values for all measurements. The computation of the eigen vectors have been performed by fitting the measurements with the function family given by equation (5) and solving the linear system linking the eigen vectors to the coefficients \(A_{rs}(k)\). The resulting eigen values and mode shapes are given in figure (1)

**Figure 1 : Modes shapes and eigen frequencies**

<table>
<thead>
<tr>
<th>Mode 1: 122 Hz, Damping 50 s(^{-1})</th>
<th>Mode 2: 160 Hz, Damping 30 s(^{-1})</th>
<th>Mode 3: 189 Hz, Damping 49 s(^{-1})</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Mode 1" /></td>
<td><img src="image2.png" alt="Mode 2" /></td>
<td><img src="image3.png" alt="Mode 3" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mode 4: 199 Hz, Damping 47 s(^{-1})</th>
<th>Mode 5: 231 Hz, Damping 42 s(^{-1})</th>
<th>Mode 6: 247 Hz, Damping 40 s(^{-1})</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image4.png" alt="Mode 4" /></td>
<td><img src="image5.png" alt="Mode 5" /></td>
<td><img src="image6.png" alt="Mode 6" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mode 7: 282 Hz, Damping 47 s(^{-1})</th>
<th>Mode 8: 308 Hz, Damping 38 s(^{-1})</th>
<th>Mode 9: 326 Hz, Damping 45 s(^{-1})</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7.png" alt="Mode 7" /></td>
<td><img src="image8.png" alt="Mode 8" /></td>
<td><img src="image9.png" alt="Mode 9" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mode 10: 350 Hz, Damping 43 s(^{-1})</th>
<th>Mode 11: 378 Hz, Damping 60 s(^{-1})</th>
<th>Mode 12: 404 Hz, Damping 47 s(^{-1})</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image10.png" alt="Mode 10" /></td>
<td><img src="image11.png" alt="Mode 11" /></td>
<td><img src="image12.png" alt="Mode 12" /></td>
</tr>
</tbody>
</table>

We see that the first mode shapes are close to that of an isotropic supported plate. Indeed the first and the second, the third, the fourth, the fifth and sixth are close to the [2,1], [3,1], [1,2], [4,1], and
the [3,2] modes of an isotropic supported plate respectively. These results are similar to those found by Nakamura [2] and by Brelay [3] for upright pianos with strings.

SOUND RADIATION

The sound pressure $p(r)$ and the acoustic velocity $v(r)$ resulting from the vibrations of a baffled plate moving with normal acceleration $\gamma_n$ are given by

$$p(r) = \rho_0 \int \int \frac{e^{-jk|r-r_0|}}{2\pi|r-r_0|} dxdy_0$$

$$v(r) = \frac{1}{2} \int \int \gamma_n(r_0) \nabla \left( \frac{e^{-jk|r-r_0|}}{2\pi|r-r_0|} \right) dxdy_0$$

(6)

where $r$ is the observation point, $r_0$ the integration variable, $a$ and $b$ the dimensions of the plate and $\rho_0$ the air density. From $p$ and $v$ we can compute the active intensity vector at the point $r$ and by an integration over a closed surface (an hemisphere) we can compute the power $W_a$ radiated by the plate.

$$W_a = \int \int \dot{n} \cdot dS$$

with

$$\dot{n} = \frac{1}{2} Re(pv^*)$$

(7)

Equation (4) allows us to determine the displacement of the soundboard resulting from any excitation. In this way we have computed the displacement resulting from the excitation of several points on the bridge: one point at the middle of the bass bridge (A1: 55 Hz) one at the low edge of the main bridge (A2: 110 Hz) and one a fifth higher (E2: 164 Hz). Then equations (6) and (7) allowed us to compute the power (in Watt) radiated by the soundboard for an excitation of 1N. The results are shown in Figure(2). We note that the best point for radiating low frequencies is point #2. The difference between the radiated power resulting from the excitation of the points #1 and #2 suggests that there could be a timbre change when passing from the bass bridge to the main bridge. Suzuki [4] measured the radiated power for a 6-ft grand piano excited by a 1N driving force. Similarly we found curves having sharp peaks at the resonance frequencies but with slightly lower amplitudes. This shows that upright pianos are poor in radiating low frequencies.

**Figure 2 : Radiated power for several points excited by a 1N force**

Point #1: Bass bridge A: 55 Hz  
Point #2: Main bridge A2: 110 Hz  
Point #3: Main bridge E2: 164 Hz

It is therefore interesting to compute the radiation efficiency of the soundboard. We show these results in figure (3) for the point (6,6) of the mesh (the point where the measurements were performed). Unfortunately our measurements were not sufficiently accurate to obtain good information on the phase for the other points. We see that below 450 Hz, the radiated power represents about 20% of the input power corresponding to an 80% loss in the soundboard. This is also poorer than the radiation efficiency found by Suzuki [4] for a 6-ft grand piano.
Figure 3: Radiation efficiency

Acoustic power.

Mechanical power.

Radiation efficiency.

Figure 4:
Estimated radiation efficiency

By looking at the shape of the first modes of the soundboard we see that they look like that of a plate. However the eigen frequencies of the soundboard modes appear to lie below the coincidence frequencies of that of a plate having similar dimensions. These modes can therefore be excited but radiate poorly. In figure(4) we show the radiation efficiency of a soundboard having eigen values (frequency and damping) a factor 1.5 higher. We see that the radiation efficiency is greatly improved. This curve is however overestimated (efficiency above 100%) because the increase of the damping coefficient due to the fluid loading was not taken into account.

CONCLUSION
The results of a modal analysis of the soundboard of an upright piano from below 100Hz-440Hz have been presented. In a low frequency range the soundboard vibrates like an isotropic plate as shown by Nakamura [2]. Then using the modal theory the displacement of the soundboard resulting from various point excitations have been computed and used to deduce the acoustic power radiated. The power radiated by the soundboard was shown to depend to the excitation point. The computation of the radiation efficiency have been performed for one point excitation. The radiation efficiency appears to be rather poor. Indeed the eigen frequencies are too low in comparison to the soundboard dimensions and therefore the acoustic field is very reactive. In order to highlight this fact, the radiation efficiency has been calculated for an imaginary soundboard having the same mode shapes but with higher eigen values. In this way the radiation efficiency is largely improved.

REFERENCES

- 412 -
AN EVENT RECORDING SYSTEM FOR THE AUTOMATIC TRANSCRIPTION OF MUSIC PLAYED ON IDIOPHONES APPLICATION TO THE GAMELAN GENDER WAYANG

Pierre Dutilleux, Thomas Ruoff*
Center for Art and Mediatechnology Karlsruhe, Institute for Music and Acoustics, Ritterstrasse 42, D-76137 Karlsruhe, Germany

SUMMARY

A system is described to record the performance of an orchestra of idiophones, it was used for recording a gender orchestra in Bali. A contact microphone is stuck under each of the metal plates of the instruments. The amplitude envelope of each of the microphones is electronically computed and then sampled and stored in a computer. As a way of controlling the validity of the acquired data, an event analysis is performed. It delivers MIDI information that is sent to a synthesizer. The synthesized sequence can then be compared to the recorded music. The collected information will be used to produce scores of the repertoire of the gamelan gender wayang. Spectra and tuning of gender sets are also presented.

INTRODUCTION

Before describing the data acquisition system, we must give a short introduction to the cultural domain for which the system was developed. Then we will focus on the technology, then comment about the field-operation of the system and the data that we have collected.

It is known that Bali is a very attractive place for tourism, that its music is very lively and exotic, but who in the western countries has an in depth knowledge of the balinese music? The lay man would say "it sounds strange", the lecture of the musicologist would be clouded with exotic words, the musician, proficient in our "classical music", would be annoyed by the steadily mistuned instruments. Percussionists would be amazed by the original patterns and playing methods that have developed in Bali. Some people though are familiar with the music of Bali. These are musicians that got trained at playing along with the balinese performers. They have learnt directly from a master, in a way that is unusual to us. They almost never use any paper or transcriptions for pedagogical purposes. The music is not written, it is learnt by hearing and imitating, again and again.

Among the various genres of the balinese music we are especially concerned by that played using genders. A gender is a "metallophone constructed so that the keys are suspended over bamboo resonators; more specifically the term is used to connote genders that are played with two mallets, one in each hand". They are used in the gamelan gender wayang : "a four-piece, slendro-tuned ensemble used mainly to accompany the shadow play" [Tenzer]. The four genders are : two larger genders and two smaller genders that sound one octave above the larger ones. Each gender covers 2 octaves. The slendro tuning has 5 equal intervals in each octave. The shadow play is a theatrical performance where the characters are puppets who's shadows are projected onto a screen.

In order to broaden the understanding and the knowledge of the gamelan gender wayang within western countries, a project led by the composer Johannes Goebel, intends to collect and transcribe the repertoire of a prominent master in bali : I wayan Loceng.

Even the skilled European gender players have troubles at deciphering the different musical parts from a recording. To help them in the transcription process, a special recording equipment was
desired. Such a system must be able to keep a separate track of each musical part of the interlocked performance. It is not the first time that the musicologists promote the development of an original investigation technique: see [Arom] in the realm of African ethnomusicology.

From the recording of the sound of an orchestra it is very difficult to trace back the various musical parts. We chose therefore a technique for recording the activity of each of the sounding elements in an individual recorder-channel. The system should be able to produce a note-list, indicating when and how which plate has been struck. This list would be comprised of records of the type:

<Plate number, Onset time, Onset intensity, Damping time>

where "Onset time" and "Onset intensity" describe how the player has hit the "Plate", and "Damping time" tells when he has stopped the vibration of the plate with his thumb-ball.

The instruments of the gamelan gender wayang are idiophones, that means that they vibrate by themselves when they are struck. Idiophones are all those musical instruments which are in general struck or shaken, like bells, xylophones, shakers. With a contact microphone it is hence possible to get the information about when and how strong they are excited. So the principle of the technique is to stick a contact microphone under each of the sounding elements of the orchestra. In our case, we have to deal with a set of four genders, each of them having 10 plates. In addition we should be able to record also the mallet of the puppet player, the dalang. With his mallet, the dalang gives informations to the musicians; this enables a tight synchronisation between theatrical performance and music. That means that we need at least 41 channels of data-acquisition.

We did not find any suitable standard equipment to solve our problem. So we decided to build a recorder for the amplitude curves of a large amount of contact microphones. We named the device an Idiagograph. This new name stands for: "device for transcribing the music played on idiophones and taking into account the agogy".

IDIAGOGRAPH HARDWARE

The system is designed for recording the amplitude curves of the signals from contact microphones onto a computer hard disk. The microphones should disturb neither the musician nor the sound of the instrument. The most practicable place for sticking the microphone is under the plate at the outer edge. Sticking the microphone damps the vibration of the plate, this effect should be kept very low. The suspension of the microphone must be very flexible not to introduce mechanical crosstalk between acquisition channels. Among the microphones that we tried, we retained the plain piezoelectric film because it damps very little the plates and it can be connected using a very flexible cable.

Charge amplifiers convert the charges, developed into the piezoelectric film by the vibration of the plate, into usable low impedance voltages. Logarithmic root-mean-square converters (LOG-RMS) are used to compute the amplitude envelopes and express them in decibels. The conditioned signals are multiplexed and then digitized using a single 8-bit analog-to-digital flash converter (ADC). The data from the ADC are stored onto a computer disk. A data acquisition board was built for a PC-AT compatible. This board hosts the ADC, the clock and the address generators for the multiplexers, and a memory to buffer the transfers between the ADC and the hard disk. The charge amplifiers, the LOG-RMS converters and the multiplexers are in external conditioning modules.

To allow for flexibility in cabling the system within the gamelan gender wayang, a conditioning module was built for each gender. Each module can handle 12 channels, and the interfaces between modules and acquisition board are designed to allow for larger distances.

IDIAGOGRAPH SOFTWARE

A special program was written for each of the operating modes of the idiagograph. Before recording a piece, it is necessary to make sure that each channel works properly and to establish
reference values for a fortissimo and a pianissimo. For each acquisition run, it is possible to adjust the acquisition rate. In practice we have used rates from 100 Hz to 400 Hz. The lower rates are valid for larger groups or for longer pieces. The higher rates are preferred for recording solo or virtuoso performances.

A program is available for displaying the measured information, 2 channels at a time. It is most useful for appreciating the quality of the acquisition channel: noise floor, scaling, crosstalk, and for manually checking the individual events. A program detects automatically all the onsets and dyingouts of notes and it outputs a sorted note-list. This program is most critical for the transcription. If it is too sensitive, it will generate a lot of false detections, if it is too coarse it will miss refinements of the performance. By comparing the note-lists and the acquired data with the audio and video recordings, it is possible to judge the quality of the detection and eventually to improve the detection program itself, hence the note-list.

The note-list contains absolute times, amplitudes, and channel numbers. A possibility to use this list is to convert it into a MIDI format. It can then be sent to a synthesizer. The sound produced by the synthesizer, although it is a very coarse representation of the gender performance, allows to check whether the data-acquisition was significant and consistent. Back to Europe, the note-lists in the MIDI format can be further processed by state of the art programs for musicians. The transcription of the note-lists is under progress. Additional software tools are under development by Heinrich Taube.

RECORDING IN BALI

In August 1994 the repertoire of the gender master I wayan Loceng was recorded in Bali, in the Sukawati village. Since, in a gamelan gender wayang, the 2 smaller genders merely double the musical parts played by the larger gender, only the 2 larger gender were recorded. 30 pieces have been recorded, amounting to about 2 hours of music. In addition a whole shadow play theater performance has been recorded, lasting 2 hours and 45 minutes.

MUSICAL SCALE AND SPECTRUM OF A PLATE

To compare the resynthesized music to the original one, it is important to know about the tuning and the spectra of genders. I wayan Loceng has several sets of genders available. Each set has its own tuning. To tune a set, one proceeds as follows: one of the larger gender is first tuned. It will be the "larger larger" gender. The second larger gender is tuned a few hertz above the lower larger gender: each plate sounds about 7 Hz higher than the corresponding plate of the lower gender. When the plates are tuned, then the bamboo resonators are tuned to the fundamental frequency of the plate. Finally the smaller genders are tuned one octave above the larger ones, apart from a few herz, to allow for shimmering. We notice that the octaves are stretched (Table 1), we neither found this information in [McPhee] nor in [Tenzer]. The values here are derived from measurements made on a set of new genders that is rated "very good" by Loceng.

<table>
<thead>
<tr>
<th>Tone name</th>
<th>dong0</th>
<th>deng0</th>
<th>ding0</th>
<th>deng0</th>
<th>ding0</th>
<th>deng0</th>
<th>ding0</th>
<th>deng0</th>
<th>ding0</th>
<th>deng0</th>
<th>ding0</th>
<th>deng0</th>
<th>ding0</th>
<th>deng0</th>
<th>ding0</th>
<th>deng0</th>
<th>ding0</th>
</tr>
</thead>
<tbody>
<tr>
<td>f (+/- 1.5 Hz)</td>
<td>172</td>
<td>196</td>
<td>229</td>
<td>266</td>
<td>310</td>
<td>354</td>
<td>412</td>
<td>472</td>
<td>541</td>
<td>630</td>
<td>718</td>
<td>848</td>
<td>962</td>
<td>1102</td>
<td>1302</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervals (cent)</td>
<td>226</td>
<td>269</td>
<td>259</td>
<td>265</td>
<td>236</td>
<td>233</td>
<td>236</td>
<td>264</td>
<td>226</td>
<td>288</td>
<td>218</td>
<td>235</td>
<td>289</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>octaves (cent)</td>
<td>1250</td>
<td>1286</td>
<td>1252</td>
<td>1229</td>
<td>1228</td>
<td>1224</td>
<td>1250</td>
<td>1233</td>
<td>1255</td>
<td>1257</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Fundamental frequencies, intervals and octaves, as measured on a larger and a smaller gender.

The spectra of several tones have been measured. The tones are highly non-harmonic. As an example, let us consider the spectrum of the deng0 plate. The fundamental frequency (229 Hz as given in Table 1) appears here in the 232 Hz analysis bin, it is not the strongest partial. The higher partials often appear as doublets. The spectra are very dependent on how hard and on where the plate is struck. Loceng uses this feature to produce subtle timbre variations.
Spectrum of the plate duung0, medium struck at the center, using a hard-wood mallet.

TRANSCRIPTION

The note-lists collected in Bali are being processed to appear in a form that would be easy to use by western musicans. The critical eye and ear of a trained gender player, Andreas Herdy, is necessary to edit the notes detected by the computer.

CONCLUSIONS

A data acquisition system has been built and programmed to record the performance of a gamelan gender wayang. During one month in Bali, the repertoire of the master Iwayan Loceng and a shadow play performance have been recorded. The collected data is of great value for musicologists and musicians. We hope that it will contribute to a better understanding of the Balinese music and to its preservation.

The investigation of the scales and spectra shows that the octaves are stretched and that the genders of the same gamelan are mistuned by about 7 Hz. Along with the doublets in the spectrum of each plate, these features contribute to a lively shimmering sound.

The idiograpth system has been tested with a gamelan gender wayang, but it could be used for other gamelans as well as for other idiophone orchestras. The 48-channels configuration that has been used in Bali could be extended to 360 channels, allowing to record large ensembles.

ACKNOWLEDGEMENT

We are indebted to I wayan Loceng for his hospitality and his excellent temper. To our resynthesized musical sequences, he said: "It sounds as if you were playing like beginners!". Thanks to Ako, the Japanese student who played one part of the repertoire along with Loceng.

REFERENCES

RUOFF Thomas, Messwerterfassungs-System für die Automatische Transkription von Musik aus Idiophonen, Diplomarbeit, Fachhochschule Karlsruhe, Feinwerktechnik, 1993.
Banjar BABAKAN desa Sukawati, Wayan LOCEENG, Wayan SARGA, Ketut BALIK, Wayan NARTHA, Gamelan gender wayang, music of Bali, Ricks records, Maharani, RCD-03.
FINITE ELEMENT SIMULATION AND MEASUREMENTS OF CURVED SWISS HORN

R.-F. Fehlmann
Norwegian Institute of Technology, Faculty of Electrical Engineering and Computer Science, Dep. Telecommunications, Acoustics, Trondheim, Swiss Engineering School CAD/CAM Department, Berne, Switzerland.

SUMMARY

This paper deals with the simulation of a curved Swiss horn by the finite element method (FEM). The resulting resonance modes are presented in form of the overlap of three models. It is then shown that a special curvature of the shape of the bell makes a conical horn harmonic analogous to the Indian drums. Measurements with and without mouthpiece confirm the harmonicity of the curved horn.

INTRODUCTION

It has been shown that the resonance modes of an axisymmetric cone with circular cross section \( S(z) = S_c z^2 \) following the B.E.L. horn wave equation [BERNOULLI (1740/62), EULER (1759/66), LAGRANGE (1761)]

\[
\frac{1}{c^2} \frac{\partial^2 \phi}{\partial z^2} - \frac{\partial^2 \phi}{\partial x^2} = \frac{1}{S} \frac{dS}{dx} \frac{\partial \phi}{\partial x} \tag{1}
\]

with closed-open Cauchy (coC) boundary conditions are not harmonic [FEHLMANN (1994 and 1995)]. Since it can be heard that the timbre of one of the four ethnic horns traditionally built in Switzerland is special, the author has been tempted to investigate the effect of the so-called 'chrump', the special curvature of the curved Swiss horn's bell.

RESONANCE MODES OF THE CURVED SWISS HORN

Today it is known that the FEM can be applied to any physical problems - be it possible to solve analytically or not - which can be represented by partial differential

![Fig. 1: Geometrical display of curved Swiss horn with hemispherical radiation space](image)

- 417 -
equations. Especially for problems involving complicated structures the FEM is most suitable: the 3D-problem of a curved non-axisymmetric horn with an unknown cross-sectional change and center line is thus a good object to be treated by this numerical method. Similar assumptions - as in the case of the cone - [see FEHLMANN (1994 and 1995)] were made for the FEM-calculations: for determining the resonance frequencies we divide the total volume \( V \) with both rigid and flexible walls as boundary conditions (BC) into a waveguide part \( V_{\text{wg}} \) and a radiation part \( V_{\text{rad}} \) with

\[
V = V_{\text{wg}} + V_{\text{rad}}
\]

and the total surface \( S \) enclosing this volume is split into a flexible, an acoustically hard (rigid) and an acoustically soft (absorptive) interface such that

\[
S = S_{(r)} + S_{\text{wg}} + S_{\text{rad}} + S_{h}
\]

Assuming this irregularly shaped cavity \( V \) enclosed by \( S \) to contain an inviscid, compressible fluid, the acoustic pressure \( p \) must satisfy globally and locally the Helmholtz equation

\[
(\nabla^2 + k^2)p = 0
\]

Integrating the Helmholtz equation, applying the necessary BC's, using Galerkin weighting and transforming by Green's integral theorem leads to [see e.g. FILIPPI (1983)]:

\[
\int_V (\nabla N_j^2) dp dV - k^2 \int_V N_j^2 dp dV - \int_{S_h} N_j dS_{(r)} - \epsilon_{ijk} \int_{S_{(r)}} \nabla dS_{(r)j} = 0
\]

Multiplicating all terms by \( 1/(\rho_0 c^2) \) and isolating \( p \) from the integrals yields a first term representing the kinetic energy (or stiffness matrix), the second integral is the potential energy (or mass matrix \( M \)) and the last term on the l.h.s. corresponds to the energy crossing the hemispherical surface \( S_h \) (boundary matrix). On the r.h.s. we have the source term for the wave generating surface \( S_{(r)} \). Remember that the infinite impedances or zero admittances at wall and baffle surfaces give zero contribution to the above equation. This set of \( m \) linear equations with complex coefficients which has been produced by dividing the total volume \( V \) into \( m \) finite elements yields after computation files with the complex pressure for each node. (For results see fig. 2):

**MEASUREMENTS**

In order to check the numerical results measurements of the resonance modes were made by setting the microphone 1[m] away from the horn opening. Using an FFT-analyzer the frequency resonance mode could be determined. According to our model the measurements had to be made - strictly speaking - without a mouthpiece since we have
assumed a constant velocity oscillation of the input throat cross-section of the horn. In using a mouthpiece one actually has to take into account the frequency shift effect for the higher modes [cf. HALL (1990)]. But since it was impossible for the author to play the fundamental mode and the highest modes in the 5th octave without a mouthpiece, different measurement series for both with and without mouthpiece were made (see Fig. 3). Now, up to the 16th mode both curves do agree nicely with the harmonic line. Considering first the overtones blown without the mouthpiece mode nr. 15 is 20 [Hz] off, all the rest of the modes enormously well follow a harmonic behaviour! Even this drop-out mode is just 3%, i.e. a quarter tone off from an ideal behaviour.

Even though it seems to be the first time that a recording of modes up to nr. 32 could be made it has to be noted, however, that the stability of the mode nr. 25 and upward is quite shaky so in order to guarantee really controlled and confidential frequency measurements of the mode nrs. 25-32 a blowing machine should be used. However, up to the 23 $^{th}$/24 $^{th}$ resonance solid blowing can be warranted and within this region even Hall's downward shift effect of the mouthpiece is ascertained.
It is worthwhile mentioning that in contrast to classical knowledge neither the natural septime (7th mode) is flat nor is the 11th mode (the horn 'fa', 'tonus diabolicus') too sharp nor is the 13th too flat: be it blown with or without the mouthpiece! The results presented here thus de-diabolize these scapegoat tones [see e.g. Van der MAAS (1985)] of classical European music theory!

CONCLUSIONS

It has been shown that with the 'chrump' contour the inharmonic resonant modes of a cone are made harmonic. Thus the secret of harmonicity lies within the design of the center line and its cross-sectional change. So, the 'chrump' has an analogous effect like the loading of the Indian drums which also successfully converge an inharmonic sequence of tones into a harmonic one [RAMAN (1920 and 1934)].

ACKNOWLEDGEMENTS

I would like to thank the Proff. A. Krokstad and U. Kristiansen (both NTH, Trondheim) for fruitful discussions and advice.

REFERENCES

Euler, L. (1759/66): Supplément aux continuation des recherches sur la propagation
du son, Mem. Acad. Sci. Berlin 15, 210
Theoretical acoustics and numerical techniques, Springer, Wien
Computer simulations and application of numerical techniques
for acoustic waves in curved Swiss horn, Ph.D. thesis, NTH,
Trondheim, ISBN 82-7119-652-9

Finite element simulation of conical waveguide, Proceedings of
ICA 95, Trondheim, June 26-30, Norway

Musical acoustics, 2nd edition, Pacific Grove, California

Lagrange, J.L. (1761):
Nouvelles recherches sur la nature et la propagation du son, in:
Serret, J.-A. (ed.). Oeuvres de Lagrange, tome 1, p. 151-316,
Gauthier-Villars, Paris

Raman, C.V. (1920):

Raman, C.V. (1934):
The Indian musical drums, Proceedings of the Indian Academy
of Sciences, A1, p. 179-188

Van der Maas, Jan (1985):
Das monochord, Das instrument des harmonikers, Schriften über
harmonik, nr. 9, Berne
FINITE ELEMENT SIMULATION OF CONICAL WAVEGUIDE

R.-F, Fehlman
Swiss Engineering School, CAD/CAM Department, Bern, Switzerland
* on the leave from

SUMMARY

This paper deals with the simulation of an axi-symmetric conical horn with circular cross section by the finite element method (FEM). A low, mid and high frequency model has been constructed. The resulting resonance modes are presented in form of the overlap of these models. The calculations show that the resonance modes deviate to a large extent from desired harmonicity.

INTRODUCTION

The dawn of analytical treatment of the cone goes back to a musical interest and has led to the B.E.L. [BERNOULLI (1740/62), EULER (1759/66), LAGRANGE (1761)] horn wave equation.

\[
\frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} - \frac{\partial p}{\partial z} = \int S(z) \frac{\partial p}{\partial z} \, dz
\]

(1)

In the evolution of the analytical treatment of axisymmetric conical ducts with circular cross sections \(S(z) = S_o z^2\) (for variable definition, see fig. 1), DUHAMEL (1849) and BOUASSE (1929) give the solution for this waveguide for the three boundary condition (BC) cases of closed-closed (Neumann), open-open (Dirichlet) and closed-open (Cauchy) waveguide

Fig.1: Variable definition and geometry of conical waveguide.

as the following natural frequencies (eigenmodes) of the idealized (no end corrections) system:

- 423 -
\[ \text{ccN: } \tan(k\Delta R) = \frac{k\Delta R}{1+2\kappa \Delta R} \]  
(2.1)

\[ \text{ooD: } k_s = \frac{n\pi}{\Delta R} \Rightarrow f_s = \frac{nc}{2L} \quad \text{(for } \Delta R = \Delta z = L) \]  
(2.2)

\[ \text{coC: } \tan(k\Delta R) = kR_{\text{in}}, \quad \tan(k\Delta R) = -kR_{\text{out}} \]  
(2.3)

depending on closed or open at \( R_0 \) or \( R_{\text{in}} \), respectively. For the complete cone, i.e. \( R_0 = 0 \Rightarrow \tan kR_{\text{in}} = -kR_{\text{in}} \).

For the musically non interesting ccN-case RAYLEIGH (1896) gives an approximation:

\[ \tan(k\Delta R) = \frac{k\Delta R}{1 + \frac{1}{4}k^2(\Sigma R)^2 - (\Delta R)^2} \]  
(2.1.1)

Assuming \( \Sigma R = R_0 + R_{\text{out}} \gg \Delta R = R_{\text{in}} - R_0 \), then

\[ f_s = \frac{nc}{2\Delta R \left(1 - \frac{4(\Delta R)^2}{n^2\pi^2(\Sigma R)^2}\right)} = \frac{nc}{2L} \quad \text{(for } \Delta R = \Delta z = L) \]  
(2.1.2)

In other words the influence of conicity upon pitch is of 2\textsuperscript{nd} order. Thus, within first order approximation, the ccN- and ooD BC's yield the cone to have harmonic eigenmodes and are therefore half-wavelength resonators.

The interesting case for musical horns (coC) has no analytical solution. It is this case which is studied in the following.

**RESULTS FROM THE NUMERICAL COMPUTATION OF THE RESONANCE MODES OF THE CONE BY THE FINITE ELEMENT METHOD**

For the FEM-calculations the cone has been assumed to be a lossless system in the stationary state with an input according to a driving frequency of a spherical piston with a time-harmonic vibrating velocity \( u \) such that

\[ u_s = u_0 e^{j\omega t} \]  
(3)

with \( u_0 = 1 \) [m/s]. In the stationary state one part of the energy is dissipated out into free space whereas the other part is reflected back to the mouth. We assume a steady state such that the
average power being supplied by the driving force is just equal to the power being dissipated by radiation through the horn's aperture. The horn walls are thus assumed to be rigid and therefore have a boundary impedance $Z_{\text{bound}} \to \infty$.

In order to determine the resonance frequencies of the system, a model with a radiation space had to be constructed [see FEHLMANN (1994 or 1995)]. Thus the waveguide had to be extended by introducing a 'hypothetical' spherical boundary with a characteristic free space impedance. In this way resonance modes could be calculated taking into account the sympathetically vibrating air mass at the discontinuous mouth opening of the conical horn. Since it is not known where to stop or to set a termination the following reasonable assumption about the wave behaviour has been made: The free space part is modelled up to a distance $R_{\text{rad}} = R_b$ - usually called the Rayleigh distance - defining a hemisphere, i.e. in the 2D case a hemicircle. The well-known complex wave impedance for a spherical acoustic field is given by

$$Z_{\text{sp}} = \frac{q \lambda (2\pi \frac{R}{\lambda})}{1 + j2\pi \frac{R}{\lambda}}$$

which depends on the ratio between the distance $R$ from the source and the wavelength $\lambda$, i.e. $R/\lambda$. But as we can see already at $R=\lambda$, the spherical wave impedance and the plane real wave (or free space) impedance differ only by about 2.25% which gives us a sensible argument to set $R=R_{\text{rad}} = R_b$ and thus $Z_{\text{sp}} = Z_{\text{p}} = q \lambda$. It is therefore supposed that the radiation field starting at the aperture's midpoint is a monopole source producing spherical wavefronts into the free space until they develop into plane waves as long as we move far enough out or if we consider them just locally. A small inaccuracy is introduced by assuming this distance to be at $R = \lambda$ and not at $R \to \infty$.

Sweeping through all the frequencies between 0-1600 [Hz] (corresponding to 32 modes) with a 1 [Hz] resolution gives then the resulting resonance spectrum. The chosen resolution is reasonable in the light of the JND (just noticeable difference) which is equal to about 1 [Hz] up to 5 [kHz] for sine waves whereas for complex waveforms we have JND = 0.1 [Hz] (see [HALL, 1990]). Precision considerations concerning the sufficient number of elements per wavelength when modelling the reader is referred to KRISTIANSEN (1975) and FEHLMANN (1994). The following four diagrams are represented in fig. 2:

i) Peaks and dips:

Calculation of the peaks and dips from the pressure response give on one hand the resonance frequencies and on the other hand the frequency dips of the system. Representing the envelope of these peaks and dips yields the tendency of the resonances and the anti-resonances, respectively, relative to the characteristic impedance $q \lambda$ drawn as a straight line in the plots at 411.4 [Pa].

ii) Harmonic vs. inharmonic modes:

In case of having harmonic resonances (the frequency of mode nr. $n$ is $f_n = nf_1$, where $f_1$ is the fundamental frequency) the impedance peaks are equally spaced in the frequency
domain. Therefore, plotting the mode frequency as a function of mode number we get a straight line (the harmonic line). Inharmonic resonances deviate then more or less from this line; thus the diagram is a visual measure of the harmonicity of the resonance peaks being important for any horn designer [SPADA (1990), STOESSEL (1977)].

ii) **Absolute harmonicity:**

The absolute deviation of the resonance modes tells us how many [Hz] the horn’s resonance modes are off from the ideal.

iv) **Relative harmonicity:**

This plot is a more practical plot from a horn manufacturers viewpoint because it tells how much each resonance has to be corrected relatively. So we can take a worst case analysis and make a step by step correction.

**Fig. 2: Resonance modes and their inharmonicity of a conical waveguide**

**CONCLUSIONS**

Besides having used the FEM as a practical method for the solution of non-analytical problems in acoustics, there is set up graphs for the musical instruments manufacturer determining the harmonicity of an instrument. From the calculations presented here it can be concluded that the axisymmetric cone with circular cross-section is - in contrast to the usual belief - not a horn shape producing harmonic resonance modes.

**ACKNOWLEDGEMENTS**

I would like to thank the Proff. A. Krokstad and U. Kristiansen (both NTH, Trondheim) for fruitful discussions and advice.
COMPUTER OPERATED STRING INVESTIGATION UNDER MUSICAL ASPECTS

Helmut Frank, Wolfgang Vogl, Christoph Muss, Horst Wagner
Thomastik-Infeld Musik Strings, Research & Development, Vienna, Austria.

SUMMARY

This paper is about a testing device in which strings can be set in motion under precisely defined and reproducible test conditions. A sound recorder transfers these vibrations via an analog-digital converter into a "hard disk recording system", where they are stored in a sound data base for further analysis. In this data base material data of raw materials, in addition to the acoustic information, is to be found, as well as statistically attained data on subjective sound reception by musicians. All of this is stored, which allows to evaluate a sound event according to the latest psychoacoustic insights.

GENERAL

Empirical knowledge about the use of materials such as steel, tungsten, silver, gold and different synthetics, that has been derived in the course of decades, and the constructive use of these materials in manufacturing music strings will now be related to physical laws. By using modern signal processing technology and computer techniques, it is possible to cast light on aspects of string sound that are mastered intuitively by musicians, but, on the other hand, could not be approached systematically by engineers. The hard- and software, based on an industry-compatible computer system equipped with signal processors, is an "Acoustic Work Station" to provide a platform for our research. The heart of the system is a testing device (monochord) in which strings can be set in motion under precisely defined and reproducible test conditions by a computer controlled bowing machine. The bridge force is picked up by a piezo transducer and recorded by a hard disk recording system. A multimedia data base containing datafiles is used for testing samples out of the production line and also for testing prototypes of new strings.

FIGURES

A. Blockdiagramm COSIMA Hardware

B. Test report of a DOMINANT 138m Viola G string
Abbildung 5: Thomastik - Infeld Netzwerk: Konstruktion / Forschung & Entwicklung
(mit freundlicher Genehmigung durch H.Frank)
Dominant 138m  
Viola  
Directory: L:Viola\g

14.02.1995 07:50:52  
Experiment ID 405  
KlangID 1231  
Nr: 138m  
Versuchdatum: 13-Feb-95  
Abstrich 20cm/s

Nennspannung (kp): 4.8  
Stimmspannung (kp): 4.99  
Streicherl (mm): 16

NennDurchmesser(mm): 0  
Durchmesser(mm): 0.74  
Auflagekraft (g): 70  
Strechgeschwindigkeit (cm/s): -20

NennDichte (mg/mm): 0  
Dichte (mg/mm): 2.300  
Temperature (°C): 20  
Luftfeuchtigkeit(%): 40

RMS Vergleich Saite-Norm

RMS (dB): -6.99  
Norm RMS (dB): -6.50 +/- 0.97

Spektrumvergleich Saite-Norm

Normiert auf die RMS

Einstimmverhalten

Einstimmzeit: 3.5 Stunden

Zeitsignal

Soll-Grundfrequenz: 196 Hz

- 429 -
CONCLUSIONS

With the help of analysis and the re-synthesis of sounds, which are then evaluated through listening by musicians, it is possible to test the specific use of materials and adapt them individually for each artist and instrument. Thus the results of numerous test runs are directly incorporated into the production of music strings, which are in turn tested by musicians in field experiments.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Werner Deutsch, head of the Department of Sound Research of the Austrian Academy of Sciences for his important comments on Psychoacoustics and numerous hours of discussions about the topic.

REFERENCES

Schumacher RT. Aperiodicity, Subharmonics, and Noise in Bowed String Instruments. Pittsburgh: Department of Physics, Carnegie Mellon University, 1993:
Schumacher RT. Measurements of Some Parameters of Bowing. Chaos 1994;
Schumacher RT, Woodhouse J. The transient behaviour of models of bowed-string motion. 1993; (UnPub)
THE KTH RULES FOR MUSICAL PERFORMANCE: OVERVIEW AND RECENT ADDITIONS

Anders Friberg, Johan Sundberg, Lars Frydén
KTH, Stockholm, Sweden.

INTRODUCTION

Interpretation is an essential part of music performance. It includes the transformation from a written score expressed in symbolic terms to a performance, expressed in terms of physical parameters such as duration and amplitude. This transformation has been investigated by means of the analysis-by-synthesis method, the result being a set of performance rules. These rules have been under development since 1976 and presently cover a wide range of aspects of music performance such as phrasing, punctuation, harmonic and melodic structure, micro timing, accents, intonation and final ritard.

The scheme for the score-to-performance transformation is shown in Fig. 1. The input is the score complemented by symbols reflecting the phrase and harmonic structures obtained by manual analysis. The analysis of melodic gesture is done automatically as described below. Then, the melodic and harmonic charges are computed. The former is used to quantify the tension between the tone and the underlying harmony. The latter reflects the tension between the harmonies and the key in tonal music. In atonal contexts chromatic charge replaces the harmonic and melodic charges. Finally, the performance rules execute the transformation to the physical domain, changing duration, amplitude, articulation, vibrato, etc. As indicated in the figure the input of this transformation is taken from the score, the analysis, the charge values and also by the $k$-values. These decide the overall magnitude of the effects induced by each rule. By varying these $k$-values, one can change a performance within musically acceptable limits.

![Diagram](image)

Fig. 1. Overview of the different parts translating a score to a performance.

The rule system is implemented in the Director Musices program (former Rulle). This program, written in Common Lisp, is running on a Macintosh computer and was specifically designed for the purpose of developing performance rules. In the bibliography, a partial list of our previous papers

- 431 -
covering the rules and related issues are given. This paper presents in preliminary form some of the recent work. Two new rules are presented, Punctuation and Phrasing, and one listening test measuring preferred k-value as function of tempo in the Inégaux rule.

PUNCTUATION

This is an attempt to automatically, from the score, identify the musical gestures and transform them to the performance. These gestures are melodic gestalts consisting of 1 and up to approximately 5-8 tones.

The gesture analysis part is roughly analogous to the grouping analysis at the lowest hierarchical level in the Lerdahl and Jackendoff theory (1983), although our analysis includes also the lowest gestalt level which may consist of one single tone. Our purpose is a little different from theirs. The purpose of the Punctuation rule is to find tone gestalts at the end of which it is appropriate to insert a micropause, while the aim of their theory was to describe the perceived musical structure. Lerdahl and Jackendoff give some rules for how this segmentation can be realized, but these are not sufficiently explicit for our application; no quantification is given and some rules assume a real performance as the input.

Our musical gesture analysis was developed from principles used in the previous rules Leap articulation and Accents. The analysis-by-synthesis method has successively increased the sophistication of the analysis. It is currently a complete, self contained rule system with totally 14 finder or eliminator rules (These rules are protected by a patent). Finder rules mark potential positions of boundaries between musical gestures. Eliminator rules indicate positions where boundary markers should not appear or modifies the interaction between different finder rules. The finder rules use weight values to estimate the importance of the inserted boundary mark.

The main principles for the finder rules are: (1) in melodic leaps, with different weights for different contexts, (2) after longest of five, (3) after appoggiatura, (4) before a note surrounded by longer notes, and (5) after a note followed by two or more shorter notes of equal duration.

The eliminator rules remove the marks or reduce the weights in the following cases: (1) after very short notes, (2) in melodic step motion, (3) when several duration rules interact and (4) at two adjoining marks in a tone repetition. A real boundary is assumed to exist if the sum of the weights in that position exceeds a certain percentage of the total average of all inserted weight values.

These boundary marks are introduced in the performance by transforming them to micropauses plus lengthenings of the preceding tone. The duration of the micropause and of the magnitude of the lengthening are proportional to the preceding note's duration. The weight values are not taken into account in this translation. A graph of the length and position of the resulting micropauses are shown in Fig. 2.

The system has been developed and tested on 60 different melodic excerpts mostly Western classical music but also including folk, jazz and contemporary atonal music. According to the judgment of our professional musician (coauthor LF), about 90% of the resulting marks occur in musically appropriate positions.

![Fig. 2. Micropauses produced by the Punctuation rule. The graph shows the length of the micropause in ms, expressed as the offset of the note below the graph to the onset of next note. The music excerpt is taken from the string quartet in A minor by F. Berwald.](image)

PHRASING

Arch-like shapes of tempo as a function of nominal time have been observed in many investigations (e.g. Sundberg et al., 1995; Gabrielsson, 1987; Repp, 1992). However, the appearance of these
arches often varies considerably between pieces played by the same performer and between performers playing the same piece. The exact shape can be hard to determine but in modeling the timing deviations of one phrase averaged over many performers Repp (1992) obtained a good fit with a quadratic curve.

The design of the rule has been inspired by Todd's phrasing model (1985), later revised (Todd, 1989). He used the time-span reduction (Lerdahl & Jackendoff 1983) as the input and applied quadratic curves using the hierarchical level of the phrase ending to determine the amount of the ritard. Todd's model works quite well for many examples but was not developed for "the prediction of individual performances as such, but the principled explanation of performance data". For a generally applicable rule, however, a more flexible model is needed. While the parameters of Todd's model are determined by the time-span reduction, our parameters can be varied by the user. Our goal was a model of phrasing which was capable of describing the major part of the variety of phrasing occurring in real performances.

The rule is currently implemented in the following ways: (1) Marks showing the hierarchical phrase structure are inserted by hand in the score. (2) The basic shape of the tempo curve for a phrase is quadratic, containing an accelerando immediately followed by a ritardando. (3) The turning point of the parabola (where the acc. ends and the rit. starts) and the end-value of the ritardando are adjustable. (4) Duration of the final note can be a fraction of the end-value it received according to the ritardando. (5) The rule can be applied to different hierarchical levels independently. (6) The end-value of the ritardando can be set higher for endings on higher structural levels. (7) The loudness varies in proportion to the tempo according to an adjustable factor. Fig. 3 shows an example from Sundberg et al (1995) where the phrasing rule is fitted to two melodies sung by a professional Swedish baritone singer. In the example, the rule Duration contrast has also been applied with a negative quantity, i.e. lengthening of short notes and shortening of long notes.

![Fig. 3. Two applications of the phrasing rule. The straight lines are two performances by the singer. The dotted lines indicate the matching of the Phrasing rule. In these examples, Duration contrast was also applied with a negative k-value. The music excerpts were taken from Dichterliebe VI, bars 21-29 by R. Schumann (left) and Paulus, Aria #18, bars 5-11 by F. Mendelssohn (right).](image)

The phrase rule gives a rather good fit in many cases but it is also evident that refinements can be done. Most notably is the amplitude that probably has to be more independent of the timing. According to Gabrielson (1987), the amplitude tends to start with a crescendo and to end with a diminuendo but not necessarily in phase with the timing deviations. The common practice to insert a micropause after each phrase is not included in this rule since it is taken care of in the Punctuation rule described above.

PREFERRED K-VALUES IN THE INÉGALES RULE

Equally long notes are sometimes realized as pairs of long+short tones. In Baroque music this phenomenon was referred to as "notes inégales" while in jazz it is sometimes called the swing-factor. The rule inégales states that any tone onset occurring in unstressed position is delayed while the preceding tone occurring in stressed position is lengthened by the same amount of ms. The definition of the rule is found in Friberg (1991). The ratio between the lengthening and shortening is regulated by the quantity k. For k = 1, this ratio is 1.56 : 1. An experiment was carried out to determine preferred k-values at different tempi.
A total of 34 subjects of varied musical background were asked to adjust the k-value until the best possible swing-feel was obtained. This was repeated 2 times at 5 different tempi. The stimulus, the first 8 bars of Yardbird Suite by Charlie Parker, was performed on a synthesizer with electric organ, bass and drums.

The result (Fig. 4) reveals an approximately linear dependence of tempo as measured in quarter notes per minute (MM). Since the swing-factor is a ratio and not constant this means that a jazz performance is not invariant under tempo transposition, i.e. it can not be transposed in tempo by multiplying all durations with a constant factor. Another remark is that since the swing-factor is a linear function of MM, the use of MM instead of beat duration might be more appropriate for describing tempo variations in music.

![Graph](image)

Fig. 4. Preferred k-values as a function of tempo for the lnégales rule. The left graph displays k as a function of beats per minute and the right graph k as a function of quarter note interonset duration. The bars indicate the 95% confidence interval.

BIBLIOGRAPHY AND REFERENCES


SOUNDS OVERLAPPING IN MELODY PLAYED ON A PIANO

Alexander Galembo
Sechenov Institute, Laboratory N32, Russian Academy of Sciences, prospect Torenza 44, 194223 St. Petersburg, Russia.

SUMMARY
It is well known that notes in piano passages overlap, especially when played legato. Experiments show that this overlapping can be much higher than might be expected without interfering with the perception of pitch. Some piano sounds, which are perceived as having a distinctive and clear pitch, are shown to consist of a mixture of almost equally strong components from the preceding and new note, respectively. It is concluded that the phenomenon probably is connected with the particular quality of pianists usually referred to as "the pianist's touch."

INTRODUCTION
The presence of overlapping tones in a melody is not a surprising fact, it is one of the important properties of performed music. The overlapping is a part of the performer's interpretation, in which one parameter controls the *staccato-legato* dimension. The degree of this overlapping seems to be influenced by instrument design and room reverberation, among other things. In piano performance, the damper does not interrupt the tone immediately, but just increases the decay rate. This additional damping is characterized by a noticeable frequency selectivity set by the design of the damper in combination with the position of the damper along the string [1].

When discussing the human element in piano tone production, it is sometimes mentioned that the pianist is able to change the character of the tones by controlling the moment when the dampers touch the strings. This seems to be a strictly theoretical conclusion as no experimental results related to problem can be found in the literature. Also, the general opinion seems to be that overlapping of tones is not significant, because a melody played on the piano is perceived as having clear pitches.

EXPERIMENTS
If one listens carefully to melodic *legato*-phrases played on the piano and try to separate the notes in the phrase (or follow a single note to complete decay), one could notice something strange in the perceived pitch of some sounds. This observation, which could be described as a "dirty" pitch, could be assumed to be due to the overlapping of tones. A more striking way of detecting the presence of overlapping sounds is to use artificial listening conditions, for example by listening to a passage reversed in time, or reproducing the phrase slower.

In this experiment a phrase in the middle range of a piano (G₂ - A₄ - B₂ - C₅ - B₄ - A₂ - G₃) played legato *mezzo forte* with a rate of about 5 tones per second was recorded for detailed analysis.
Listening to the passage confirmed that there was
- a distinctive and clear pitch sensation for every note in the phrase
- no evident audible difference in the character of the same tone between ascending and descending passages, respectively.

Inspection of the waveform (see Fig. 1) showed that
- the notes are overlapping considerably. At the onset of a new note, the preceding note has not at all decayed completely, and in some cases even the second last note remains rather strong
- the repeated notes in the ascending and descending passages, respectively (A₄ = 2nd and 6th note, B₄ = 3rd and 5th note) show distinctly different envelopes. This is apparently due to the position in the phrase, as they interact with different preceding notes.

Sonagraph analysis of the phrase showed a striking degree of overlapping. For instance, the fundamental of the first tone was present all the way up to the fourth note in the phrase (see Fig. 2).

Selective listening at one note at a time by means of a sound editing program (thus separating each note from its predecessor) revealed a distinctive mixture of two notes of comparable loudness during overlap.

In short, the analysis showed that there is a considerable overlap between adjacent notes in a piano phrase, but that this overlapping is not perceived under normal listening conditions. Using a different wording, it seems fair to conclude that piano music is highly dependent on an auditory illusion.

DISCUSSION
No simple explanation of the observations seems to be at hand, but it is likely that the following factors could contribute to the effect.

Masking
A weak sound will be masked by a strong sound. There is some doubt against this explanation in this case, however, because the two tones can be of comparable intensity at the onset of the second note without causing an unpleasant beating mixture. Even an artificial reduction of the level of the attack part of the second note, so that the decay part of the first note and the onset of the second note are equalized in level at the point they intersect, does not influence the perception. The second note is still perceived with a clear, distinctive pitch, free from contamination of the preceding note.

"Windowing"
When listening selectively at a note by cutting it out from the context, the overlap was heard as a profound beating. This could be due to a "windowing effect," in that we add a new fast "attack" to the tail of the decaying preceding note by applying the window. A hypothesis would be that this sudden onset increases the audibility of the decaying tail when it overlaps with the second note.
Also against this hypothesis there are some doubts. A sound file was prepared in which the notes were separated by silent intervals of 0, 3, 5, 8, 12, 20, 30, and 40 ms duration, respectively. For silent intervals shorter than 5 ms approximately, the melody was perceived with clear pitches of all notes, but for longer silent durations the mixture of tones during overlap was heard. Possibly, an adaptation process could be imagined at the high-level neural processing, by which tones following in a continuous sequence are processed differently than isolated notes, as discussed in the following paragraph.

**High-level auditory processing**

In general, our perception is more attentive to transients than to stationary processes, which means that we are more keen on new events than to permanent situations. The new tone might be masking the preceding note not by intensity, but just as a new event of the same modality, a new acoustical figure. The auditory processing has a limitation in time resolution, which could result in that fast changes in the signal (for example defined by the envelope of piano tone attack) are averaged and form an impression of a specific quality. In contrast, if the signal is changing slowly we can follow the envelope and extract details. This would explain why we hear overlapping pitches when a piano passage is reversed in time. This hypothesis also seems relevant to the experiment with interleaved intervals of silence described above.

**CONCLUSIONS**

Regardless of the mechanisms behind the auditory illusion of overlapping tones, it seems reasonable to conclude that the use of the dampers, either at the release of the key or by the use of the pedal, should be included in the discussion of the *individual touch* of pianists. Not only the beginning of the note, but also the termination is of importance to the perceived sound, due to the variable degree of overlapping. The relevant perceptual parameter in this case might be the "clearness of pitch" or *pitch strength* [2] during the overlapping sound segments, which, according to the reasoning above, to some extent would be under the control of the pianist.

**ACKNOWLEDGMENTS**

The author is indebted to Prof. Gunnar Fant for valuable discussions on auditory perception and to Anders Askfelt for kind assistance in the editing of the manuscript. This study was supported by the Wenner-Gren Center Foundation.

**REFERENCES**

Fig. 1. Microphone signal of a phrase in the middle range of a piano \((G_4 - A_4 - B_4 - C_5 - B_4 - A_4 - G_4)\) played legato at mezzo forte level with a rate of about 5 tones per second. Observe the marked overlap of notes, and the different beat processes for the same note in ascending and descending direction, respectively (2nd and 6th note, 3rd and 5th note).

Fig. 2. Sonagraphe analysis of the same passage as in Fig. 1, showing considerable overlap of tones, in particular for the lower partials.
SPECTRAL VISIBILITY OF TIMBRAL DIFFERENCES IN TWO REALIZATIONS OF A. SCHOENBERG'S ORCHESTRA PIECE OP. 16 NO.3

Alexandra Hettergott
Institut für Musikwissenschaft der Universität Wien,
Universitätstraße 7, Austria

SUMMARY

The paper focuses two realizations/orchestrations of Arnold Schoenberg's orchestra piece opus 16/3, called "Colours" (earlier "Chord Colouring"). About two years after the first transcription, the composer sketched this term when considering the so-called "tone-color melody" in his Harmony. [1] Depending on the same original version, the five pieces of opus 16 first were revised and later reduced for chamber orchestra. The varying scores and orchestrations thus yield to the pieces becoming different in texture – auditorily as well as visually, as presented by spectrograms. Arisen in this context, the question if the "continuous change of colours, rhythms, and impressions" [3] would remain unchanged in the realization by (reduced) chamber orchestra, could be answered in a negative way, since as compared with the complete orchestra's realization, in that by chamber orchestra the differing prominence of single instrumental elements leads to the predominant tone colour of several bars becoming changed. Moreover, in case of full orchestra it can be seen that the greater number of instrumentalists has not to necessarily result in a more intensive sound, but on the other hand appears more complex and spatial. Furthermore, as regards duration it could be stated that in each realization the 44 bars are performed in a musically, but not in a mathematically exact way.

INTRODUCTION

The opus 16 of Arnold Schoenberg consists of five orchestra pieces which are not connected; therefore the almost four minutes lasting third piece can be interpreted as a whole being musically autonomous. The original opus, composed in 1909 and revised by Schoenberg himself in 1922, provides about seventy instruments as opposed to 12 instruments (string quintet, single woodwind instruments, horn, harmonium, and piano) of the version reduced by F.Greissle in agreement with the composer (1949). [2a/b]

In this context the compositional setting of opus 16/3 has its very meaning, since on the cognitive level the notation and therefore the composition of the orchestra is connected with it, yielding to differences in the resulting sound; musical information, which partially can be seen with the visual scenes given by spectrograms. The texture visually representing the musical opus as a whole, allows to determine boundaries and, in a way, to deduce several peculiarities of the different sections. As for the musical model, for the structural (syntactical) analysis it is thus necessary to differentiate characteristics, which can be grasped visually/auditorily and which also lead to separate foreground elements from the background area.

---

1 S.TOOLS© by the Lab of Sound Research of the Austrian Academy of Sciences.
REALIZATION AND FINDINGS

Each realization of the third orchestra piece op.16 [2a/b] was segmented in bars along the belonging notation model, and verbally described by a peculiar profile. Furthermore, the comparison of the profiles of both realizations by instrumental elements and predominant chord colourings was supposed to confirm the hypothesis that the different importance of elements in both realizations would result in a different sound of the corresponding bar.

As instructed by the music score, the first eleven bars of Schoenberg's opus 16/3 represent a static and almost motionless surface: intendedly, there is no significant break in the texture. The differences are internal and subtle timbraⅢ, caused by the instrumental substance changing in the middle of almost any bar, where "the change of chords has to be so sensitive that not stress, but only a different colour becomes apparent." [2a] As follows, such units of one tone colour often last only the half of a bar: in the middle of a bar there mostly occurs a change of instruments on the same pitch or a change of tone colour, where the instruments (instrument groups) can take over from or and overlap with one another. By extension, by either exactly, or ± playing a half note, as given by a half plus an eighth note [+] or a quarter plus an eighth note [-].

The term "colours" refers to the peculiar timbres of single instruments (types) in so far, as they are the constitutive elements of the resulting homogeneous sound as a whole. On the one hand combined acting single instruments, mostly marked by "principal voice," at provided times are, frequently for reasons of momentary tension, intendedly contrasting with the background of the (more or less) even sound of the (chamber) orchestra. "Where an instrumental voice should become more prominent, it is organized in this way, and the sounds have not to be lessened." [2a] On the other hand also groups of heterogeneous instruments are synchronously acting to make a mostly short increasing or decreasing movement more striking to provoke and to underline moments of tension, but by that are also covering the timbre potentially typical for one certain instrument. Or different instruments (groups) alternate while sustaining the same tone, and thus prevent a certain instrumental timbre from becoming predominant. Obviously, this all is intended by the organization of the composition; in both versions the ensemble is heterogeneously composed of instruments of various types. But by means of the structure the substance of heterogeneous instruments is condensed to result in a homogenous and quasi-static sound.

For which static is suggested by sustained complex tone/chord clusters, particularly in wind instruments or harmonium (whereas sustained tones/chords in groups of string instruments serve mostly to voluminize sound), and the tone colour characteristic of mostly at the first half of a bar. This is most striking probably in case of triad tones, synchronously being played in parallel octaves by different instruments (groups). To momentarily increase tension on this predominantly static-homogeneous basis there are used several compositional methods of style, as a successively built seventh chord (bar nos.12/13), staccato-tones of the same pitch (16/17), crescendo/decrescendo (25-28), or a tremolous scale descent (28/29) or single tremolo-tones (40) in the string instruments.

2 Event analysis, with event being an occurring sound element contrasting with a quasi static-homogeneous sound surface. This was imagined by Schoenberg, when choosing the subtitle for his third piece, "Summer Morning At A Sea" (Transec in Austria): the scene, in which auditive single sound elements be distinguished within an otherwise homogeneous area, should visually resemble the quasi motionless surface of the sea, on which reflexes or motions become most striking (motion/shape, an effect of gestalt psychology).

3 As at the beginning, where one predominant tone colouring successively changes into another: bar no.3 A-colouring, bar no.4 (still) A, bar no.5 – B.
Figure: Long-time spectrograms of both versions of opus 16/3.

The epic and through-composed third orchestra piece can be said being grossly organized in three parts: The first one (bar nos.1-11) with a homogeneous and static structure, the eventful second one (bar nos.12-29), being marked by a quasi continuously increasing tension leading to a climax in bar no.29, and a third (bar nos.32-44), introduced by bar nos.30 (resting) and 31 (preparing), with again a predominantly static-homogeneous structure. Visually limited by $f_{\text{max}} = 5 \text{ kHz}$, the two long-time spectrograms represent both realizations of orchestra piece opus 16/3, including bars and sections. Most striking in the visually depicted musical scenes is the difference in the internal caesuras in bar nos.11/12 (fermata, indicated by "\n") between the 1st and 2nd, and in bar nos.30/31, between the 2nd and 3rd parts. The realization by complete orchestra is structured more noticeable into three parts than is that by chamber orchestra, doubtlessly dependent on a greater degree of dynamics and therefore also caused by interpretation. Furthermore, the massive tone clusters of the harmonium in the caesura bar nos.11 and particularly 30 (here plus piano) are also responsible for a less decreasing dynamics. The different conciseness of the internal caesuras shows also, that a more voluminous orchestra does not necessarily result in a more intensive sound.
Because of the greater number of instruments the realization of the complete orchestra yields to a higher degree of complexity and appears to be more spatial as compared with that by chamber orchestra. Correspondingly, events given by solo or synchronously acting instruments are of different importance within a sound composed by a chamber orchestra, single instrument timbres are more apparent than in the more complex and more spatial surface area of a full orchestra. As a percentage, within the chamber orchestra the single instruments take a greater part of the resulting sound than in case of the complete orchestra. Consequently, in the resulting sound of chamber orchestra dissonances also appear "more dissonant": within the small ensemble of twelve instruments a dissonant interval yields to the resulting sound becoming a dissonant timbre, whereas it only adds an additional colouring to the sound of an orchestra with about seventy instrumentalists.

CONCLUSIONS

As presumed, in conclusion it can be confirmed that "the continuous change of colours, rhythms, and impressions" [3] as postulated by Schoenberg, in the reduced version does not remain unchanged: in the realization by chamber orchestra partially single instrumental elements different from that of the complete orchestra version become important. This includes that in the smaller ensemble the colouring by dissonances seems to be more prominent and, as follows, the predominant tone colour of several bars changes, a fact that originally was probably not intended by the reduced version. Moreover, the overall impression of the piece changes, since the chamber sound with string quintet single woodwind instruments, horn, harmonium, and piano appears less spatial as compared with the sound of the complete orchestra. However, substantially the nature of the orchestra piece opus 16/3 does not change. Generally, the three-dimensional \((P,A,T)\) spectrogram is more likely to represent the multidimensionality of the perception of musical and sound space than purely does the music score. Visually depicting the concrete musical scenes the spectrogram shows both the texture of the movement as a whole (long-time spectrogram) and the structural peculiarities of single parts (microstructurally), where the density of informations in sound corresponds in a way to the spectrographical density. As visually in the spectrogram, auditive it is differentiated between contrasting sound events on the one hand, and the basic and even sound area on the other hand; this sound perspective is essentially, when perceiving sound space (aspect of gestalt). But admittedly, the cognitive perception of sound space cannot be adequately represented.

ACKNOWLEDGEMENTS

This investigation was supported by the Jubiläumsfond of the National Bank of Austria (project no.4862).

REFERENCES


[2] A. Schoenberg, "Fünf Orchester Stücke Op.16" (Five Orchestra Pieces),
[a]: Eulenburg No.1328 [original version, revised by the composer in 1922], ©1922 & renewed 1950 C.F.Peters [p. 31];
[b]: No.3828 L [Version for Chamber Orchestra by F.Creisle], ©1949 C.F.Peters: authorized copy of no.III, "Farben," (Colours) C.F.Peters, Frankfurt/M.


- 442 -
ABSOLUTE ANGLE AND JND OF AUDITORY SOURCE WIDTH FOR A MUSIC MOTIF

Kazuhiro Iida, Masayuki Morimoto*
AV&C Research Laboratory, Matsushita Communication Industrial Co., Ltd.,
Saedo, Tsuzuki, Yokohama 224, Japan;
* Environmental Acoustics Laboratory, Faculty of Engineering, Kobe University,
Nada, Kobe Japan.

SUMMARY

In this paper, following two topics relevant to the practical acoustical design of concert halls are investigated; (1) the absolute angle of auditory source width (ASW) for a music motif, and (2) jnd of DICC with regard to ASW of a music motif. Psychological experiments were carried out and an equation for estimation of the absolute angle of ASW as variables of DICC and BSPL was obtained. The equation indicates that the increase of DICC by 0.1 decreases ASW about 4° and the increase of BSPL by 1dB increases ASW about 1.6°. The measured jnd becomes smaller, as DICC becomes higher, and the Weber's ratio, ADICC/(1-DICC), is almost constant in the range from 0.2 to 0.3. Consequently, Weber's law is applicable to perception of ASW.

INTRODUCTION

The degree of interaural cross correlation (ICC), which is well known as a physical measure for auditory source width (ASW), depends not only on a room impulse response, but also on a source signal and a head-related impulse response. Therefore, it is impossible to discuss the usefulness of ICC without limiting source signals. In this paper, for putting DICC (degree of interaural cross correlation measured by using a dummy head without artificial ear simulators and without A-weighting), which is useful to evaluate ASW for music[1], to practical use as a single number physical measure for evaluation of ASW in concert halls, absolute angle and jnd of ASW for a music motif, which is same as that used in Barron and Marshall's experiment[2], are investigated.

ABSOLUTE ANGLE OF ASW AS A FUNCTION OF DICC AND BINAURAL SPL

The purpose of the experiments is to obtain an equation and a chart for estimation of the absolute angles of ASW as variables of DICC and the binaural summation of sound pressure level (BSPL)[3].

Method

The motif used in the experiments was a 6s section from bar 94 of the 4th movement of Mozart's "Jupiter" Symphony(No.41) recorded in an anechoic chamber. The sound field consists of a direct sound radiated from the front of a subject and two reflections radiated from the azimuth angles of ±45°. Reflection delays were 25 and 45ms. The sound pressure levels of reflections were made equal to each other.
DICC was set from 0.4 to 0.9 in step of 0.1. DICC was adjusted by controlling the energy ratio of reflections to a direct sound. The BSPL defined by Eq.(1)[3] was measured using the KEMAR dummy head without artificial ear simulators and was set from 50 to 80dBA in step of 10dB. In total, 24 kinds of stimuli (6 DICC × 4 BSPL) were used in the experiments.

\[ BSPL = 6 \log_{2}(2Ll/6 + 2lr/6) \]  

where \( Ll \) and \( lr \) are the sound pressure level at the left and the right ears, respectively.

Two kinds of psychological experiments were performed. In Experiment A, the relative values of ASW were obtained by paired comparison test. Each subject was tested individually and 20 times for each BSPL, while seated, with head fixed. The task of the subject was to judge which ASW is wider.

In experiment B, the absolute angles of ASW was judged quantitatively. Seventy-two light-emitting diodes (LED) were arranged at every 2.5° in the frontal semicircle of a subject on the horizontal plane including his aural axis. Eight of twenty-four stimuli were used in this experiment. Their DICC were 0.4 and 0.9, and their BSPL were 50, 60, 70 and 80dBA. The subject's task was to adjust the positions of the shining LED at the both horizontal ends of a sound image which he perceived. The angle in degree between two shining LED was regarded as the absolute angle of ASW. Each subject was tested individually and 20 times for each stimulus, while seated, with his head fixed. Five and seven male students acted as subjects for Experiment A and B, respectively. Five males acted as subjects for both experiments.

Results and discussion

In Experiment A, 100 responses to each pair (5 subjects × 20 times) were obtained in total. The psychological scales of ASW were obtained using Thurstone Case V model.

In Experiment B, 140 responses to each stimulus (7 subjects × 20 times) were obtained in total. The average value and the standard deviation (SD) of the absolute angle of ASW were calculated for each stimulus. For each BSPL, absolute angles of ASW for stimuli of DICC = 0.5, 0.6, 0.7, and 0.8 were estimated by sharing the difference between the average value of the absolute angle of ASW for DICC = 0.4 and that for DICC = 0.9 obtained in the Experiment B, according to the differences between psychological scales of ASW for stimuli of DICC = 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 obtained in Experiment A.

To obtain an equation and a chart for evaluation of the absolute angles of ASW in degree as variables of DICC and BSPL, the multiple regression analysis was applied using the data. The criterion variables was ASW and the predictor variables were DICC and BSPL. Equation (2) shows the multiple regression equation.

\[ ASW = -39.6X + 1.55Y - 31.9 \text{ (degree)} \]  

where \( X \) is DICC and \( Y \) is BSPL (dBA).

The correlation coefficient between the perceived ASW and the estimated ASW from Eq.(2) is 0.977. Consequently, it is clear that Eq.(2) is valid to estimate the absolute angle of ASW in degree. From this equation, the increase of DICC by 0.1 decreases ASW by about 4° and the
increase of BSPL by 1dB increases about 1.5°. An increase of DICC by 0.1 is equivalent to
decrease of BSPL by 2.6dB in ASW. The ratio +1.6/+1dB coincides with the result of Kect[4].
Figure 1 is the equal ASW contours for various DICC and BSPL obtained from Eq.(2).

jnd OF DICC AS A PHYSICAL MEASURE FOR ASW

An earlier study on jnd of the degree of interaural cross correlation used various noises as a source
signals[5]. But, now ASW for music motif cannot be estimated from the data for noises. jnd of
DICC as a physical measure for ASW of a music motif is measured as a criterion for practical
acoustical design.

Method

The psychological experiment was performed by the method of constant, using the paired
comparison method between the reference sound field with a fixed DICC and the comparison
sound fields with different DICC. The music motif and the structures of both the reference field
and comparison field are same as those used in the experiment on absolute angle of ASW
mentioned above.

DICC of comparison field were set at 14 steps from 0.38 to 0.68, at 14 steps from 0.52 to 0.84,
and at 12 steps from 0.84 to 0.94 for the reference fields with DICC=0.5, 0.7, and 0.9,
respectively. From preliminary tests, these ranges were found to cover the region where the subject
can easily discriminate ASW for two sound fields. The sound pressure levels of all fields were
constant at 70dBA, slow peak, measured at the left ear of the KEMAR dummy head without an
artificial ear simulator.

Paired comparison tests of ASW were carried out. A pair consisted of the reference field with a
fixed DICC and one of the comparison fields with different DICC. Each subject was tested 16
times for each pair. The task of the subject was to judge which ASW was wider. Eight male
students acted as subjects for the experiment.

Results and discussion

In total, 128 responses to each pair were obtained. For each reference field, the percentage of
responses for which ASW for the comparison field was wider than that for the reference field was
obtained. The distribution of responses to any reference field can be regarded as the normal one.
Then, the regression equation was obtained by the least square method. jnds were defined as DICC
at which the percentages were 75% and 25% and they were obtained from regression line for each
reference field. The results are shown in Table 1. jnds of ASW perceived to be wider(75%) and
narrower(25%) than ASW for the reference field are almost identical for any reference field. But as
DICC of the reference field becomes higher, jnd becomes smaller. This behavior looks like
Weber's law. Weber's ratios defined as Eq.(3) are also shown in Table 1.

\[ K = \frac{\Delta \text{DICC}}{1 - \text{DICC}}, \]

where \( K \) is Weber's ratio, DICC is degree of interaural cross correlation of a reference field, and
\( \Delta \text{DICC} \) is jnd. The ratios for all reference fields can be considered to be almost constant in the
range from 0.2 to 0.3. Consequently, Weber's law is applicable to the perception of ASW at least
in the region of DICC between 0.5 and 0.9.
Table 1. jnd of DICC and Weber's ratio with regard to ASW.

<table>
<thead>
<tr>
<th>DICC of reference field</th>
<th>jnd of DICC wider</th>
<th>jnd of DICC narrower</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASW = 40°</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>30°</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>20°</td>
<td>0.03</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Fig.1. Equal ASW contours for various DICC and BSPL.

CONCLUSIONS

Absolute angles of ASW and jnds of ASW were investigated, for a music motif. The results of psychological experiments lead to the conclusions that:

1. The absolute angle of ASW can be estimated by the following equation as functions of DICC and binaural sound pressure level (BSPL): 
   \[ ASW = -39.6X + 1.55Y - 31.9, \]
   where X is DICC and Y is BSPL(dBA).
2. Weber's law is applicable to the perception of ASW. The Weber's ratio as expressed by \( \frac{\Delta DICC}{(1-DICC)} \), is found to be almost constant in the range from 0.2 to 0.3, where DICC is for the reference field and \( \Delta DICC \) is jnd.

ACKNOWLEDGEMENTS

The authors would like to thank Mr. M. Matsuda and Mr. M. Kashiwa for their cooperation in the experiments.

REFERENCES

PHYSICAL MODELS OF MUSICAL INSTRUMENTS IN REAL-TIME BINAURAL ROOM SIMULATION

Jyri Huopanenl, Matti Karjalainen2, Vesa Välimäki1
1 Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Otakari 5 A, Finland.
2 Stanford University, CCRMA, Stanford, USA.
3 CARTES, Ahertajankuja 4, Espoo, Finland.

SUMMARY

We introduce methods for designing a binaural listening environment using computer-generated sounds. The system is capable of creating a natural and realistic spatial simulation in real time. This environment has been developed by combining high-quality model-based sound synthesis of plucked string and woodwind instruments with binaural room simulation using a multiprocessor DSP programming environment. A new method for approximating head-related transfer functions (HRTF) using cepstral smoothing is proposed. The system and the techniques presented here are intended for applications of virtual reality, multimedia, and music technology.

MODEL-BASED SOUND SYNTHESIS

Physical modeling of musical instrument families such as plucked string and woodwind instruments has evolved to a new sound synthesis technique, called model-based sound synthesis [1]. The goal in this synthesis method is to model the acoustic-mechanical principles found in the instrument and to apply modern digital signal processing (DSP) techniques that allow for real-time processing. We have applied physical modeling techniques to plucked string and woodwind instruments. An overview of the modeling techniques can be found in [2] and [3].

The directivity patterns of musical instruments can also be taken into account in model-based sound synthesis. In a companion paper [4] we have considered three methods suitable for real-time processing. The method that we call directional filtering uses low-order digital filters that simulate the directivity of the instrument model. This set of filters is applied according to the direction of interest at the output of the instrument model.

Modeling of the directivity patterns of musical instruments is attractive both from the research and application point of view. Virtual reality environments such as the system described in this paper are capable of simulating moving and rotating sound sources and producing real-time 3D audio output for headphone or loudspeaker reproduction.

REAL-TIME MODELING OF ROOM ACOUSTICS

The goal of real-time processing places several constraints on the amount of detail that can be achieved in the simulation of room acoustics. We have concentrated on applying the image-source method for approximating the early reflections in a rectangular room. The image-source method is based on the theory of using phantom images of the desired source position in a bounded space. The $M$th-order reflections of a sound source can be calculated by adding $N = L(L-1)^{M-1}$ images of the room with $L$ boundaries and calculating the length of the path between the two points.

The frequency-dependent attenuation of the sound radiating from the source and its images is caused by three major factors: 1) absorption of the boundary materials, 2) the frequency-depen-
dent air absorption, and 3) the attenuation due to the distance between the source and the listener. In DSP terms frequency-dependent filtering and attenuation can be accomplished with low-order digital filters. The image-source method that we have implemented assumes the modeled room to be rectangular and all the reflections to be specular.

An artificial reverberation model for approximating late reverberation in a room has been proposed in [2]. The fundamental idea is 1) to use the direct sound and image-source method for accurate modeling of the first reflections, and 2) to use outputs from the direct and imaged sources as inputs to the late reverberation filter. The late reverberation filter is a recursive structure consisting of comb filters and an allpass section. The combined image-source and late reverberation model simulates the impulse response of a rectangular room to the extent needed in realistic room simulation.

BINAURAL AURALIZATION

Auralization refers to processing of sounds in such a way that a three-dimensional illusion of a sound space in a room is created [5]. Modeling of human spatial hearing involves localization cues such as the interaural time difference (ITD) and the interaural amplitude difference (IAD). Frequency-dependent coloration due to reflections and diffraction from the torso and pinnae are modeled by measuring head-related transfer functions (HRTF). An HRTF represents a free-field transfer function from a fixed point in a space to the entrance of the test person's ear canal. In the real-time environment design we have used HRTFs based on measurements carried out on human subjects [2] and a dummy head [6].

Filter Design from Measured HRTF Data

It is necessary from the computational point of view to reduce the order of the HRTF filters when designing real-time simulation. Several methods for approximating measured HRTF data have been presented in the literature over the past few years. The methods have mainly concentrated on the design of finite-impulse response (FIR) filters, which can exactly preserve the phase behavior. This implies that the impulse response of, e.g., an HRTF measurement can be directly applied as an FIR filter. Suggested techniques for approximating HRTF magnitude have included the use of sliding third-octave window averaging [7] and principal components analysis (PCA) [8]. We have studied a method for reducing the order of the FIR filters using cepstral smoothing, which has been widely used in speech processing (see, e.g., [9]). The cepstrum \( c(n) \) of the measured HRTF impulse response \( h(n) \) has been defined by:

\[
c(n) = \mathcal{F}^{-1}\{\mathcal{F}\{\log|H(\omega)|\}\}
\]

where \( \mathcal{F} \) and \( \mathcal{F}^{-1} \) denote the operation of discrete Fourier transform (DFT) and inverse DFT, respectively, and \( H(\omega) \) is the frequency response of \( h(n) \). The cepstrum is multiplied by a window function (e.g., a Hanning window) to remove those peaks that correspond to rapid deviations in the frequency response.

\[
\hat{c}(n) = c(n)w(n), \quad \text{where} \quad w(n) = 0.5 - 0.5\cos(2\pi n/(2M + 1)), \quad -M \leq n \leq M
\]

Thereafter, the smoothed magnitude response is obtained by applying the DFT to the windowed cepstrum:

\[
\hat{C}(\omega) = \mathcal{F}\{\hat{c}(n)\}
\]

The smoothed HRTF impulse response \( g(n) \) is then calculated by canceling the logarithmic operation and by adding the original phase function \( \phi(\omega) \):

\[
g(n) = \mathcal{F}^{-1}\left\{10\hat{C}(\omega)e^{-j\phi(\omega)}\right\}
\]
Fig. 1 Magnitude response (upper left) and the impulse response (lower left) of the measured HRTF at 0° azimuth and 0° elevation, and the corresponding smoothed magnitude response (upper right) and the smoothed and windowed impulse response (lower right).

Shown in Fig. 1 are the magnitude spectrum and the impulse response of an original HRTF measurement (of a dummy head [6]) and the corresponding cepstral smoothed magnitude and impulse response. It can be clearly seen that the overall shape of the magnitude response is well preserved, but the ripple resulting from noise and anomalies in the measurement are removed. The phase behavior of the HRTF shows nearly a minimum-phase characteristic, in the example case of order 128 some 9 zeros were located outside the unit circle. However, the original phase function of the HRTF is retained as seen in Eq. (4) to ensure the best possible result in the simulation.

OVERVIEW OF THE SIMULATION SYSTEM

The designed simulation system is currently functional in the Apple Macintosh platform. The hardware is based on a multiprocessing environment consisting of Texas Instruments TMS320C40 signal processors [2]. An overview of the complete system is illustrated in Fig. 2. The current system contains two C40 processors, one being dedicated to the room and auralization modeling and the other for running the physical models. The room simulation system is operated by a mouse-controlled user interface and the physical models are controlled via MIDI. The binaural audio output is directed to headphone listening. The current sampling rate used is 22.05 kHz. We are experimenting on an 8x8-matrix A/D to a converter which enables processing for multiple loudspeaker auralization.

APPLICATIONS OF VIRTUAL AUDIO REALITY

The aim of this project has been to develop an environment that would serve both as an expandable simulation system for model-based sound synthesis and as a stand-alone room simulation system for real-time auralization. The combination of three-dimensional sound and model-based sound synthesis has attractions to several hardware and software applications: 1) virtual reality environments, 2) computer games, 3) computer sound cards, and 4) stand-alone synthesizers. The computational efficiency is a strong constraint and compromises have to be made to achieve real-time processing in affordable hardware environments. Nevertheless, virtual acoustic environments containing physical models of musical instruments as sound sources are a step towards virtual audio reality.
CONCLUSIONS

A real-time room simulating environment for model-based sound synthesizers has been proposed. The system produces real-time virtual three-dimensional audio output for stereo headphone listening. Radiation directivity of the musical instruments has been incorporated by applying direction-dependent low-order digital filters. Room simulation is performed with a combined image-source and late reverberation model. Binaural processing is carried out using digital filter approximations of measured HRTFs. The HRTF approximation is achieved by applying cepstral smoothing and windowing the smoothed impulse responses.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Unto K. Laine for his valuable comments on the use of cepstral smoothing in the reduction of the HRTF filter order. Mr. Tommi Huotilainen is acknowledged for the design and implementation of the hardware for the simulation system. This study was financed by the Academy of Finland.

REFERENCES

DIRECTION-DEPENDENT PHYSICAL MODELING OF MUSICAL INSTRUMENTS

Matti Karjalainen 1, Jyri Huopaniemi 1, Vesa Välimäki 1,2
1 Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Otakaari 5A, Espoo, Finland
2 CARTES, Ahertajamlija 4, Espoo, Finland
3 Stanford University, CCRMA, Stanford, USA

SUMMARY

Modeling the directional behavior of musical instruments has several attractions both in research and application of sound synthesis and room acoustics. It is known that due to the directional properties the tone quality of the instrument can vary remarkably as a function of direction. To our knowledge, directivity has not been incorporated before in model-based sound synthesis of musical instruments. We have designed a simulation system using physical models of plucked string instruments and wind instruments. Both real-time and non-real-time simulation environments include radiational characteristics of these musical instruments. As an example, we have measured and modeled the directional properties of the trumpet. Several methods for incorporating radiation directivity in sound synthesis models are introduced.

MODELING OF MUSICAL INSTRUMENTS USING DIGITAL WAVEGUIDES

The term physical modeling is used for simulation of acoustic-mechanical principles found in musical instruments. By means of physical models it is possible to simulate quite detailed effects of sound generation. Digital waveguide modeling employs digital filter representation of wave propagation [1]. This allows for real-time synthesis on modern signal processors. The physical modeling approach may be extended to include the directional characteristics of the musical instrument. This is of great interest, e.g., in room simulation and virtual reality environments, where the sound source (the physical model) can be moved in an acoustic space [2].

Model for Plucked String Instruments

The main elements of a plucked string instrument are as shown in Fig. 1. Each string is a distributed subsystem that starts to vibrate when excited (e.g., plucked). The strings are coupled to the body and may also interact with each other (sympathetic vibrations). The body or a soundboard is a complicated resonator that is needed for acoustic amplification, sound radiation, and coloring of the sound.

The general solution of the wave equation for a string is composed of two independent transversal waves traveling in opposite directions (see, e.g., [3]). At the string terminations the waves reflect back with inverted polarity and form standing waves. The losses in the system damp the

![Fig. 1. Model for a plucked string instrument.](image1)

![Fig. 2. Digital waveguide string model.](image2)
almost periodic vibration of the string. All losses and other linear non-idealities may be lumped to the termination and excitation or pickup points. The string itself is then described as an ideal lossless waveguide [1]. The system may be modeled using a pair of delay lines and a pair of termination filters as illustrated in Fig. 2.

A practical implementation is a digital waveguide with two digital filters which may often be combined into a single one—called the loop filter—and optional excitation and pickup filters. The lossless delay line in a waveguide filter can be implemented very efficiently by a circular buffer. The waveguide model for a plucked string instrument is a linear system consisting of the excitation, the string model, and a model for the body. Due to linearity, it is possible to change the order of the parts. In practice, we use a combined excitation that includes both the string excitation and the impulse response of the body [4] [5]. This method reduces the computational load by several orders of magnitude.

Model for Wind Instruments

A general waveguide model for wind instruments is illustrated in Fig. 3. It can be divided into linear and nonlinear parts. The linear part represents the bore of the instrument and the reflection from the open end of the bore or from the first open tone hole (in the case of woodwind instruments). The excitation model simulates the interaction of the pressure input and the wave that propagates in the bore. This part of the system includes a nonlinearity which is characteristic to each wind instrument family. The input signal \( e(n) \) can be a white noise sequence or a DC signal. The model includes two outputs, \( y_1(n) \) and \( y_2(n) \). The former corresponds to the sound that radiates from the mouthpiece and the latter to that radiated from the end of the bore.

MODELING THE DIRECTIVITY OF MUSICAL INSTRUMENTS

Plucked string instruments exhibit complex sound radiation patterns due to various reasons. The resonant mode frequencies of the instrument body account for most of the sound radiation (see, e.g., [3]). Each mode frequency of the body has a directivity pattern such as monopole, dipole, quadrupole, or their combination. The sound radiated from the vibrating strings, however, is weak and can be neglected in the simulation.

In wind instruments, particularly in the flute, the radiation properties are dominated by outstanding sound from various parts of the instrument (the embouchure hole, the finger holes, the bell). Another noticeable factor in the modeling of directivity is masking and reflection caused by the player of the instrument. Masking plays an important role in virtual environments where the listener and sound sources are freely moving in a space.

Detailed computational modeling of directivity patterns of musical instrument sound radiation is out of the capacity of real-time DSP sound synthesis. It is therefore necessary to find simplified models that are efficient from the signal processing point of view and as good as possible from the perceptual point of view. We have considered three different strategies [2]: 1) directional filtering, 2) a set of elementary sources, and 3) a direction-dependent excitation.

Directional Filtering

A set of direction-dependent digital filters may be attached to the output of the physical model as illustrated in Fig. 4a. The output of each filter represents the response of the instrument to a particular direction. This method was studied for the acoustic guitar (see [2]) and the trumpet. The trumpet measurement was carried out by exciting the instrument by an impulse sound source and
Fig. 4. Three methods for incorporating directivity into physical models. a) Directional filtering, b) a set of elementary sources, and c) a direction-dependent excitation.

by registering the reference response at 0° and the related response in various directions. The measured responses were fitted separately with first-order AR models: the transfer functions \( H_{\text{ref}}(z) \) and \( H_{\text{dir}}(z, \theta_i) \) were designed to match the frequency responses of the reference at 0° azimuth, and the directional response at azimuth angles \( \theta_i \) (for \( i = 1, 2, 3, \ldots, M \)), respectively. Pole-zero directivity filters \( R(z, \theta_i) \) were obtained by division of the transfer functions:

\[
R(z, \theta_i) = \frac{H_{\text{dir}}(z, \theta_i)}{H_{\text{ref}}(z)}, \quad i = 1, 2, 3, \ldots, M
\]

Figure 5 depicts the modeling of direction-dependent radiation of the trumpet (in the horizontal plane) relative to the main axis radiation. Shown in the figure are magnitude responses of first-order IIR filters at azimuth angles 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5° and 180°. The reference magnitude spectrum at 0° is assumed to be flat.

In Fig. 5, the lowpass characteristic of the filters is noticeably increased as the relative angle becomes greater. This result agrees well with the theory found in literature (e.g., [3], pp. 373–375). There are, however, some deviations from this trend. They can be caused by noise in the measurements or by nulls in the radiation pattern. Note that the model presented here includes the masking effect of the player. The reliability of the filter estimates may be increased by applying auditory smoothing to the responses before designing the filters. This is well motivated due to the critical band frequency resolution of the human hearing.

Set of Elementary Sources

The radiation pattern of a musical instrument may be approximated by a small number of elementary sources such as monopoles or dipoles. These sources are incorporated in the physical model and each of them produces an output signal \( y_i(n) \) as illustrated in Fig. 4b. This approach is
particularly well suited to woodwind instruments, where there are inherently two point sources of sound radiation, the embouchure hole and the first open tone hole. We have applied this method to the modeling of the flute as shown in Fig. 3 (see [6]).

**Direction-Dependent Excitation**

The directivity filtering may be included in the combined excitation as shown in Fig 4c. The same approach has been suggested by Smith [7] for inclusion of the early room response in a physical model. This method is useful when it is desired to synthesize the sound of an instrument at one direction only. However, this approach is inefficient when sound radiation to several directions is simulated. This is because each modeled direction requires an additional physical model. The considerations above as well as our experiments have shown that the directional filtering technique is normally the most efficient one. A first or second-order filter approximation is often a satisfactory solution in a real-time implementation.

**Application to Virtual Acoustic Reality**

The methods presented above are applicable to virtual acoustic environments, where physical models are used as sound sources. Modern room simulation and auralization systems are capable of dynamic source and listener position changes. Moving and rotating sources can be modeled by changing the filter parameters of the paths in a proper way (e.g., the Leslie effect of a rotating loudspeaker can be simulated). Methods for designing virtual acoustic environments are discussed in a companion paper [8].

**CONCLUSIONS**

In this paper we have presented methods for incorporating directional characteristics of musical instrument sound radiation to model-based sound synthesis. This has been achieved by measuring and analyzing acoustical instruments, and using DSP techniques. We found three different methods for retaining the directional radiation information during spatial sound synthesis: 1) directional filtering, 2) a set of elementary sources, and 3) a direction-dependent excitation. The results are useful, e.g., in the design and implementation of room simulation and virtual reality environments for physical models of musical instruments [8].

**REFERENCES**

HOW DOES THE BANDWIDTH OF A HARMONIC COMPLEX-TONE MASKER INFLUENCE ITS MASKING BEHAVIOR?

Armin Kohlrausch, Andreas Sander*
Institute for Perception Research (IPO), Eindhoven, The Netherlands
* Graduiertenkolleg Psychoakustik, Universität Oldenburg, Germany.

SUMMARY
In this contribution, we investigate the relation between masker bandwidth and masked thresholds for complex-tone maskers. In contrast to what is known from noise maskers, thresholds decrease for increasing masker bandwidths. Changes in threshold are observed, even if spectral components several Barks away from the signal frequency are added to the masker. These results support the conclusion from other experiments that the masking behaviour of complex-tone maskers is strongly dependent on the degree of masker envelope modulation.

INTRODUCTION
The fundamental concept of critical bands in auditory perception is supported by a variety of experimental paradigms. Masking experiments using noise maskers of constant spectrum level have shown that thresholds increase with increasing noise bandwidths up to a 'critical bandwidth' and remain constant for larger bandwidths. This led to a description of the auditory system as an array of overlapping bandpass filters (auditory filters) with an effective bandwidth corresponding to the critical bandwidth (Zwicker and Feldtkeller, 1967). The threshold of a sinusoid in a broadband noise masker is determined by the averaged noise power at the output of the auditory filter that provides the highest signal-to-noise ratio. Masker energy outside the critical band does little to enhance or degrade signal detection. In comparable experiments complex-tone maskers behave differently from noise maskers (Kohlrausch and Sander, 1995). In order to predict masked thresholds of a sinusoid in such maskers it is not sufficient to only know the masker intensity within a critical band, but one has to consider the temporal waveform after an appropriate filtering process as well. The experiments presented here investigate the relation between masker bandwidth and masked thresholds for complex-tone maskers.

METHOD
Masked thresholds were measured using an adaptive three-interval forced-choice (3IFC) procedure. The masker was always presented in three consecutive 320-ms intervals. In one randomly chosen interval, the 260-ms signal was added temporally centered with respect to the masker. All acoustic stimuli were generated digitally. After D/A conversion, they were presented diotically over headphones. Three subjects participated in the experiments. Further details of the method are described in Kohlrausch and Sander (1995).

The maskers used in this experiment were complex tones with equal-amplitude components of common fundamental $f_0$. The starting phases of all masker components were set to zero. The level of a single masker component was 50 dB SPL. The signal was at the same frequency
as the central masker component, and, if not stated explicitly, its starting phase was set to zero. Masker and signal were both gated with 10-ms raised-cosine ramps.

EXPERIMENT 1
In the first experiment masked thresholds were determined for a wide range of masker bandwidths. The spectra of the maskers were always chosen to be symmetrical around one of the four signal frequencies (550 Hz, 1100 Hz, 2200 Hz and 4400 Hz). The fundamental frequency of the maskers varied proportionally with the signal frequency. It was 12.5, 25, 50, and 100 Hz, respectively. At each signal frequency masked thresholds were determined for 10 different masker bandwidths. Each masker is characterized by its relative bandwidth, which was calculated by dividing masker bandwidth by the center frequency of the masker.

The results of this experiment are shown in Fig. 1. The data points are given as median values and interquartile ranges of 12 individual threshold values (four values for each of the three listeners). The four different target frequencies are indicated by the different symbols: 550 Hz (diamonds), 1100 Hz (squares), 2200 Hz (crosses) and 4400 Hz (triangles). The masked thresholds are plotted as a function of the relative masker bandwidth and are expressed relative to the level of a single masker component.

The influence of masker bandwidth is qualitatively the same at all signal frequencies. At a relative bandwidth of .05 (masker consisting of three components) the curves have a local maximum. Towards larger bandwidths the curves decrease by between 10 dB (4400-Hz signal, triangles) and 40 dB (550-Hz signal, diamonds). Since the period of the masker is inversely related to the signal frequency, we can conclude that thresholds generally increase for decreasing durations of the masker period. Thresholds change even for large bandwidths that correspond to several Barks or ERBs. On the scale of relative bandwidths 1 ERB corresponds to about 0.12, and 1 Bark to about 0.2.

The following measurements served to investigate the non-monotonic behaviour of the curves at narrow bandwidths. The thresholds for one and three masker components differ on average by 8 dB. On the basis of the increase in overall masker level from one to three components one would expect a difference of only 4.7 dB. One source for this larger difference could be that only for the one-tone masker the fine structures of masker and signal are added coherently. This measurement was performed for a signal frequency of 1100 Hz.

In Fig. 2 the averaged thresholds of the three subjects for a signal with a starting phase of 90 degrees are shown by the diamonds. For comparison the averaged thresholds from the

Fig. 1: Masked thresholds of sinusoidal signals in harmonic complexes of various bandwidths. The four curves represent signal frequencies of 550 Hz, 1100 Hz, 2200 Hz and 4400 Hz (from bottom to top). Masker bandwidth varied symmetrically around the signal frequency and is expressed relative to this center frequency. Results of three subjects.
previous measurement for a zero-phase signal are included in the figure (squares). The change of the signal phase has the strongest influence on the thresholds for narrow bandwidths of the masker. The threshold for the one-component masker (data point to the left) increases from -18 dB to about -5 dB. This increase reflects the 90-degree phase shift between masker and signal. In order to yield a 1-dB level increment, the (incoherently added) signal must have a relative threshold of -6 dB. For a coherently added signal, a 1-dB level increment corresponds to a relative threshold of -18 dB. For wider maskers, the difference between the two signal phases decreases and for relative bandwidths of more than 0.5, the two curves do not differ significantly. This results suggests that for maskers with a high degree of masker envelope modulation, the thresholds are only influenced by the shape of the masker envelope and not by the waveform interaction between masker and signal.

For the three narrowest masker bandwidths, on the other hand, the thresholds are affected in a nontrivial way by the starting phase of the target. For these maskers, the specific fine-structure interactions between the signal and the masker have a significant influence on the signal threshold.

EXPERIMENT 2

The second experiment focuses on spectral properties of the complex-tone maskers. If we calculate a running short-time spectrum of a bandlimited harmonic complex with a sufficiently short temporal window, we see maxima at the spectral borders of the complex (Fig. 3). These edge frequencies can be perceived as a clear, sinusoid-like pitch superimposed on the fundamental pitch of the complex (Kohlransch and Houtsma, 1992). With varying bandwidth of the complex, the spectral distance between the signal and the two edge tones varies. The aim of the second experiment was to test whether the previously observed bandwidth effects reflect some sort of spectral masking by these edge tones.

Figure 3 shows short-time spectra of a harmonic complex at the center frequency 1100 Hz with a bandwidth of 1800 Hz (relative bandwidth of 1.6). The short pulses in the time functions of the sine phase masker are reflected in this representation as broad spectral maxima. Temporally interleaved with these broad maxima, the upper and lower edge tone of the complex emerge clearly. Distinct minima lie spectrally between the edge frequencies and temporally between the broadband maxima. With decreasing bandwidth the edge frequencies move closer together. So the corresponding increase of the masked thresholds with decreasing masker bandwidth (cf. Figs. 1 and 2) could be described as an increasing spectral
Fig. 3: Short-time spectra of a harmonic complex with $f_o = 25$ Hz and a center frequency of 1100 Hz. The edge frequencies are 200 and 2000 Hz. The spectra were calculated using a Hanning window of 5 ms.

masking effect caused by the two edge frequencies.
To test this consideration we measured the thresholds of the 1100-Hz target in two-tone maskers consisting only of the edge frequencies of the broadband maskers. These data are compared with the data of complex-tone maskers (fundamental frequency 25 Hz) consisting of all harmonics between (and including) the two edge components. The results for three subjects are shown in Fig. 4. Results for the two-tone maskers are indicated by the diamonds, those for the bandpass complex-tone maskers by the squares. For narrow bandwidths the thresholds for the two masker types are nearly identical. Here the spectral overlap of the edge-tone ramps seems to be the dominant effect. For wide bandwidths the thresholds for the complex-tone masker are higher than those for the two-tone masker. This result indicates the increasing role of temporal, i.e., forward masking for the harmonic maskers with a highly-modulated temporal waveform. Such temporal masking effects are absent for the two-tone maskers and therefore, these maskers lead to lower thresholds than the complex tones for large spectral bandwidths.

Fig. 4: Masked thresholds of an 1100-Hz signal as a function of the relative bandwidth of the masker. Squares indicate data for a complex-tone masker with $f_o = 25$ Hz. Diamonds indicate data for a two-tone masker. Results of three subjects.

REFERENCES
THE PERCEPTION OF RHYTHMIC STRUCTURE IN EXPRESSIVE MUSICAL PERFORMANCE

Neil P. McAngus Todd, Eric F. Clarke*
Department of Psychology, University of Manchester, Manchester, UK
* Department of Music, University of Sheffield, Sheffield, UK

SUMMARY

In this paper we discuss two algorithms which model the performance and perception of musical expression. The first algorithm takes as input a hierarchical phrase structure, which may be represented as a tree, and generates as output a sequence of timing and dynamic data which may be used to synthesise an expressive piano performance. The second algorithm as input a sampled sound signal and generates a tree-like pattern which is referred to as a rhythmogram. As an example we demonstrate a synthetic performance of the first 16 bars of Schubert’s *Impromptu* in B-Flat Major and the recovery of the input structure by the rhythmogram algorithm.

INTRODUCTION

In general one may discuss expression in musical performance from three different viewpoints (Gabrielsson, 1982): structural, motional and emotional. The structural view emphasises the idea that a major function of musical expression is the communication of the performer’s structural interpretation. This view has been the most influential over the last two decades or so and has resulted in a number of systems or algorithms for the synthesis of expressive variation in timing and dynamics (Clynes, 1987; Sundberg, 1988; Todd, 1992; Longuet-Higgins, 1994). The motional view emphasises the idea that the subtle variations in timing and dynamics are related to physical and biological movement (Clynes, 1987; Todd, 1992, 1995). The third viewpoint sees the function of expression to be the communication of emotional meaning (Clynes, 1987).

Although at first glance apparently incompatible, it is possible to combine two or more of these views into a single view. For example, Clynes takes the view that motions have a particular dynamic temporal form which more often than not may be manifested in the form of a movement. Todd (1992) has proposed a model of expressive timing and dynamics in which takes a structural description of a piece of music as input but any particular *accelerando* or *crescendo* is constrained to vary as a natural motion.

Whilst research continues on the exact nature of these structural/motional phenomena (Repp, 1992) and their possible emotional implications, until recently little or no work had been done on attempting to model the recovery of structure from the expressive performance. In this paper we demonstrate just such a model of recovery of the structure (Todd, 1994; Todd and Brown, 1995) by applying it to a synthetic performance of the first 16 bars of Schubert’s *Impromptu* in B-flat for which the input is a binary phrase structure. The perceptual model is based on the idea that following transduction the central auditory system carries out a multi-scale decomposition of the auditory nerve response (Todd, 1994). The output of this model may be visualised in the form of a rhythmogram which strongly resembles the “trees” commonly used by music theorists to represent rhythmic structure (Longuet-Higgins and Lee, 1982; Leerdahl and Jackendoff, 1983). The rhythmogram algorithm has also been applied to speech rhythms (Todd and Brown, 1995).
PERFORMANCE

The performance algorithm (Todd, 1992) is based on two main principles. The first principle is that the main function of musical expression is the communication of the performers intended structural interpretation. For music of the romantic era in particular which involves a deep rubato, the dominant structural component appears to be the phrasing structure. Other musical styles may emphasise other structural components such as metre or melody. However, for the purposes of this paper we shall restrict ourselves to phrasing. The second principle is that a musical phrase is indicated by an accelerating/ritardando shape in tempo and a corresponding crescendo/diminuendo shape in the dynamics. Further, this shape has the form of a single motion or gesture so that a complete performance may be thought of as a hierarchically organised sequence of gestures.

![Binary phrase structure and resultant tempo profile](image)

Figure 1. (top) The binary phrase structure for the first 16 bars of the Schubert Impromptus. (bottom) The resultant tempo profile. The dynamics is linearly proportional to the tempo, i.e. "the faster the louder".

There has been much debate recently about the exact nature of these expressive gestures (Kronman and Sundberg, 1987; Repp, 1992; Todd, 1995). However, one view is that within a phrase tempo varies linearly with time (Longuet-Higgins, 1994) and thus has a kinematic similarity to that of a ballistic movement (Todd, 1995). As an example we demonstrate a synthetic performance of the first 16 bars of Schubert's Impromptu in B-flat. The input to the performance algorithm was a simple binary phrase structure as indicated in Figure 1 (top). The output, in the form of a tempo curve, is shown in Figure 1 (bottom). Note that for each of the structural levels the tempo/dynamic gesture is computed independently so that the resultant output is obtained by the linear superposition of these components.
The perceptual algorithm is a component in a recent auditory-motor theory of rhythm perception (Todd, 1994; Todd and Brown, 1995). This algorithm was inspired by the theory of edge detection in vision (Marr, 1982). The basic idea, in the case of vision, is that the optical signal is blurred over a range of spatial scales by a number of Gaussian low-pass filters. Edges are detected by looking for zero-crossings in the second derivative of the Gaussian which has a characteristic “Mexican Hat” shape. In the case of hearing it is possible to do a similar computation on a simulation of the auditory nerve response. The essential difference is that the low-pass filtering is done in time rather than space.

![Image of sound signal and rhythmogram]

**Figure 2.** (top) The sampled sound signal of the synthetic performance. (bottom) The rhythmogram output by the perceptual algorithm.

The output of such an analysis may be visualised in a number of ways. One way is to plot the output of each low-pass filter to form an “energy flux surface”. This form of visualisation tends to emphasise the motional view of expression. Another way is to plot zero-crossing points of the derivatives of the low-pass response. This can be done either in the energy/time plane or the time-constant/time plane. Such a representation, referred to as a rhythmogram, strongly resembles the “trees” used in both music theory and metrical phonology (Longuet-Higgins and Lee, 1982; Lerdahl and Jackendoff, 1983; Liberman and Prince, 1977). Indeed, the rhythmogram has also been shown to be effective in speech as well as music (Todd and Brown, 1995). Figure 2 (bottom) shows the rhythmogram of the synthetic performance of the Shubert Impromptu. The input to the rhythmogram algorithm was the sound sample as shown in Figure 2 (top). The Gaussian filters are logarithmically spaced at 12 per octave with time-constants ranging from 1-35 seconds. The rhythmogram clearly reflects the binary structure of the performance, at least down to the level of the two-bar phrase. Below the two bar phrase the melodic structure interferes with the simple binarity.
CONCLUSIONS

In this paper we have shown that it possible to synthesise a convincing artificially expressive performance on the following principles: (1) that the main function of expression is the communication of the performers structural/motional interpretation; (2) that in the case of music of the romantic era the dominant structural feature is the phrasing structure which may be represented in the form of a tree; (3) that associated with a phrase is a motional form which has an accelerando/ritardando shape in tempo and a corresponding crescendo/diminuendo shape in dynamics; (4) that the motional form may be modelled on natural biological and physical motions. We further showed that it is possible to recover such a structural/motional form from a signal and thus to model the perception of rhythm in expressive performance. The perceptual algorithm was based on the following main principles: (1) that following transduction in the peripheral auditory system the central auditory system carries out a multi-scale analysis of the auditory nerve response; (2) that the hierarchic motional forms associated with the phrasing structure are reflected in the multi-scale energy flux; (3) that the hierarchical structure may be recovered from the pattern of zero-crossings in the derivatives of the energy flux, which may be visualised in the form of a rhythmogram.

ACKNOWLEDGEMENTS

This work was carried out during a grant from the MRC

REFERENCES


THE ACOUSTICS OF THE ORCHESTRA

J. Meyer

The tone quality of an orchestra rests on the cooperation of a multitude of differing instruments which are used partly in solo fashion and partly gathered in groups; it is furthermore shaped by the spatial extension of the ensemble and by the influence of room acoustics. Acoustic measurement procedures permit an analysis and description of tonal characteristics of individual instruments in regards to sound power output, spectral composition, time dependent fine structure as well as the directivity of sound radiation. Appreciation of the full orchestra sound, on the other hand, requires consideration of sound transit times within the orchestral sound, as well as variations in wall reflection times for different instrument position. In addition to these considerations of physics principles, the effect of dynamics and masking, the chorus effect and the seating arrangement of the strings along with spectral and time dependent tone construction will be illustrated with the assistance of the orchestra. Sound examples are taken largely from the overture to the opera "Die Meistersinger von Nürnberg" by Richard Wagner. In conclusion the orchestra will perform the entire overture.
A "MASKED FREQUENCY SPECTRUM" OF A BROAD-BAND NOISE AND ITS TIMBRE

Kenji Ozawa, Yōiti Suzuki, Toshio Sone
Research Institute of Electrical Communication,
Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-77 Japan.

SUMMARY

Similarity in timbre among 15 broad-band noise stimuli obtained from a psychophysical experiment was analyzed with the method of multidimensional scaling, and a three-dimensional solution was derived. The relation between dissimilarities in timbre and differences in spectrum was considered. Consequently, it was suggested that the masked frequency spectrum rather than the physical frequency spectrum correlates with timbre.

INTRODUCTION

Timbre of a steady sound is closely related to its frequency spectrum [1–4]. For example, Plomp and Steeneken [1] indicated that dissimilarity of broad-band harmonic complex tones in timbre is correlated well with the difference in their frequency spectra. Bismarck [2, 3] revealed that "sharpness" is determined by the frequency location of the overall energy concentration of the spectrum. For broad-band noise, any other features than sharpness have not been reported.

In this paper, timbre of a broad-band noise is experimentally examined, and the relation between timbre and two kinds of frequency spectrum of the noise is considered.

PSYCHOPHYSICAL EXPERIMENT ON DISSIMILARITY IN TIMBRE

In the experiment, similarities in timbre among the 15 broad-band noise stimuli shown in Fig. 1 were examined. The lower cut-off frequencies were 200 Hz for all the stimuli, and the higher ones were one of three frequencies: 4, 6, or 8 kHz. Slopes of the spectral envelope of the stimuli Nos. from 1 to 3, from 4 to 6, from 7 to 9, and from 10 to 12 were 0, −3, −6, and −9 dB/octave, respectively. While outlines of the spectral envelopes of the stimuli Nos. from 13 to

Fig. 1 Physical spectra of the stimuli. Abscissa is relative level, and a grid represents 20 dB.
Fig. 2 Three–dimensional configuration of the stimuli derived from the experimental results. dB/octave, respectively. While outlines of the spectral envelopes of the stimuli Nos. from 13 to 15 were the same as that of No. 6, there were some notches; the notches were located at frequencies of (2n+1)*200 Hz, where n is an integer ranging from 1 to 19. To avoid the effects of difference in loudness on the results, the overall loudness of the stimuli was kept at 10 sone by means of ISO 532B [5].

The stimuli were presented diotically to a subject via a headphone (Yamaha YHD–3). Subjects were twenty university students with normal hearing. Similarity/dissimilarity in timbre for all possible pairs of the stimuli was rated in seven categories of the scale (Category 1: indistinguishable, Category 7: entirely different). Each combination was presented four times in random order to each subject, and the average of the judgments was employed as a subjective distance in the pair.

The obtained distances among the stimuli were analyzed by ALSCAL, a computation program for multidimensional scaling, to represent the stimuli as points in Euclidean space. The number of dimensions of this space was determined by evaluating the stress, which indicates goodness of fit. The three–dimensional solution shown in Fig. 2 was employed, where the stress was 13% (one dimension: 42%; two: 19%; four: 11%).

The points corresponding to the stimuli Nos. from 13 to 15 degenerated around the point of stimulus No. 6 on the I–II plane. The configuration is interpreted according to two features of the physical spectra, i.e., the slope of the spectral envelopes and the higher cut–off frequencies. These indicate that the I–II plane correlates with the outlines of the spectral envelopes of the stimuli. On the I–III plane, the configuration of the stimuli Nos. from 13 to 18 having notches in their spectral envelopes spreads along axis III. This suggests that axis III is correlated with the fine structure of the spectral envelopes of stimuli.

DISSIMILARITY IN TIMBRE AND DIFFERENCE IN PHYSICAL SPECTRA

As mentioned in the introduction, it is indicated that dissimilarity in timbre of stimuli having broad–band spectra is correlated with the difference in their frequency spectra [1]. The authors intended to test whether this is applicable to the present results. The difference, $d_{ij}$, between the spectra of the stimuli Nos. $i$ and $j$ is given by

$$d_{ij} = \sqrt{\frac{1}{k} \sum (L_{ik}-L_{jk})^2},$$  \hspace{1cm} (1)

where $L_{ik}$ is SPL (Sound Pressure Level) of tone $i$ in 1/3-octave frequency band $k$ [1].

The differences were also analyzed by ALSCAL, and a three–dimensional solution was obtained as shown in Fig. 3. The configuration on the I–II plane in Fig. 3 agrees with that in Fig. 2 to some extent. As to the I–III plane, however, a clear discrepancy between the configurations in Fig. 2 and 3 is seen; the points corresponding to the stimuli Nos. 13 to 15 also degenerate on the plane in Fig. 3. The spectral difference given by Eq. (1) cannot explain the timbre dissimilarity which originated from the difference in the fine structure of the spectral envelopes.
**Fig. 3** Configuration of the stimuli derived from the differences in the frequency spectra.

**DISSIMILARITY IN TIMBRE AND DIFFERENCE IN MASKED FREQUENCY SPECTRA**

It is assumed that some characteristics of the auditory system such as non-linear transmission, lateral inhibition, and band-pass filtering translate the physical frequency spectrum of a sound into an "internal spectrum," which is the description of a sound spectrum in the auditory pathway [6, 7]. The authors assume that the internal spectrum is then converted into a "subjective spectrum" in terms of subjective (psychological) magnitude. Since this "subjective spectrum" is assumed to be used as the input to the central stage of timbre perception, the timbre would be closely correlated with it rather than with the physical spectrum. Furthermore, the authors suppose that the subjective spectrum could be approximated by a "masked frequency spectrum," which is the frequency spectrum in terms of loudness of the narrow-band frequency components of the input signal partially masked by each other [8, 9].

For a complex tone consisting of a small number of components, the masked loudness of each component of the tone can be measured, and the masked frequency spectrum of the tone may be easily derived. For a broad-band sound, however, it is difficult to measure the masked loudness of each narrow-band frequency component. For a sound such as that used in this experiment, the authors proposed to estimate the masked frequency spectrum as the frequency characteristic of the masked loudness for every 1/3-octave band component by using the method of ISO 532B [9, 10]. In the estimation, an overlap in the excitation patterns caused by two frequency bands with different frequencies causes masking given in the higher frequency band by the lower frequency band, and the overlap should be subtracted from the original loudness of

**Fig. 4** Masked frequency spectra of the stimuli. Abscissa indicates masked loudness in a band, and a grid represents 1 sone.
Fig. 5 Configuration of the stimuli derived from the differences in masked frequency spectra.

The higher frequency band.

Figure 4 shows the masked frequency spectra of the stimuli. The relation between dissimilarities in timbre and differences in the masked frequency spectra shown there was considered. To apply Eq. (1) to the masked frequency spectra, the masked loudness of each band was converted into PL (Perceived Level) by the equation

$$ PL = 10 \log_{10} S + 40, $$

where $S$ is the masked loudness of each band. The spectral difference between the stimuli was obtained by substituting the PL for the SPL in Eq. (1). The differences were also analyzed by ALSACL, and the configuration shown in Fig. 5 was obtained. The configuration on the I–II plane in Fig. 5 agrees with that in Fig. 2 to a certain degree. On the I–III plane in Fig. 5, on the other hand, the points corresponding to the stimuli Nos. 13 to 15 are separated from those of Nos. 1 to 12; this implies that, although it is qualitative, the differences in the masked frequency spectra represent the timbre dissimilarities which originated from the differences not only in the outlines but also in the fine structures of the physical spectra. This is very interesting because the masked frequency spectrum was calculated from the 1/3-octave band SPLs that did not reflect the timbre dissimilarities induced by the fine structure of the spectral envelopes.

CONCLUSION

The relation between the frequency spectrum of a broad–band noise and its timbre was considered. The masked frequency spectrum of sound correlates better with its timbre than does the physical spectrum. Further quantitative consideration is required to obtain a better understanding of the relation.

ACKNOWLEDGEMENT This study was partially supported by a Grant–in–Aid for Scientific Research (No. 0670057) from the Ministry of Education, Science and Culture of Japan.

REFERENCES
MODELLING THE STRING-FINGER INTERACTION IN THE CLASSICAL GUITAR

Maria Pavlidou, Bernard E. Richardson
University of Wales College of Cardiff.

SUMMARY

The sound of a guitar is determined partly by its construction and partly through the player's control via the interaction between his fingertip and the string. This is obvious since the same instrument can sound very different in the hands of different guitarists. By modelling the interaction between the string and the player we are able to understand how the physical properties of the fingertip, such as stiffness and nail shape, affect the sound, as well as investigating the important role of friction between the string and fingernail.

For good sound production, it is essential for the guitarist to release the string in such a way that the vibrating string imparts a strong vertical force component at the bridge. The string is made to slide over the fingernail, a process which lasts for a few milliseconds. During this time, transverse and torsional travelling waves are created on the string, which alter the local forces at the plucking point and dynamically determine the subsequent movement of the contact point and the final spatial distribution and velocities of the string elements on their release. The model uses finite difference methods to solve the dynamic interaction of the string and fingertip. The computational method can also include movement of the end supports. Predictions of the duration of the interaction and the initial amplitude and velocity profile for the free string are shown for different finger parameters.

INTRODUCTION

The sound of an individual guitar is determined by the way the string is excited by the player. Different guitarists can produce different sounds on the same guitar because they alter the way the string is plucked and subsequently the initial conditions with which the string is released. The string–finger interaction lasts for approximately 0.1 sec; it starts when the finger touches the string and it ends when the string is released to vibrate freely. The initial amplitude and velocity distributions of the string upon release are determined by the properties of the finger and string, the construction of the body and the interaction process.

The role of friction between fingertip and string during the interaction is very similar to the role of friction in the string−bow interaction in the violin. The string element which touches the bow sticks and slips continuously on the bow hair, a process which is governed dynamically by the friction between them. McIntyre et al. [1]–[4] have modelled the string−bow interaction using the assumption that for a given bow speed and normal bow force, the friction is a function of the relative velocity of bow and string only. In our model the friction is treated in a very similar way.

The string−finger interaction also has similarities with the string−hammer interaction on the piano. Hall [5] [6] has investigated the string−hammer interaction by modelling the hammer as a mechanical system and by including the travelling waves created during the interaction. The existence of these waves in the piano is more profound since on their return to the striking point they bounce the hammer back and are responsible for any further contacts between string and hammer. In our model travelling waves are included and the finger is modelled as a mechanical system.

THE MODEL

Although a two−dimensional model is still under development, we present the theory for an one−dimensional system (Figure 1a). A flexible guitar string of linear density ρ, length l and tension T lies on the x-axis and transverse vibrations are restricted to the y-axis. The bridge and the nut of the guitar are fixed at the points (0, 0) and (l, 0); in the final model information will be included about the input admittance of

- 469 -
the body. The guitarist's fingertip touches the string at the plucking position \( p \) and along a width \( w \). The tension is assumed to remain constant during the process, so longitudinal waves are not created. For the one-dimensional model torsional waves are not included. The damping is assumed to be independent of frequency and is given by the temporal absorption coefficient \( \beta_t \), where \( \beta_t = 2/\tau_t \) and \( \tau_t \) is the decay time.

The finger is modelled as a mass \( m_f \) connected to a spring of known stiffness \( k_f \) and damping \( q_f \) which can move perpendicularly to the string's length along the \( y \)-axis. In the two-dimensional model the shape of the nail will also be included as a known function and the finger will be able to move on the plain \( (y,z) \). The displacement of the spring's end is assumed to be a known function \( y(t) \) which simulates the guitarist's control of his finger. The existence of the spring allows the fingertip's displacement \( y_f(t) \) to be influenced also by the string's movement since it can be different from \( y(t) \). On the surface of the mass \( m_f \) friction \( \Phi \) exists which is assumed to be a known function of the relative velocity between string element and fingertip.

![Diagram](image)

Figure 1: (a): One-dimensional model. (b): Relation between friction and relative velocity (after McIntyre and Woodhouse [1]).

The function which connects friction with relative velocity consists of two parts, one for the sticking process where the friction is static and can have any value between zero and maximum \( \Phi_{\text{max}} = \mu_s N \), and one for the slipping process where the friction is dynamic and its modulus is calculated from the relative velocity between fingertip and string element (Figure 1b). If \( y_s(x,t) \) is the string's displacement at position \( x \) and time \( t \), the relative velocity between fingertip and string's element is \( v_{rel}(t) = \frac{\partial y_s(x,t)}{\partial t} - \frac{dy_f(t)}{dt} \) and the friction is given by eq.(1):

\[
\Phi(t) = \begin{cases} 
\frac{N(-D\mu_s - \mu_d v_{rel}(t))}{v_{rel}(t) + D} & \text{for } v_{rel}(t) > 0 \\
\frac{N(D\mu_s - \mu_d v_{rel}(t))}{D - v_{rel}(t)} & \text{for } v_{rel}(t) < 0,
\end{cases}
\]

where \( \mu_s \) and \( \mu_d \) are the coefficients of static and dynamic friction and \( D \) gives the one asymptote \( x = D \). The other asymptote is \( y = N\mu_s \), where \( N \) is the force normal to the frictional surface.

At time \( t = 0 \) string and finger are both at rest and the fingertip touches the string between the points \( (p-w/2,0) \) and \( (p+w/2,0) \). The displacement of the spring's free end is \( y(0) = l_f \) where \( l_f \) is the natural length of the spring and the friction is \( \Phi(0) = 0 \). The interaction starts when the free end of the spring starts moving, the spring expands and the mass \( m_f \) starts moving too. For the points along the string, apart from the plucking point, the equation of motion is the wave equation eq.(2):

\[
\frac{\partial^2 y_s(x,t)}{\partial t^2} = c^2 \frac{\partial^2 y_s(x,t)}{\partial x^2} - \beta_s \frac{\partial y_s(x,t)}{\partial t},
\]

where \( c \) is the velocity of the transverse waves, equal to \( c = \sqrt{\ell/T} \). The equation of motion for the string element in contact with the fingertip is given by eq.(3):

\[
m_s \frac{\partial^2 y_s(p,t)}{\partial t^2} = T \frac{\partial^2 y_s(p,t)}{\partial x^2} + \Phi(t) - \beta_s m_s \frac{\partial y_s(p,t)}{\partial t},
\]

where \( m_s = c w_s \) is the string element mass. The equation of motion for the fingertip alone is given by eq.(4):

\[
m_f \frac{\partial^2 y_f(t)}{\partial t^2} = k_f(y_f(t) - y_f(t) - l_f) - \Phi(t) - q_f \frac{dy_f(t)}{dt}.
\]
\[
\begin{align*}
\alpha_1 &= (m_e + 2m_f - k_f \Delta t^2 - q_f \Delta t)/(m_e + m_f) & \alpha_9 &= -1 + \beta_3 \Delta t \\
\alpha_2 &= -(m_f + q_f \Delta t)/(m_e + m_f) & \alpha_{10} &= T \Delta t/(m_e \Delta x) \\
\alpha_3 &= (m_e \Delta x - 2T \Delta t^2 - \beta_3 m_e \Delta x \Delta t)/(\Delta x (m_e + m_f)) & \alpha_{11} &= \Delta t^2/m_e \\
\alpha_4 &= (m_e (1 + \beta_3 \Delta t)/(m_e + m_f)) & \alpha_{12} &= 2 - \Delta t/k_f/m_f - q_f \Delta t/m_f \\
\alpha_5 &= -q_f \Delta t/(m_e + m_f) & \alpha_{13} &= -1 + q_f \Delta t/m_f \\
\alpha_6 &= k_f \Delta t^2/(m_e + m_f) & \alpha_{14} &= 2T \Delta t/k_f/m_f \\
\alpha_7 &= -k_f \Delta t^2/(m_e + m_f) & \alpha_{15} &= -k_f \Delta t^2/m_f \\
\alpha_8 &= 2 - 2T \Delta t^2/(m_e \Delta x) - \beta_3 \Delta t & \alpha_{16} &= -\Delta t^2/m_f
\end{align*}
\]

Table 1: Computational constants.

**COMPUTATIONAL METHOD**

We use the finite difference method to solve eqs. (2), (3), and (4) simultaneously. Our aim is a) to simulate the string’s movement while it is in contact with the finger as well as when it is released, b) to calculate the duration of the interaction and how this is affected for different finger parameters and c) to predict the shape and the velocity distribution the string has when it is released.

The computer program consists of five time steps. In the first the string element sticks on the fingertip, on the second the string element slips along the fingertip, and on the third the string is vibrating freely. The process starts from the first loop and stays in it until the calculated friction reaches its maximum value; after that, it continues on the second loop. The process can go back to the first loop if the velocity of the string element becomes almost equal to the fingertip’s velocity or it can go to the third loop if the fingertip leaves the string. The string is divided into K segments, the spatial step is \( \Delta x = l/K \) and \( m_e = \varepsilon \Delta x \). The time step is calculated as \( \Delta t = l/(Kc) \) (see Chaine [7]). The position along the string’s length is given in terms of an integer \( n \) (\( n = 0 \) for \( x = 0 \), \( n = P \) for \( x = p \) and \( n = K \) for \( x = l \)), and the time increment is given in terms of an integer \( \tau \). All the computational constants are given in Table 1.

The recurrent equation for points along the string apart from the plucking point, is \( y_i[n, \tau + 1] = \alpha_0 y_i[n, \tau] + \alpha_1 y_i[n, \tau - 1] + \alpha_2 y_i[n + 1, \tau] + \alpha_3 y_i[n, \tau + 1] \).

On the first loop, the recurrent equation for the fingertip’s displacement is \( y_f[\tau + 1] = \alpha_1 y_f[\tau] + \alpha_2 y_f[\tau - 1] + \alpha_3 y_f[P, \tau] + \alpha_4 y_f[P, \tau - 1] + \alpha_5 y_f[P + 1, \tau] + \alpha_6 y_f[P - 1, \tau] + \alpha_7 y_f[\tau] + \alpha_8 \) and for the string-element’s displacement is \( y_i[P, \tau + 1] = y_i[P, \tau] + y_i[\tau + 1] - y_i[\tau] \).

On the second loop, the recurrent equation for the fingertip’s displacement is \( y_f[\tau + 1] = \alpha_1 y_f[\tau] + \alpha_2 y_f[\tau - 1] + \alpha_3 y_f[P, \tau] + \alpha_4 y_f[P, \tau - 1] + \alpha_5 y_f[P + 1, \tau] + \alpha_6 y_f[P - 1, \tau] + \alpha_7 y_f[\tau] \).

On the third loop, the recurrent equation for any point, including the plucking point, is \( y_i[n, \tau + 1] = \alpha_0 y_i[n, \tau] + \alpha_1 y_i[n, \tau - 1] + \alpha_2 y_i[n + 1, \tau] + \alpha_3 y_i[n - 1, \tau] \).

**RESULTS**

The finger’s parameters \( m_f, k_f, q_f \) are chosen so as the finger’s movement when it is not in contact with the string looks like Figure 2a and its mode frequency is only a few Hz. For the results below, the function \( y(t) \) was chosen to be proportional to \( t^2 \) (i.e. we assumed that the guitarist applied a constant force on his finger; this assumption could be modified) and the plucking position was set to \( x = 0.16m \). Figure 2b shows the displacement of the string and the fingertip on the plucking position, for \( m_f = .001 kg, k_f = 4 N/m, q_f = .09 Nsec/m \) for the open low E-string. For these parameters, the duration of the interaction was found equal to .14sec. The initial amplitude and velocity distribution of the string upon release are shown in Figure 3.

The duration of the interaction was found to be strongly dependent on the finger’s parameters. Increasing the finger’s mass or damping resulted in a longer time duration while increasing the finger’s stiffness resulted in a shorter duration. In addition, the duration did not change for different plucking positions.

Results for the amplitude and velocity distribution of the string when it is released show that for different finger parameters the string’s shape does not vary a lot and roughly looks like a triangle with the corner at the plucking position. In contrast, the velocity distribution along the string’s length changes a lot for different plucking positions and finger’s parameters.
Figure 2: (a): Movement of the finger when it does not interact with the string. (b): Displacement of string and fingertip at the plucking position.

Figure 3: (a): Initial shape and (b): velocity distribution of the string upon release.

CONCLUSIONS

Modeling the interaction is a part of ongoing work. The model will soon be expanded in two dimensions (string vibrating in y-axis and z-axis) allowing for the shape of the fingernail to be defined through a function. The string element will be able to stick, slip or roll along the fingernail and the created torsional waves will be included. The model will also allow movement of the string’s end supports giving the opportunity to include information for the guitar’s body.

Thus far, functions like the guitarist’s control $y(t)$ and the spring’s restoring force were defined very simply (for instance the spring’s restoring force was assumed to increase linearly with the string’s expansion, which is not realistic since the muscle of a finger allows only for limited movements). These functions can be modified after comparison with experimental data.

The model will allow us to investigate how an individual guitarist can produce a variety of sounds by changing the movement $y(t)$ of his finger. In addition we will be able to investigate the string’s movement for different finger’s parameters (i.e. for different guitarists) and particularly for different nail shapes. We will also be able to predict the net force the guitarist applies on the string during the interaction.

Results from this model will be used to investigate how the properties of the string and the finger influence the sound produced by a guitar as well as to improve existing models in Cardiff by inclusion of the player.

REFERENCES


- 472 -
STUDY OF VIBRATO EXTENT AND F0 VARIATION AMONG SINGERS

Eric Prame
Dept. of Speech Communication and Music Acoustics, KTH, Stockholm, Sweden

SUMMARY

Data on vibrato extent and F0 variations are calculated from measurements on 25 tones from Schubert's Ave Maria recorded by ten singers on commercially available CD records. Mean vibrato extent amounted to ±71 cent with extremes at 123 and 34 cent. The greatest difference between a singer's sharpest and flattest tones was no less than 69 cent, the mean across artists being 54 cent. The difference between the artist's mean intonation and the accompaniment varied considerably, from +12 to -20 cent. A negative relation was found between vibrato extent and tone duration.

INTRODUCTION

Vocal vibrato is basically a frequency modulation of F0, which in turn is causing an amplitude modulation of individual spectrum partials. Its rate has been analysed in detail in a previous study (Prame 1994). Here we will consider vibrato extent plus the running cycle average of F0, henceforth the MF0.

An extensive scientific analysis of vibrato was carried through in the early 30th by Carl E. Seashore and his team (Seashore 1937 and 1938) and later by many others (for reviews, see e. g., Sundberg 1987; Titze 1994). The study of singers' MF0 variation, however, is almost an undiscovered field even though Seashore 1937 provided a good starting point with two articles by Harold G. Seashore and Ray S. Miller (for a discussion of MF0, see also Sundberg 1987, 177-181).

MEASUREMENTS AND RESULTS

The same material was used for this study of vibrato extent and MF0 variation as in my previous investigation of vibrato rate, i. e., Franz Schubert's Ave Maria recorded on CD by 10 prominent artists (5 sopranos, one mezzo, one alto, and 3 tenors), all representing Western classical music tradition (Prame, 1994). The same 25 tones as before were selected for the analysis using time spectrograms on a KAY DSP Sona-graph, Model 5500. F0 was determined at all crests and troughs. Typical data are shown in Fig. 1a, b, c. In thirteen tones no F0 values could be obtained because of too low sound level etc. The total material included about 5400 measurements representing 237 tones.

From F0 values of the type shown in Fig. 1a the vibrato extent and MF0 were calculated, see Fig. 1b and c. The values of the vibrato extent has been presented in different ways in the literature. Here we will use cent and maximum positive and negative deviation from cyclic mean, henceforth ± amplitude. MF0 is given in cent as the deviation from the target value according to the equally tempered scale using the tuning of the accompaniment as the reference.
Fig. 2 shows the beginning and the end of the average vibrato extent curve (with SD) for tones with a length of at least 8 vibrato periods. The rather smooth curves are arch shaped and start and end at 42 and 55 cent, respectively, thus well above zero. However, the variation is great.

Fig. 1. Measured and calculated values for a tone. Notice the different scales! 1a. Measured points converted from Hz to cent; 1b. Vibrato extent envelope; 1c. MFO variation (lower left).

Fig. 2. Mean vibrato extent at start and end of tones (lower, right).

The tuning of the accompaniment was determined by a listening experiment. The first bar of the accompaniment was repeatedly played over loudspeaker. A professional violinist was asked to adjust the frequency of a complex tone so that it matched the pitch of the root of the chord. Three attempts were made for each recording, typically differing by less than 0.5 Hz. The mean of the three readings was used as a measure of the tuning.

Tables I and II show each artists’ values for a number of vibrato and MFO variation parameters.
Table I. Vibrato extent values for the ten artists. Column 1: Vibrato extent (cent) first averaged over all vibrato periods within each tone (tone mean) and then averaged across all 25 tones (artist average of tone means); 2-3. Extreme values (cent) for tone means; 4. SD (cent) of the 25 tone means; 5. Mean SD (cent) across the 25 tones; 6. Mean vibrato rate (Hz) according to previous investigation.

<table>
<thead>
<tr>
<th>Artist</th>
<th>Mean (cent)</th>
<th>Max (cent)</th>
<th>Min (cent)</th>
<th>Inter (cent)</th>
<th>Intra (cent)</th>
<th>Rate Mean (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72</td>
<td>106</td>
<td>38</td>
<td>17</td>
<td>14</td>
<td>5.6</td>
</tr>
<tr>
<td>2</td>
<td>84</td>
<td>105</td>
<td>59</td>
<td>12</td>
<td>18</td>
<td>6.5</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>99</td>
<td>42</td>
<td>16</td>
<td>13</td>
<td>6.3</td>
</tr>
<tr>
<td>4</td>
<td>73</td>
<td>102</td>
<td>50</td>
<td>16</td>
<td>13</td>
<td>6.0</td>
</tr>
<tr>
<td>5</td>
<td>64</td>
<td>85</td>
<td>36</td>
<td>13</td>
<td>10</td>
<td>5.9</td>
</tr>
<tr>
<td>6</td>
<td>57</td>
<td>71</td>
<td>44</td>
<td>8</td>
<td>10</td>
<td>6.7</td>
</tr>
<tr>
<td>7</td>
<td>72</td>
<td>88</td>
<td>51</td>
<td>9</td>
<td>10</td>
<td>5.9</td>
</tr>
<tr>
<td>8</td>
<td>86</td>
<td>123</td>
<td>66</td>
<td>15</td>
<td>14</td>
<td>5.4</td>
</tr>
<tr>
<td>9</td>
<td>60</td>
<td>80</td>
<td>34</td>
<td>11</td>
<td>10</td>
<td>5.6</td>
</tr>
<tr>
<td>10</td>
<td>70</td>
<td>111</td>
<td>48</td>
<td>14</td>
<td>12</td>
<td>5.8</td>
</tr>
<tr>
<td>Aver:</td>
<td>71</td>
<td>97</td>
<td>47</td>
<td>13</td>
<td>13</td>
<td>6.0</td>
</tr>
<tr>
<td>SD:</td>
<td>9.3</td>
<td>16</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Table II. MF0 data for the ten artists. Column 1 and 2: MF0 (in cent re 440 Hz and in Hz, respectively) of accompaniment; 3. Mean deviation of MF0 (cent) from accompaniment, first averaged over all vibrato periods within each tone (tone mean) and then averaged across all 25 tones (artist average of tone mean); 4. Difference between sharpest and flattest tone mean (cent); 5. SD (cent) across the 25 tone means; 6. Mean SD (cent) across the 25 tones.

<table>
<thead>
<tr>
<th>Artist</th>
<th>Acc. (cent)</th>
<th>Acc. eq. (Hz)</th>
<th>Deviation (cent)</th>
<th>Max - min (cent)</th>
<th>Inter (cent)</th>
<th>Intra (cent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>443.9</td>
<td>5</td>
<td>60</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>442.8</td>
<td>7</td>
<td>66</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>442.3</td>
<td>-7</td>
<td>45</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>440.9</td>
<td>10</td>
<td>47</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>445.1</td>
<td>-20</td>
<td>45</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>-2</td>
<td>439.6</td>
<td>-20</td>
<td>57</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>442.6</td>
<td>5</td>
<td>42</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>34</td>
<td>448.7</td>
<td>9</td>
<td>69</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>440.3</td>
<td>-14</td>
<td>52</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>442.0</td>
<td>12</td>
<td>60</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Aver:</td>
<td>11</td>
<td>442.8</td>
<td>-1</td>
<td>54</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>SD:</td>
<td>10</td>
<td>2.6</td>
<td>13</td>
<td>9</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

As can be seen in Table I the vibrato extent, averaged across all artists and tones, is 71 cent with individual tone extremes at 34 and 123 cent. Similar values were obtained for the extent variation within a tone (intra-variation) as well as between tones (inter-variation). Table II shows that the mean difference across artists between sharpest and flattest intonation was 54 cent, the individual maximum amounting to 69 cent. The intra and inter tone MF0 variation is again almost the same. The artist mean deviation from the accompaniment varied between +12 and -20 cent.
When comparing the tone means of vibrato extent averaged over the ten artists a negative relation was found between these 25 values and corresponding values for tone durations. No such relation seem to exist between vibrato extent and vibrato rate neither among the tones nor among the artists.

DISCUSSION

The collected material provides a solid basis for our knowledge of vibrato extent. The values are not far off from what could be expected from the many values found in the literature, but are slightly higher than Seashore’s commonly quoted average of ±50 cent with a typical variation between 30 and 70 cent. In addition to tone duration, other factors, related to musical expressivity, can be expected to influence vibrato extent.

The data on MF0 variation are more surprising. It is truly amazing that in a gorgeous performance of a song the intonation of tones can differ as much as 69 cent between flattest and sharpest. Also the high values of some artists’ mean MF0 deviation from the accompaniment (+12 and -20 cent) is astonishing, especially as the difference to some of the tones must reach much higher values. Many of these deviations are almost of a different order of magnitude than the JND for pitch perceived of vibrato tones (Sundberg 1987). An equally interesting question is what causes the variations of MF0 within the tone.

Some rather strong correlations were found between mean vibrato extent and various MF0 variation parameters. These relations will be further analyzed in the future.

The field of MF0 variation in sung performance seems to hide interesting and as yet unknown phenomena which invite further investigation. The assembled material of data on vibrato extent and MF0 variation would be a valuable asset in such an enterprise.

CONCLUSIONS

A measuring procedure was used to assemble a great number of data on vibrato extent and MF0 variations from recordings of 10 singers of world fame. The mean vibrato extent of the selected material amounted to ±71 cent and the maximum and minimum values were 123 and 34 cent respectively. The greatest difference between the highest and lowest tuned tone of a singer was no less than 69 cent while the mean across artists was 54 cent. The difference between the artist’s mean intonation and the accompaniment varied considerably, from +12 to -20 cent. Vibrato extent was found to have a negative relation to tone duration.

ACKNOWLEDGEMENTS.

Johan Sundberg and Johan Liljencrantz are gratefully acknowledged for their helpful and stimulating discussions and Lars Frydén for patiently, accurately, and kindly determining the tuning of the accompaniment.

REFERENCES


NUMERICAL MODELLING OF VIBRATIONS AND SOUND RADIATION FIELDS OF STRINGED MUSICAL INSTRUMENTS

Bernard E. Richardson
Department of Physics and Astronomy, University of Wales College of Cardiff, P.o. Box 913, Cardiff CF2 3YB, UK.

SUMMARY

One of the most important and illusive skills which needs to be acquired by the makers of stringed instruments is a working knowledge of how the materials, dimensions and design of an instrument affects its final sound quality. The maker's art is, however, not a quantitative science, and his empirical approach to this problem creates difficulties in achieving consistency in manufacture and an inability to copy existing instruments.

In order to gain a better understanding of the relationships between construction and sound quality, we have been developing techniques which enable us to model the acoustical function of instruments such as guitars and violins. Techniques include (i) the finite element method for computation of the vibrational modes of the structure, (ii) coupling between the structural modes, air-cavity modes and the transverse vibrations of the string, and (iii) the boundary element method for calculation of the sound radiated by the complete instrument. Application of a suitable force at the string can be used to synthesise a sound, allowing the instrument to be "played". These synthesised tones can then be used in formal psychoacoustic listening tests.

This paper will briefly review each stage of the modelling process, using the guitar as an example. In particular, it will concentrate on some of the numerical procedures involved in the boundary element method, a technique which is less well known to musical acousticians.

SCOPE OF THE MODEL

The motivation for the development of the model described here is to attempt to determine quantitative relationships between the construction and sound quality of stringed instruments. We have consequently not sought techniques which give extreme accuracy, nor have we been concerned to make the model operate in real-time. Instead we have been more concerned with establishing trends and identifying those aspects of construction which have the greatest influence on the sound of the instrument as a whole. We have been conscious to create a model which can be "played", so that the effects of physical changes can be assessed by ear.

The model comprises a vibrating string coupled to a soundboard, back plate and Helmholtz air cavity. These in turn are coupled to the surrounding air. Details of the coupling theory are given in an earlier paper[1]. The model yields interesting data concerning modes of vibration of the structure and their associated sound radiation fields. More importantly, however, it allows us to compute transfer response functions which link forces applied at the plucking point on the string to the sound pressure radiated to an arbitrary point in space. Application of suitable time-dependent forces at the string can then be used to synthesise sounds from the model. Preliminary listening tests[2] have already revealed the importance of low-order structural modes in radiating energy from the instrument at both low and high frequencies, and we are beginning to formulate
acoustic criteria which we believe could be useful to makers in achieving consistency in their manufacture of instruments. Results of further psychoacoustical experiments on synthesised guitar tones will be reported in another paper at this conference (Wright and Richardson).

STRUCTURAL VIBRATIONS

Numerical techniques such as the finite difference method and finite element analysis (FEA) have been used with great success for modelling vibrations of stringed musical instruments. The latter has been used extensively to model structural modes in violins and guitars, and even for modelling the internal air cavity modes of these same instruments[3]. Figure 1 shows some typical FEA output. We conclude from this figure that variations in constructional parameters, in this case the thickness of the soundboard, induce changes in the resonant frequencies and shapes of the modes. The format of the data, however, gives us little insight into the musical implications of the changes.

![Figure 1. The third mode of vibration of a guitar top plate calculated using FEA. The figures show contour plots of the mode shapes. The plate varies in thickness from (a) 2.2 mm, (b) 2.6 mm, to (c) 3.0 mm. Adapted from Walker[3].](image)

After examining a large amount of mode data from both real instruments and from FEA, it is easy to make the superficial observation that mode shapes are very consistent from instrument to instrument but that mode frequencies vary quite considerably. The absolute frequencies of modes can, and do, affect the musical response of the instrument, particularly in cases where individual string and body modes are strongly coupled. However, most string modes drive the body off-resonance, and good energy transfer between the string and the outside world depends on maintaining good coupling between the string and the body and the body and the surrounding air at all frequencies. The best sound radiators are the "air-pumping modes" with strong monopole components. But these modes are only of any use if the string is able to drive the mode effectively. Careful examination of Figure 1 shows subtle changes in the relative position of the nodal line (the heavy line crossing the bridge area) and the string driving position (marked by a dot); for this particular instrument, the mode becomes significantly more difficult to drive as the plate is made thicker. We note, however, that the radiation fields of each of these modes is virtually identical. If changes in the structure are to have any significant musical implications, there must be changes either in the sound radiation from the mode or in the coupling between the string and the body (or in both, of course). The latter coupling can be determined by calculating (or measuring) the effective mass of the structural mode at the string driving position.
SOUND RADIATION AND THE BOUNDARY ELEMENT METHOD (BEM)

The simplest method for modelling radiation fields is to compute the volume displacement of each mode and treat them as monopole sources[5]. Volume displacements are readily determined from the FEA output. It is simple to extend this method to include dipole components from the antisymmetric modes. For many of the low-order modes, this technique gives surprisingly good accuracy with the benefit of minimal computation. Further detail can be obtained by treating each element as a simple source of sound[11]. This gives reasonable approximations for the radiation directly to the front of the instrument, but it is unable to account for radiation from the back and sides and is less reliable for predicting the sound radiated towards the player, for example.

Significant improvements can be made by incorporating the boundary element method (BEM) for the calculation of sound radiation from the complete structure. The BEM is based on a numerical implementation of the Helmholtz Integral Equation (HIE):

\[ \int_I \left( p(x_j) \frac{\partial G(r)}{\partial n_i} - G(r) \frac{\partial p(x_j)}{\partial n_i} \right) dS = \varepsilon p(x_j) . \]

Here, \( p(x_j) \) and \( p(x_s) \) are the complex acoustic pressures at the field points \( f \) and \( s \), which respectively define the "listening" and source points with position vectors \( x_f \) and \( x_s \). \( G(r) \) is the free-space Green's function, where \( r = (x_f - x_s) \). The value of \( \varepsilon \) depends on the position of the "listening" point. For regions exterior to the surface \( \varepsilon = 1 \), but for points on the vibrating surface \( \varepsilon = \frac{1}{2} \). Edges and corners present a special problem. In these positions \( \varepsilon = \Omega/4\pi \), where \( \Omega \) is the solid angle subtended in the exterior field. If \( n_s \) is the outward unit vector normal to the surface, \( v_s \) the surface normal velocity, and \( \beta \) the angle between \( n_s \) and \( r \), then

\[ \frac{\partial G}{\partial n_s} = G(ik + \frac{1}{2}) \cos \beta, \text{ and } \frac{\partial p}{\partial n_s} = \nabla p \cdot n = -i \omega p v_s. \]

The numerical form of the HIE follows directly. If the surface is divided into \( N \) discrete elements, each of area \( A_j \), then

\[ \sum_{j=1}^{N} \left[ p_j (ik + \frac{1}{2}) \cos \beta + i \omega p v_j \right] A_j = \varepsilon p_f, \text{ where } G_j = \frac{\exp(-ikr)}{4\pi r}. \]

The equation is not intrinsically difficult to solve. The mode shapes and frequencies determined from FEA allow calculation of the average normal velocity for each surface element, including the velocity of the plug of air in the sound-hole (the latter being derived from coupled equations).

At this stage both \( p_j \) and \( p_f \) are unknown. However, by locating the listening point at each surface element in turn, a set of \( N \) simultaneous equations in \( N \) unknowns can be set up and solved to determine values for the associated surface pressures. Once the values for \( p_f \) are known, the HIE can then be evaluated at any exterior point. There are, of course, numerical problems associated with these calculations. For example, when the listening point coincides...
with the source point the equation becomes singular. There are also problems with non-uniqueness of the solutions and with the correct determination of $\varepsilon$ at edges or at surface points of high curvature. Procedures for dealing with these problems, plus an extensive coverage of the method as a whole, are given by Brooke\cite{2}.

**DISCUSSION**

The combination of FEA and the BEM for modelling musical instruments is extremely powerful. For computational ease in our guitar model, we are modelling the soundboard and back plate of the guitar independently. Nevertheless, the BEM is able to compute the sound radiated from the complete structure, including radiation from the sound-hole and the masking effects of the ribs. With more powerful processing power, it would be possible to model the vibrations and sound radiation fields of the complete wooden structure. The BEM yields information about the pressure and velocity fields over a closed surface, so in principle it is possible to calculate total power output, which is arguably a more useful quantity than precise knowledge of radiation fields. It should be noted that the BEM is as effective in the near field as it is in the far field.

Attempts have also been made to couple FEA and the BEM for calculation of radiation damping\cite{6}. Such calculations would be highly relevant to low-order modes in stringed instruments, which interact strongly with the air. Similar techniques might also allow calculations of the coupling between structural and internal (largely non-radiative) air-cavity modes.

The coupling of a string to this model allows direct evaluation of the musical importance of changes in the structure of the instrument. The model clearly identifies the benefits which are to be gained by "looking into" the instruments through the string. Psychoacoustical tests are beginning to identify the relative musical importance of the different mode parameters. Generally speaking, mode frequencies and Q-values are less important than effective masses and monopole source strengths. It is with great interest that we are now looking at the role played by each mode and at the co-operation between the various modes. It may well transpire that there is advantage in reducing the coupling of the strings to some modes.

**ACKNOWLEDGEMENTS**

Much of the work presented here is the result of the hard work and dedication of two former research students, Dr M Brooke and Dr G P Walker. It is their skills which helped to turn ideas in reality. I would also like to thank Professor P Blood for his continued support of musical acoustics activities at Cardiff.

**REFERENCES**

BORE RECONSTRUCTION BY PULSE REFLECTOMETRY AND ITS POTENTIAL FOR THE TAXONOMY OF BRASS INSTRUMENTS

David Sharp, Arnold Myers, Raymond Parks and Murray Campbell
University of Edinburgh, U.K.

SUMMARY

The character of a brass musical instrument is primarily dependent on its bore profile. Many historic brass instruments are so constructed that substantial parts of the air column are inaccessible to direct measurement. This paper discusses the factors involved and reports recent advances in the use of pulse reflectometry for bore reconstruction, in particular compensation for losses and allowance for multiple reflections. Measurements of historic musical instruments are presented and their significance evaluated.

TAXONOMIC OBJECTIVES

By a ‘brass instrument’ we mean a lip-vibrated aerophone consisting of a tube narrow in relation to its length, with or without a mechanism for changing the sounding length. Here we limit ourselves to unperforated tubes, i.e. instruments without finger-holes. The bore profiles which have proved to be viable for brass instruments have at one end a mouthpiece, which may be cupped or funnel-shaped, followed by the narrowest part of the windway, the ‘mouthpiece throat’. From the throat, the windway passes through the mouthpiece backbore, the mouthpipe or leadpipe, any tuning-slide or valves, finally coupling with the free air at the widest part of the tube, the bell. The area of cross-section in most cases increases monotonically; in some cases there is a localised narrowing (e.g. at slides and at valves, or due to some slight damage). The traditional bore cross-section is circular, where (rarely) a deliberately elliptical cross-section is introduced, or in the slight deformation at bends in the tube, the bore profile can for most purposes be considered to be equivalent to that of a cylindrical tube of the same cross-sectional area.

Few classification systems go beyond a division of brass instrument types into ‘conical’ and ‘cylindrical’; both concepts have intuitive meaning, but are not capable of rigorous definition. In order to approach an acoustically-based system of taxonomy, it is helpful to look separately at the beginning, middle, and the end of a brass instrument. The choice of a mouthpiece can affect the character of an instrument, though not completely determine it (Myers & Campbell 1993). The design of the bell flare is of critical importance in the sound radiation properties and in the relative intonation of the resonating modes (Benade and Jansson 1974; Jansson and Benade 1974). However, it does not appear that the mouthpiece and the bell flare even considered together can account for the more subtle distinctions between instrument types, or between earlier and later historical models of the same instrument type. It is therefore necessary to examine the complete bore of an instrument. A mouthpiece, where present, can be physically measured, as can the tapered mouthpiece receiver on the body of the instrument. The bell can also be physically measured, or at least enough of it to allow calculation of the horn function in the region of its peak value (in the case of flaring bells). Much of the sounding length of the instrument can, however, pose severe problems for direct physical measurement, particularly in instruments with many coils.

For the purposes of comparison between instruments, a coiled instrument can be treated as equivalent to a perfectly straight instrument with the same cross-sectional area at each point along a line drawn through the geometric centre of the bore, the ‘mid-line’. The effects of bends in the tubing of wind instruments were considered theoretically by Keefe and Benade (1983) and its practical implications
for taxonomy considered by Myers and Parks (1995). The bends encountered in the great majority of actual instruments will give rise to second-order discrepancies only. Other factors which affect brasswind character, although important in performance, have to be regarded as being of second (or higher) order for taxonomic purposes. These include the properties of a particular player’s lips and vocal tract, wall thicknesses, bore perturbations (e.g. water-keys, dents, valve misalignments) temperature and humidity gradients in playing conditions, and unsteadiness in the flow of air through the instrument.

**PULSE REFLECTANCE TECHNIQUES**

The extension of the pulse reflectance techniques for bore reconstruction already in use in the medical field (Marshall, 1990) to brass instruments is a potentially useful means of investigating brass instruments for the purposes of classification. Existing applications to musical instruments (Watson and Bowsher 1987) have been directed to other purposes. A particular advantage would be in establishing the bore profile of instruments with substantial portions of coiled tubing without the difficulties of making large numbers of precise physical measurements of curved tube: a smaller number of direct physical measurements could be made and bore reconstruction techniques used for interpolation. Pulse reflectometry is a non-invasive technique and hence is very useful in the measurement of instruments with a degree of inaccessibility.

In work carried out at the department of Physics at the University of Edinburgh, an electrical pulse (containing frequencies from 0–12 kHz) was produced, amplified and used to drive a loudspeaker. The resultant sound pressure pulse was passed along a source tube of diameter 9.6mm (of the same order of magnitude as the tubing of the narrower parts of brass musical instruments). A microphone, embedded partway along the tube, recorded the input pulse as it passed. A short time later, it recorded the reflections returning from the instrument under test, which (without its detachable
mouthpiece) was coupled to the far end of the source tube.

For an ideal delta function sound pressure pulse, the reflections obtained from the instrument would be its input impulse response. However, the sound pressure pulse departs from a pure delta function in an arbitrary way, so to obtain the input impulse response, the reflections are deconvolved with the input pulse shape, using a transform size of 1024. The reflections occur at changes in impedance, such as at expansions or contractions of the instrument bore. A suitable algorithm allows the reflection coefficients arising from these impedance changes to be evaluated from the input impulse response. It is then a small step to calculate the changes in area along the bore and, assuming cylindrical symmetry, the changes in radius.

Watson and Bowsher (1978, 1988) applied the technique of pulse reflectometry to brass instruments, using the algorithm derived by Ware and Aki (1969) to reconstruct the bore profiles. However, although the Ware-Aki algorithm takes into account multiple reflections within the instrument, it does not consider attenuation of the signal along the length of the instrument. Hence, the accuracy of the bore profiles decreased with the length of the instrument (as attenuation became more significant). More recently, Amir, Rosenhouse and Shimony (1995a,b) have developed an algorithm which compensates for the attenuation along the instrument, resulting in significantly more accurate reconstructions. We have used this algorithm to examine the bore profiles of brasswind from the Edinburgh University Collection of Historic Musical Instruments.

RESULTS FOR CERTAIN HISTORICAL MODELS OF INSTRUMENT

Fig. 1 shows the bore profile of a B♭ cornet by Rudall Carte (without mouthpiece) without any valves operated. The directly measured profile and the profile reconstructed from reflectance measurements are superimposed. The initial dip is the mouthpiece receiver taper, and does not represent part of the sounding bore when a mouthpiece is inserted. The small-scale fluctuations in the reconstructed bore are mostly accounted for by features such as waterkeys and small discontinuities at the ends of tuning slides and at the valves; no attempt was made to measure the cross-sectional area at these features.

In Fig. 2 we see the reconstructed profile of the same instrument with all three valves operated; there are more small-scale fluctuations in the central part of the bore where the windway passes in and out of the valves. When the measurements were first taken, the irregular (solid line) 'profile' was plotted, deviating substantially from the measured profile beyond the third valve. Close examination of the instrument showed a leak in the tubing of the third valve tuning-slide — a leak too small to be perceptible to a player of the instrument, but 'seen' by the pulse as an enlargement of the effective cross-section of the tube. With the leak sealed, a good agreement with the actual bore of the instrument was obtained (broken line).

The question of musically inconsequential leaks having a disproportionate effect on the bore reconstruction is a potential problem. The bandwidth of the pulse spectrum utilised in reflectometry does not, of course, match the range of frequencies employed in musical use of the instrument. It is proposed to distinguish such small leaks.

Figure 3: Bore profiles of Rudall Carte cornet (EUCHMI 2988) and Boosey cornet (EUCHMI 2704), all three valves operated.
by comparing the results obtained from ‘tailored’ pulses with differently shaped spectra.

Careful inspection of the central part of the bore in the previous figure indicates that the bore through the valves is not cylindrical, but expands gradually. This is indeed the case: the Rudall Carte cornet is the ‘Patent Conical Bore’ model with incremental bore cross-section in the windways through the pistons and in the tuning-slide bows. Fig 3 shows a comparison of the Rudall Carte ‘Conical Bore’ cornet with a standard cornet by Boosey which has the usual cylindrical profile. The Boosey cornet has at some time had its playing pitch lowered by extension of each leg of its tuning-slide; this accounts both for the greater overall length and for the two peaks around 0.5 metres.

The method of pulse reflectance is of potential value for coiled tubes with bores inaccessible to direct measurement. Fig. 4 shows the successful reconstruction of a Spanish bugle. Only the mouthpipe (sliding for tuning purposes) and the bell flare could be reached for direct measurement.

CONCLUSIONS

For purposes of comparison of bore profile between instruments, bore reconstructions of the accuracy now possible with corrections for attenuation offer a useful tool for the taxonomist. The problems of measuring the mid-line in coils are avoided, since the use of sound waves in measurement ensures that the acoustically defined path is what is measured. The interpretation of the results for the great variety of instrument types is our next goal.

REFERENCES


Figure 4: Bore profile of Honshu bugle in C (BUCHMI 2343).
SUBJECTIVE MEASUREMENTS OF DENOISED MUSIC SIGNALS: REALISATION AND ANALYSIS

Laurent Simon, Jean-Christophe Valiere
Laboratoire d'Acoustique de l'Université du Maine, Avenue O. Messiaen
BP535, 72017 Le Mains Cédex, France.

SUMMARY
Different noise reduction methods are compared using listening tests in the study of signal restoration. Results are analysed and show that a recent technique [1] clearly improves the global quality of the restoration. Moreover, a general way of systematically comparing noise reduction techniques is proposed.

INTRODUCTION
In the restoration of old recordings, listening is the only way to globally compare different noise reduction methods. In this context, listening tests have been realised to compare different noise reduction techniques, using different sets of parameters. The adjustment of these parameters is often critical and poor choices lead to audible signal degradations [2]. Thus, throughout the test, questions have concerned not only the remaining noise or the general impression, but also the quality of the musical signal. Furthermore, in the literature of speech enhancement and old recording restoration, there is a little mention of listening tests.

The aim of this paper is threefold: to confirm the improvement of the recent noise reduction technique we have set up [1], called the mixed-filter-bank, compared to the classical whole band method [2], to have a general insight into the criteria used by the listeners, and to propose a general procedure of testing noise reduction methods.

FIRST LISTENING TEST: REALISATION, RESULTS AND ANALYSIS
The first set of listening tests was set up during October 1993 in the Acoustics Laboratory of the University of Maine, to compare recent noise reduction techniques applied to old recordings. The signal analysed was a thirty second long monophonic recording extract of the opera "Carmen", dating from 1929.

The purpose of this test was thus to compare noise reduction treatments, called $B_i$ (with i in the domain {1-5}), using different sets of parameters, in terms of a priori independent perceptual dimensions, and using the original undenoised signal (called A) as the reference. This listening technique is called the A/B technique or pair-comparison method [3]. Each pair ABi, for a given i, randomly appeared three times during the test. The five treatments are described in Table 1. The whole test lasted thirty minutes.

Forty-nine listeners performed the test; only thirty-one of them have been retained according to the smallest variances of the answers to identical questions. Most of them were acousticians and/or musicians. The questions concerned the following areas:
1- musical signal distortion (degradation due to the treatment),
2- remaining noise, in terms of quality (2a) and quantity (2b),
3- general subjective impression (quality of the restoration)
From the answers of the selected listeners, averages, variances, 80% confidence intervals and correlations [4] have been calculated. Averages and 80% confidence intervals are presented in Table 2 (marks are out of 10). The subjective rating yields the following general tendencies:

- treatments B1 and B2 are clearly closely estimated for every question,
- treatment B3 paradoxically appears better than treatment B1,
- treatment B4 is better than treatment B1, especially for questions 2b and 3, and
- treatment B5 is the worst one, except for question 1, where it obtains the best mark.

The analysis of the answers can then be artificially split into three different categories: the "expected results" which have been confirmed by the test, the "expected results" which have not been confirmed by the test and the unexpected results.

- **Two expected results, which have been confirmed**: 1/ Treatment B4 (based on the mixed filter bank) is clearly the most satisfactory one, particularly for the general impression and especially in comparison with B1. 2/ When comparing treatments B4 and B5, of which the musical signal is basically the same, the listener prefers the noisier treatment, in terms of quality of the musical signal. This is a confirmation of a psychoacoustic law: the signal appears richer in high-pitched harmonics when noise is present.

- **Two unexpected results, which have not been confirmed**: 1/ The notion of distortion of the musical signal has not been discriminated. In this way, the choice of parameters of the treatment B3 generates an important smoothing of transients and eliminates a lot of musical components, which is unacceptable for a commercial restoration. Therefore, the listeners do not distinguish the smoothing of transients and the elimination of components, even though they are perceptible. 2/ Similarly, treatments B1 and B4 are identically estimated with regard to the first question. The general impression is then estimated only from the remaining noise.

- **One unexpected result**: The quality and, to a lesser degree, the quantity of the remaining noise are not only the prevailing criteria for the restoration quality estimation, but also for the assessment of the subjective quality of the musical signal. Listeners have difficulty leaving aside the remaining noise when estimating the quality of the musical noise.

In conclusion, two points have been retained from this test: 1/ The quality of treatment B4 (based on the mixed filter bank) is indisputably preferred to the quality of the treatment B1 (based on the classical whole band treatment). As there is no crossover of the 80% confidence intervals for the last question, we can assert that the general quality of the restoration is improved by using the mixed filter bank. 2/ On the other hand, as the smoothing of transients was not discriminated, a supplementary test limited to this question was set up.

**SECOND LISTENING TEST: REALISATION, RESULTS AND ANALYSIS**

In order to distinguish the effect of the noise reduction methods on the smoothing of transients, a second listening test was set up. A part of the first extract was used, containing string pizzicati, for which the smoothing of attack transients can be easily discriminated. This extract lasted four seconds. Each comparison between two treatments called T1 and Tj (i, j in the domain {1-4} and i≠j) consisted of the following listening sequence:

\[ T_i \rightarrow T_j \rightarrow T_i \rightarrow T_j \rightarrow T_i \rightarrow T_j \rightarrow S, \]

where S was a twenty second silence. Each sequence lasted one minute and randomly appeared twice during the test. The whole test lasted twelve minutes. Following an idea used for loudspeaker listening tests [5], we chose to compare the previous treatments to each other. Four different treatments have been compared, including three of the previous treatments, as presented in Table 3.
The sole question was about the smoothing of transients. From the best respondents to the first test, twelve listeners were selected. The listeners corresponding to the three worst variances were eliminated. As attempted in Ref. [6], this panel of listeners presents very good variances and the results of their answers is clearly highly coherent. Averages and 50% confidence intervals have been calculated for each comparison between treatments (Table 4). Even if the whole scale (from -5 to +5) was not used, general tendencies can be deduced:

- The three treatments T1, T2 and T4 present homogeneous results: they are identically estimated when listened to two at a time. It is a highly coherent result from the listeners, especially when comparing T2 and T4.
- Treatment T3 is always judged worse than all the other treatments, with separated confidence intervals. The subjective distance between T3 and T4 is weaker than the other ones, which is coherent with the previous remark: the musical signal appears better when a noise is present, and actually, T1 and T2 are clearly noisier than T4.

In conclusion, in spite of different noise levels (for treatments T1, T2 and T4), the listeners considered that the quality of the musical signal was the same. On the other hand, it was confirmed that T3 was the treatment which generated the smoothest transients.

CONCLUSION

Listening tests have been organised to compare recent noise reduction techniques applied to the restoration of old recordings. It has been confirmed that the noise reduction method we propose [1] indisputably improves the quality of the restoration. Concerning the compromise between the remaining noise level and the quality of the musical signal, it appears that it can be preferable to leave some noise on the denoised signal to subjectively improve the quality of the musical signal. Finally, we propose a global realisation methodology of listening tests for the restoration of old recordings, as follows: from a large panel of listeners (at least 50 persons), a first general test made up of different questions (about the musical signal degradation, the remaining noise, the general impression, ...) is set up. Then, from the best answers (weaker variances), a reduced panel is selected and judicious questions are asked about more precise points (smoothing of transients, tonal equilibrium, ...) from shorter segments of the previous extract. General and more precise listeners' behaviours constitute the global rating we then have to analyse.

Supplementary tests (multiple-comparison method, listening tests realised by sound professionals) will permit the level of noise which realises the best "remaining noise level - quality of the musical signal" compromise to be more precisely quantified.

ACKNOWLEDGMENT

We wish to thank Dr Niven Brown for his helpful comments and numerous corrections to this paper.

REFERENCES

**Table 1:** First test: Treatments used. (*) subjective optimality, according to the authors.

<table>
<thead>
<tr>
<th>Name</th>
<th>Treatment</th>
<th>Optimised</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>Whole band</td>
<td>Yes</td>
<td>Best subjective realisation (*) [2]</td>
</tr>
<tr>
<td>B2</td>
<td>Whole band</td>
<td>Yes</td>
<td>B1 plus white noise (SNR=35dB)</td>
</tr>
<tr>
<td>B3</td>
<td>Whole band</td>
<td>No</td>
<td>Distortion of the musical signal</td>
</tr>
<tr>
<td>B4</td>
<td>8 band splitting</td>
<td>Yes</td>
<td>Best subjective realisation (*) [1]</td>
</tr>
<tr>
<td>B5</td>
<td>8 band splitting</td>
<td>No</td>
<td>Higher denoised band replaced by the original one</td>
</tr>
</tbody>
</table>

**Table 2:** Averages and 80% confidence intervals. 31 listeners, 93 judgments, 1 experiment. Marks are out of 10.

<table>
<thead>
<tr>
<th>Question</th>
<th>Signal B1</th>
<th>Signal B2</th>
<th>Signal B3</th>
<th>Signal B4</th>
<th>Signal B5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (musical signal distortion)</td>
<td>6.4</td>
<td>6.4</td>
<td>7.4</td>
<td>6.5</td>
<td>7.5</td>
</tr>
<tr>
<td>2a (remaining noise quality)</td>
<td>6.5</td>
<td>6.5</td>
<td>5.7</td>
<td>6.8</td>
<td>2.9</td>
</tr>
<tr>
<td>2b (remaining noise quantity)</td>
<td>4.9</td>
<td>4.5</td>
<td>7.1</td>
<td>6.7</td>
<td>1.9</td>
</tr>
<tr>
<td>3 (general impression)</td>
<td>5.2 → 5.7</td>
<td>5.3 → 5.8</td>
<td>5.7 → 6.4</td>
<td>6.4 → 6.8</td>
<td>3.2 → 3.8</td>
</tr>
</tbody>
</table>

**Table 3:** Second test: Treatments used. (*) subjective optimality, according to the authors.

<table>
<thead>
<tr>
<th>Name</th>
<th>Treatment</th>
<th>Optimality</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Whole band</td>
<td>Yes</td>
<td>Previously called B1</td>
</tr>
<tr>
<td>T2</td>
<td>8 band splitting</td>
<td>Yes</td>
<td>B4 (T4) less denoised in higher bands</td>
</tr>
<tr>
<td>T3</td>
<td>Whole band</td>
<td>No</td>
<td>Previously called B3</td>
</tr>
<tr>
<td>T4</td>
<td>8 band splitting</td>
<td>Yes</td>
<td>Previously called B4</td>
</tr>
</tbody>
</table>

**Table 4:** Averages and 50% confidence intervals. 9 listeners, 18 judgments, 1 experiment. Marks are out of 10 (-5 to +5).

<table>
<thead>
<tr>
<th>Comparison between</th>
<th>Mark</th>
<th>Comparison between</th>
<th>Mark</th>
</tr>
</thead>
<tbody>
<tr>
<td>T3 - T2</td>
<td>-1.1</td>
<td>T1 - T2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>-1.4 → -0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3 - T1</td>
<td>-1.</td>
<td>T3 - T4</td>
<td>-0.5</td>
</tr>
<tr>
<td></td>
<td>-1.3 → -0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4 - T2</td>
<td>0</td>
<td>T1 - T4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>-0.4 → 0.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 488 -
DIGITAL WAVEGUIDE MODELS FOR SOUND SYNTHESIS
BASED ON MUSICAL ACOUSTICS

Julius O. Smith III
Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University Stanford, CA 94305.

SUMMARY

This paper gives an overview of the digital waveguide modeling technique, recent results, and directions for future research.

INTRODUCTION

In recent years there has been a growing trend toward "time domain" simulations in musical acoustics [Benade 1988, Chainge 1992, Chainge and Askewfelt 1994, McIntyre et al. 1983, Ruiz 1969]. Time-domain models can accurately incorporate nonlinearities and time-varying parameters, unlike frequency domain descriptions. As a result, time-domain methods provide much greater modeling generality.

Time-domain models of speech production have been in use for many years by signal processing researchers pursuing automated speech synthesis [Markle and Gray 1976, Oppenheim and Schafer 1975]. In this application, the vocal tract is implicitly modeled as a piecewise cylindrical acoustic tube driven by a periodic and/or noise waveform corresponding to glottal excitation. Linear predictive coding (LPC) is used to estimate the inter-section reflection coefficients from the speech waveform. In practical LPC, the acoustic model is actually simplified to the point that a precise physical interpretation no longer exists. The basic connection to a true acoustic model is widely cited but hardly ever explored as such.

In the field of computer music, there has also been a recent trend toward time-domain models of musical instruments for purposes of research and digital synthesis. Some of these are based on physical modeling concepts [Cook 1991, Hirschman 1991, Karjalainen et al. 1991, Smith 1983-1992, Van Duyne and Smith 1993] while others are more properly classified as "signal models"—time-domain algorithms which happen to be effective in reproducing a certain type sound [Karplus and Strong 1983, Jaffe and Smith 1983, Laroche and Mellier 1993]. Signal models are often used in the context of "lossy" signal compression techniques [Rabiner and Schafer 1978]. There is a nice coincidence between acoustic models using sampled acoustic waveguides and signal models which are based on filtered delay loops. Commutativity of linear, time-invariant elements can often be used to convert the precise acoustic model to a much simpler filtered delay-loop equivalent [Smith 1987, 1992].

DIGITAL WAVEGUIDE MODEL FOR THE IDEAL STRING

The wave equation for the ideal (lossless, linear, flexible) vibrating string can be written as $K y'' = \epsilon y$, where $K$ is string tension, $\epsilon$ is linear mass density, $y$ is string displacement, $y = y(t, x)$, $y' = \frac{\partial y}{\partial t}(t, x)$, and $y'' = \frac{\partial^2 y}{\partial t^2}(t, x)$, where "$\Leftarrow$" means "is defined as." The same wave equation applies to any perfectly elastic medium displaced along one dimension, e.g., the air column of a clarinet, organ pipe, or human vocal tract. We refer to the general class of such media as one-dimensional waveguides. Extensions to two, three, and higher dimensions have also been investigated [Van Duyne and Smith 1993].

For a physical string model, at least three coupled waveguide models should be considered, corresponding to the horizontal and vertical transverse wave polarizations, as well as longitudinal waves.
For bowed strings, torsional waves should also be considered, since they affect bow-string dynamics [McIntyre et al. 1983].

It has been known since 1747 that the wave equation for the ideal string is solved by any string shape which travels to the left or right with speed \( c = \sqrt{K/\mu} \). If we denote right-going traveling waves by \( y_r(x - ct) \) and left-going traveling waves by \( y_l(x + ct) \), where \( y_r \) and \( y_l \) are arbitrary twice-differentiable functions, then the general class of solutions to the lossless, one-dimensional, second-order wave equation can be expressed as

\[
y(x, t) = y_r(x - ct) + y_l(x + ct)
\]

To carry the traveling-wave solution into the "digital domain," it is necessary to sample the traveling-wave amplitudes at intervals of \( T \) seconds, corresponding to a sampling rate \( f_s = 1/T \) samples per second. For CD-quality audio, we have \( f_s = 44.1 \text{kHz} \). The natural choice of spatial sampling interval \( X \) is the distance sound propagates in one temporal sampling interval \( T \), or \( X = cT \) meters. In a traveling-wave simulation, the whole wave moves left or right one spatial sample each time sample; hence, simulation only requires digital delay lines.

Formally, sampling is carried out by the change of variables \( x \to x_m = mX \) and \( t \to t_n = nT \). Substituting into the traveling-wave solution of the wave equation gives

\[
y(t_n, x_m) = y^+(n - m) + y^-(n + m)
\]

where

\[
y^+(n) \triangleq y_r(nT) \quad y^-(n) \triangleq y_l(nT)
\]

The right-going traveling-wave component \( y_r[(n - m)T] = y^+(n - m) \), can be thought of as the output of an \( m \)-sample delay line whose input is \( y^+(n) \). Since \( y^+ \) travels to the right, we draw its delay line with input \( y^+(n) \) on the left and its output \( y^+(n - m) \) on the right. This can be seen as the upper "rail" in Fig. 1.

Similarly, the term \( y_l[(n + m)T] \triangleq y^- \) can be thought of as the input to an \( m \)-sample delay line whose output is \( y^- \). Since \( y^- \) is the left-going component, it makes sense to draw the delay line with its input \( y^- \) on the right and its output \( y^- \) on the left. This can be seen as the lower "rail" in Fig. 1. Note that the position along the string, \( x_m = mX = mcT \) meters, is laid out from left to right in the diagram, giving a physical interpretation to the horizontal direction in the diagram. Finally, the left- and right-going traveling waves must be summed to produce a physical output according to

\[
y(t_n, x_m) = y^+(n - m) + y^-(n + m)
\]

We may compute the physical string displacement at any spatial sampling point \( x_m \) by simply adding the upper and lower rails together at position \( m \) along the delay-line pair. In Fig. 1, "transverse displacement outputs" have been arbitrarily placed at \( z = 0 \) and \( z = 3X \). The diagram is similar to that of well known ladder and lattice digital filter structures [Markle and Gray 1976], except for the delays along the upper rail, the lack of lossless scattering junctions, and the direct physical interpretation. We could proceed on to ladder and lattice digital filters by (1) introducing a perfectly reflecting (rigid or free) termination at the far right, (2) defining changes in the wave impedance at discrete points along \( z \), and (3) commuting the delays rightward from the upper rail down to the lower rail. In acoustical tube simulations, such as for voice [Cook 1960] or wind instruments [Hirschman 1991], lossless scattering junctions are used to model discrete changes in cross-sectional vocal-tract area, and lossy scattering junctions are used to implement tone holes.

\[
\begin{align*}
&\cdots \quad x^0 \quad x^1 \quad x^2 \quad x^3 \quad \cdots \\
&\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\
&y^0 \quad y^1 \quad y^2 \quad y^3 \quad \cdots \\
&\cdots \quad x^0 \quad x^1 \quad x^2 \quad x^3 \quad \cdots \\
&\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\
&y^0 \quad y^1 \quad y^2 \quad y^3 \quad \cdots
\end{align*}
\]

\[\text{Figure 1: Digital simulation of the ideal, lossless waveguide with observation points at } z = 0 \text{ and } z = 3X = 3cT.\]

The symbol "\( z^{-1} \)" denotes a one-sample delay.

The digital waveguide simulation is exact at the sampling instants, to within the numerical precision of the samples themselves, provided that the waveforms traveling along the string are initially bandlimited.
to less than half the sampling frequency. In other words, the highest frequencies present in the signals \( y_n(t) \) and \( y_i(t) \) may not exceed half the temporal sampling frequency \( f_s \geq 1/T \); equivalently, the highest spatial frequencies in the shapes \( y_n(x/c) \) and \( y_i(x/c) \) may not exceed half the spatial sampling frequency \( \nu_c \geq 1/X \). Bandlimited spatial interpolation may be used to construct a displacement output for an arbitrary \( z \) not a multiple of \( cT \), and bandlimited interpolation across time serves to evaluate the waveform at an arbitrary time not an integer multiple of \( T \) [Smith and Gessett 1984].

RELATION TO THE FINITE DIFFERENCE APPROXIMATION

It is interesting to compare the digital waveguide simulation technique to the recursion produced by the finite difference approximation (FDA) of the wave equation. The FDA time-update recursion for the ideal string is given by [Chaigne 1992]

\[
y(n+1, m) = y(n, m+1) + y(n, m-1) - y(n-1, m)
\]

Substituting the traveling-wave decomposition \( y(n, m) = y^+(n-m) + y^-(n+m) \) (which is exact in the ideal case at the sampling instants) into the right-hand side of the FDA recursion above gives the surprising result that the FDA recursion is also exact in the lossless case. This is surprising since the FDA involves replacing derivatives with finite differences, which is at first sight a gross approximation. This equivalence has been found to be true also in higher dimensions and on other grid types.

MULTIDIMENSIONAL WAVEGUIDE MODELING

Scott Van Duyne, a Ph.D. candidate at CCRMA, has obtained new results for simulation of membranes and volumes using 2D and 3D waveguide meshes. For example, in 2D, the membrane can be simulated by a rectilinear mesh of strings, analogous to a tennis racket. Since linearity holds and energy is preserved, the only essential approximation error is traveling-wave dispersion, i.e., sound-speed on the mesh varies with frequency and with direction along the mesh. Different mesh types exhibit different dispersion characteristics. Dispersion in the rectilinear mesh is less uniform than that in the hexagonal mesh, for example, (which uses three-way intersections instead of four). Intuitively, the elimination of straight lines in the mesh serves to make the various angles of propagation more similar to each other. In the rectilinear 2D mesh, propagation is exact along diagonals of slope \( \pm 1 \), and the most dispersive directions are along the \( x \) and \( y \) axes. In those directions, higher frequencies travel more slowly. At the upper limit of half the sampling rate, the phase velocity falls to 0.7c. In the hexagonal mesh, there are six equivalent directions of propagation which are highly non-dispersive, and the angles in between show less variation in dispersion relative to the rectilinear mesh.

Due to characteristics of human hearing, multidimensional waveguide meshes work better for audio simulations than might at first be expected. At low frequencies, there are plenty of grid-points in the mesh per wavelength, and so the low-frequency modes are well tuned. At high frequencies, the dispersion phenomenon causes mistuning of the high frequency modes; however, if this begins at a frequency at which there are several modes per critical band of hearing, one cannot hear the mistuning under normal conditions of excitation. Modal energy is not lost by dispersion.

CONCLUSIONS

The digital waveguide modeling approach is proving itself to be an especially efficient and numerically well-behaved time-domain modeling technique for acoustic systems. Good results have been obtained in one, two, and three dimensions. Future directions for research include further incorporation of nonlinearities, further characterization of and compensation for multidimensional dispersion, preservation of energy-invariance under time-varying and nonlinear conditions, artificial reverberation, new kinds of digital effects, refinement of instruments in the string, wind, and brass families, development of percussion instruments, and application to other distributed media.
REFERENCES


FLEXIBLE STRING VIBRATIONS EXCITED BY
A NONLINEAR HEREDITARY HAMMER

Anatoli Stuvelov

SUMMARY
The process of a flexible string excitation by striking with a nonlinear hammer is considered. Two types of the strongly nonlinear hammers were used for comparison. For the first hammer the force is determined by the felt compression in the form of power-law dependence. The second nonlinear model of the hammer takes into account the hysteresis-type of the force-compression characteristics of the felt deformation also. It is shown that this model gives prediction of the string energy spectrum better than the first one and without allowing for additional physical processes in string as damping.

HAMMER FELT MODEL
The timbre of sound produced by a piano mostly depends on the detailed motion of strings excited by the impact of the hammers. So, the creation of good theoretical model which permits to predict the string vibrations spectra is important problem that has been the subject of considerable research for many years. The last preceding papers in this series are [1-4]. The comparisons with laboratory measurements in [1], however show definite limitations on the success of the models considered. Their greatest weakness appears to be their failure to allow for the nonlinear compliance properties of real hammers, which have been measured and discussed earlier by several other authors [2-7]. Hall [2] addresses the problem of how to make improved predictions with models that allow for nonlinearity in the string-hammer interaction. The model which is described in that paper deals with nonlinear hammer and a perfectly flexible string.

But the experiments provided by Yanagisawa, Nakamura and Aiko [5,6] have shown the significant influence of hysteresis characteristics of the hammer. Boutillon [7] has made an attempt to explain the nonlinear hysteretic character of the force-compression curve, but his non-analytical model cannot describe the dependence of the slope of the force-compression characteristics on the rate of loading.

The analytical model of such nonlinear hysteretic hammer which takes into account all important hammers features observed in the experiments is presented in [8].

According to this model the hammer possess history-dependent properties or just as well is made of the hereditary material. In this case two hereditary parameters $\gamma$ and $\tau_0$ are involved to describe the hysteretic behaviour of the hammer. The governing equation connecting the nonlinear force $F(u)$ exerted by hammer and the felt compression $u(t)$ can be written for hereditary hammer in the form

$$F(u(t)) = F_0 \left[ u^p(t) - \frac{c}{\tau_0} \int_0^t u^p(\xi) \exp \left( \frac{t-\xi}{\tau_0} \right) d\xi \right].$$

The constant coefficient $F_0$ (with units such as N/mm$^p$) is a felt stiffness constant of the
hammer. The suitable values of hereditary constants $\varepsilon$, $\tau_0$ and $F_0$ for various hammers were obtained by numerical simulation of experiment [6] and are presented in [8].

FLEXIBLE STRING

The string vibrations are governed by the following equation

$$\frac{\partial^2 y}{\partial t^2} = c^2 \frac{\partial^2 y}{\partial x^2},$$

(2)

where $c = \sqrt{T/\mu}$ gives the wave speed in terms of tension $T$ and linear density $\mu$ of the string. This equation is satisfied by simple nondispersive waves of arbitrary waveforms $f_+$ and $f_-$ moving in both directions and a newly created outgoing wave $g(t)$ due to the hammer’s action

$$y(x, t) = \begin{cases} 
  f_+(t - x/c) + f_-(t + x/c) + g(t + x/c) & (x < 0), \\
  f_+(t - x/c) + f_-(t + x/c) + g(t - x/c) & (x > 0). 
\end{cases}$$

(3)

Let the string of total length $L$ extends from $x = -(L - l) = -\beta L$ on the left to $x = l = \alpha L$ on the right, with $\beta = 1 - \alpha$. Using the boundary conditions $y(\alpha L, t) = y(-\beta L, t) = 0$ from Eq. (3) it is easy to find the string deflection at the contact point

$$y(0, t) = g(t) + 2 \sum_{i=1}^{\infty} \left[ g \left( t - \frac{2iL}{c} \right) - \sum_{i=0}^{\infty} \left( g \left( t - \frac{2iL}{c} - \frac{2\alpha L}{c} \right) + g \left( t - \frac{2iL}{c} - \frac{2\beta L}{c} \right) \right) \right].$$

(4)

The physical interpretation of this equation shows that the outgoing wave $g$ at some earlier time has been reflected from the string ends. At the contact point $x = 0$, we must have

$$m \frac{d^2 z}{dt^2} = -F(u) = T \Delta \left( \frac{\partial y}{\partial x} (x = 0, t) \right).$$

(5)

This represents the transmission of force between the hammer body (with mass $m$ and displacement $z$) and the discontinuity in slope of the string under tension $T$.

Using Eqs. (1) and (3) from (5) we obtain the system of equations

$$\frac{dz}{dt} = -\frac{2T}{cm} g(t) + V,$$

(6)

$$\frac{dg}{dt} = \frac{cF_0}{2T} \left[ u^e(t) - \frac{\varepsilon}{\tau_0} \int_0^t u^e(\xi) \exp \left( \frac{\xi - t}{\tau_0} \right) d\xi \right].$$

(7)

The hammer felt compression $u$ is determined here by

$$u(t) = x(t) - y(0, t).$$

(8)

The initial conditions at $t = 0$, the moment when the hammer first contacts the string, are taken to be $g(0) = x(0) = 0$, and $dx(0)/dt = V$, the initial hammer speed.

SPECTRUM OF THE STRING MOTION

It is naturally of great interest to predict spectra of the string motion also. One way is to calculate the string deflection $y(x, t)$ and velocity $v(x, t)$ at the moment $t_0$ of the last loss of contact between the hammer and the string and then to use these data as initial conditions to find the mode amplitudes for the free motion of the string. The second way is the calculation of the mode energy spectrum directly from the force.
Suppose the total string disturbance is written as
\[ y(x, t) = \sum_{n=1}^{\infty} C_n(t) \sin \frac{n\pi(x - l)}{L}, \] (9)
to satisfy the boundary conditions. If the force is concentrated at a single point \( x - l \), then the force distribution upon the string is \( F(t) \delta(x) \). It is easy to show that \( C_n \) is given by
\[ C_n = A_n(t) \sin(\omega_n t) - B_n(t) \cos(\omega_n t), \] (10)
where
\[ A_n = \frac{2 \sin(\alpha \pi \xi)}{\pi c \mu} \int_0^L F(s) \cos(\omega_n s) \, ds, \quad B_n = \frac{2 \sin(\alpha \pi \xi)}{\pi c \mu} \int_0^L F(s) \sin(\omega_n s) \, ds. \] (11)
The general expression for the string mode energy is derived in [9] as
\[ E_n = \frac{M \omega_n^2}{4}(A_n^2 + B_n^2), \] (12)
and the mode energy level is determined by
\[ E L_n = 10 \log(E_n/E_0). \] (13)
Here \( M = \mu L \) is the total string mass; \( \omega_n = n\omega_0 / L = n \omega_0 \) is the string mode angular frequency; \( E_0 = mV^2/2 \) is the initial hammer energy.

A few of the spectra found in this way would be shown in the following section.

The system of Eqs. (6), (7) was solved numerically by a modified Euler's method. In case of \( \varepsilon = 0 \) we obtain first model of the interaction of nonhereditary hammer with a flexible string. This model is very popular now and it is often used by various authors [2-4, 7]. The comparison of this nonhysteretic hammer model with the second model described the interaction of the hereditary hammer (\( \varepsilon > 0 \)) with a flexible string is presented below.

TWO MODELS COMPARISON

The results of the hammer-string interaction for nonhysteretic and hereditary hammers for note A3 (\( f_0 = 220 \) Hz) are shown in Fig. 1. In this case the hammer strikes only one string of a three-string set of A3 note. The values of the primary parameters of the string and the hammer for this note are: \( L = 777 \) mm; \( l = 91 \) mm; \( T = 834 \) N; \( \mu = 7.1 \) g/m; \( m = 10.5 \) g; \( F_0 = 3 \) kN/mm²; \( p = 3.3 \); \( \tau_0 = 7 \) \mu s. For nonhysteretic hammer the hereditary constant is equal to \( \varepsilon = 0 \), and for hereditary hammer \( \varepsilon = 0.956 \). The mass of the hammer \( m \) is the effective mass, which is measured in the same way as in Ref. 7. The force-time characteristics and the spectra of the string vibrations are shown for initial hammer velocities \( V = 1 \) m/s and \( V = 4 \) m/s.

For this note A3 the difference between two models observed is rather essential. The process of the hammer-string interaction is more long in time for hereditary hammer, than for nonhysteretic one. The peaks due to the waves traveling along the string are less for hereditary hammer, and the whole curve is "smoother", than for nonhysteretic hammer. This fact may be explained by the particular features of the hereditary materials. Such materials are very deformable for the slow rates of the impact force. It is more noticeable for the hammer velocity \( V = 1 \) m/s. The result of this effect in frequency domain shows a short spectrum of the string vibrations excited by a hereditary hammer. Furthermore, the spectral slope for the hereditary hammer is approximately twice steeper than for nonhysteretic hammer.

The harmonics attenuation for flexible string excited by a nonhysteretic hammer is rather small at all. On the contrary, the number of harmonics excited in the real grand piano string is not so much. To obtain the steeper spectral slope a some more complicated models of the string with a stiff and damping terms are used [3,4] for numerical simulation of experiments.
Fig. 1. Force histories (a), (c) and spectra envelopes (b), (d) for nonlinear hammer and flexible string. △, and solid lines represent the results for nonhysteretic hammer; ◊, and dashed for hereditary hammer.

The analysis of the results obtained shows that the nonlinear hereditary model of the hammer provides a good practical way to make predictions about the vibration spectra of struck strings that come closer to measurements in real pianos and without any additional damping in the string.

REFERENCES

RECENT ADVANCES IN SPEECH PROCESSING TECHNOLOGY

Sadaoki Furui
NTT Human Interface Laboratories, 3-9-11, Midori-cho, Musashino-shi, Tokyo, 180 Japan

SUMMARY
This paper reviews recent speech processing technologies, including speech recognition, synthesis and coding, and discusses their current capabilities and appropriate applications. It also describes the most important research problems, and tries to forecast where progress will be made in the near future and what applications will become commonplace as a result of the increased capabilities. Robust performance in speech recognition and more flexibility in synthesizing speech will continue to be major problems that must be solved expeditiously.

1. INTRODUCTION
For the majority of mankind, producing and understanding speech is quite natural and these unconsciously acquired processes are performed quickly and effectively every day. Speech recognition, synthesis, and coding systems are expected to play important roles in an advanced multi-media society with user-friendly human-machine interfaces. Speech recognition systems include not only those that recognize messages but also those that recognize the identity of the speaker. Services using these systems will include voice dialing, database access and management, guidance and transactions, automated reservations, various order-made services, dictation and editing, electronic secretarial assistance, robots, automated interpreting (translating) telephony, security control, digital cellular communications, and aids for the handicapped (e.g., reading aids for the blind and hearing aids for the vocally handicapped).

Figure 1 shows a typical structure for task-specific voice control and dialog systems [8]. Although the speech recognizer, which converts spoken input into text, and the language analyzer, which extracts meaning from text, are separated into two boxes in the figure, they should work closely together, since the recognizer must use the language analysis in order to obtain the correct text. How to combine these two functions is a major problem, especially in conversational speech recognition (understanding). The meanings extracted by the language analyzer are used to drive an expert system to select the desired action, issue commands to various systems, and receive data from these systems. Replies from the expert system are sent to a text generator which constructs reply texts. These are then converted into speech by a text-to-speech synthesizer. "Synthesis from concept" is performed by the combination of the text generator and the text-to-speech synthesizer.

![Diagram of speech processing system]

Figure 1 - Typical structure for task-specific voice control and dialog systems.
Figure 2 shows hierarchical relationships among the various types of speech recognition, synthesis, and coding technology [3]. The higher the level is, the more abstract the information is. This figure is closely related to Fig. 1; speech recognition/understanding is the process extending upward from the bottom to one of the higher levels of Fig. 2, and speech synthesis is the process progressing downward from one of the higher levels to the bottom. Historically, speech technology originated from the bottom, and has developed toward the extraction and handling of higher-level information. Some of the technologies indicated in the figure remain to be investigated.

2. SPEECH RECOGNITION

2.1 Overview

Typical large-vocabulary continuous speech recognition systems currently under investigation consist of parts for signal processing, feature extraction, phoneme recognition, word recognition, and sentence recognition [20]. A speech wave is first converted into digital form in the signal processing part, and then converted into a time series of feature parameters, such as cepstra and delta-cepstra, in the feature extraction part. Although various methods for extracting and using prosodic features, that is, time functions of voice pitch (height) and amplitude, have been investigated, no satisfactory method has yet been invented.

The system predicts a sentence (hypothesis) that is likely to be spoken by the user, based on the current topic, the meaning of words, and language grammar, and represents the sentence as a sequence of words. This sequence is then converted into a sequence of phoneme models which were created beforehand in a training stage. Each phoneme model is typically represented by an HMM. The likelihood (probability) of producing the time series of feature parameters from the sequence of the phoneme models is calculated, and combined with the linguistic likelihood (appropriateness) of the hypothesized sentence to calculate the overall likelihood (probability) that the sentence was uttered by the speaker. The (overall) likelihood is calculated for other sentence hypotheses, and the sentence with the highest likelihood score is chosen as the recognition result. The recent incorporation of stochastic language modeling has greatly improved the accuracy of speaker-independent, large-vocabulary, continuous speech recognition.

In most of the current advanced systems, the recognition process is performed top-down, that is, driven by linguistic knowledge. Based on this principle, a very large vocabulary continuous speech recognition system with an 80,000-word vocabulary was built as a key element of a multi-modal dialogue system for telephone directory assistance at NTT Labs. It is based on the HMM-LR algorithm using HMMs as phoneme models and a generalized LR parser as a language model. To cope with the problem of background noise, an HMM composition technique was investigated and incorporated into the system [16].

2.2 Dynamic spectral features

Psychological and physiological research into human speech perception mechanisms shows that the human hearing organs are highly sensitive to changes in sounds, that is, to transitional (dynamic) sounds, and that the transitional features of the speech spectrum and the speech wave play crucial roles in phoneme perception [2].

- 498 -
The length of the time windows in which sound transitions are perceived have a hierarchical structure and range from the order of several milliseconds to several seconds. The hierarchical layers correspond to various speech features, such as phonemes, syllables, and prosodic features. It has also been reported that the human hearing mechanism perceives a target value estimated from the transitional information extracted using dynamic spectral features.

The representation of the dynamic characteristics of speech waves and spectra has been studied, and several useful methods have been proposed. However, the performance of these methods is not yet satisfactory, and most of the successful speech analysis methods developed thus far assume a stationary signal. It is still very difficult to relate time functions of pitch and energy to perceptual prosodic information. If good methods for representing the dynamics of speech associated with various time lengths are discovered, they should have a substantial impact on the course of speech research.

2.3 Robust speech recognition

It is crucial to establish methods that are robust against voice variation due to individuality, the physical and psychological condition of the speaker, telephone sets, microphones, network characteristics, additive background noise, speaking styles, and so on [6][11]. It is also important for the systems to impose few restrictions on tasks and vocabulary. To solve these problems, it is essential to develop automatic adaptation techniques.

Extraction and normalization of (adaptation to) voice individuality is one of the most important issues [5]. A small percentage of people occasionally cause systems to produce exceptionally low recognition rates. This is an example of the "sheep and goats" phenomenon. Speaker normalization (adaptation) methods can usually be classified into supervised (text-dependent) and unsupervised (text-independent) methods. Experiments have shown that people can adapt to a new speaker’s voice after hearing just a few syllables, irrespective of the phonetic content of the syllables [13].

2.4 Language modeling

Stochastic language modeling, such as bigrams and trigrams, has been a very powerful tool, so it would be very effective to extend its utility by incorporating semantic knowledge. It would also be useful to integrate unification grammars and context-free grammars for efficient word prediction. Adaptation of linguistic models according to tasks and topics [14] is also a very important issue, since collecting a large linguistic database for every new task is difficult and costly.

2.5 Spontaneous speech recognition

One of the most important issues for speech recognition is how to create language models (rules) for spoken language. When recognizing spontaneous speech in dialogs, it is necessary to deal with variations that are not encountered when recognizing speech that is read from texts. These variations include extraneous words, out-of-vocabulary words, ungrammatical sentences, botched utterances, restarts, repetitions, and style shifts. It is crucial to develop robust and flexible parsing algorithms that match the characteristics of spontaneous speech. How to extract contextual information, predict users’ responses, and focus on key words are very difficult and important issues.

Style shifting is also an important problem in spontaneous speech recognition. In typical laboratory experiments, speakers are reading lists of words rather than trying to accomplish a real task. Users actually trying to accomplish a task, however, use a different linguistic style.

2.6 Speaker recognition

Speaker recognition accuracy has been improved by using HMMs, likelihood normalization, and the text-prompted method [7][22]. The text-prompted method was recently proposed at NTT Labs to cope with the problem that conventional systems are easily defeated by a recorded voice [15]. In this method, a new text is prompted by the system every time the system is used. The system accepts the input utterance only when it determines that the registered speaker uttered the prompted sentence. That is, the system not only recognizes speakers, but also rejects utterances whose text differs from the prompted text, even if it is uttered by a registered speaker. Because the vocabulary is unlimited, prospective impostors cannot know in advance the sentence they will be prompted to say. A recorded and played-back voice can thus be correctly rejected.

2.7 Minimum error discriminative training

Unlike conventional maximum likelihood training, which estimates a model based only on training utterances from the same category, a discriminative training approach takes into account models of other competing categories and formulates the optimization criterion such that category separation is enhanced and the classification/recognition error rate on the training data is directly minimized. The optimization solution is obtained with a generalized probabilistic descent algorithm. This method is, therefore, called MCE/GPD (minimum classification error with generalized probabilistic descent). Unlike the Bayesian framework, it does not require estimation of probability distributions, which usually cannot be reliably obtained. This method has been successfully applied in various experimental studies for both speech and speaker recognition [12].

- 499 -
3. SPEECH SYNTHESIS
3.1 Overview
A typical structure of a text-to-speech synthesizer consists of text-preprocessing, text-to-phoneme conversion (letter-to-sound), prosodic contour generation (pitch and duration), and speech signal reconstruction (synthesis). Although a large number of text-to-speech synthesis systems are commercially available and they can generally provide intelligible speech, they all still lack naturalness. The quality of sound attained has recently jumped up, as the result of using waveform concatenation methods for speech signal reconstruction, but it is still a long way from a naturally-articulated voice. The most serious problems are in the first two stages. First, the reading of a written text is often error-prone, such as incorrect pronunciation of proper names, foreign words, abbreviations, acronyms etc. Second we do not have accurate rules for prosody control. Prosody is the major cause of the unnaturalness of current synthesized speech. It tends to be perceived as monotonous or jerky, and synthesized speech fails to transmit "meaning" [24].

3.2 Text preprocessing and text-to-phoneme conversion
In general, the first two modules for converting letters to phonemes combine heuristic rules and the utilization of an orthographic-phonetic dictionary. Actually, the pronunciations of many words and symbols, especially abbreviations, are word-dependent and so ambiguous that they are only clarified through their context. Overcoming this problem will require dictionaries and methods that are much more complex and sophisticated than those currently used. They must include rules based on a large database and natural language understanding algorithms taking into account the syntactic, semantic, and pragmatic aspects. It is obvious that the accuracy of pronunciation depends on how well the message content or the surrounding context is understood.

3.3 Prosodic contour generation
Correct prediction of stress location for accented languages needs the removal of grammatical ambiguities in utterances (e.g., verb/noun ambiguities). Although an adequate prosodic counter for each utterance could ultimately be generated by its meaning, at least partial syntactic parsing of the utterance is indispensable. The lack of sufficient syntactical information leads in many cases to prosodic segmentation errors, which are unacceptable for listeners.

Prosodic contours are generally specified by rules, either directly elaborated from real contour analysis or derived from formal models. It could also be possible to achieve further improvement in the reconstruction of prosodic contours by optimal use of prosodic databases compiled from live recordings.

3.4 Speech signal reconstruction
Most currently available commercial systems rely on the source/filter theory of speech production. This theory regards speech as being produced by the excitation of a linear filter (vocal tract) by one or more sound sources (excitation sources). Typical examples are LPC (linear predictive coding) and formant based synthesizers. However, the rules for dynamically controlling the source and filter parameters have not been adequately specified, so it is difficult to produce natural sounding synthesized speech by controlling these parameters with rules.

One way to reduce the number of rules to be specified for vocal tract evolution is to store pre-recorded speech segments, which are then concatenated to produce sentences. The problems to be solved are the following: choosing optimal speech segments to store, concatenating the segments without discontinuities at the concatenation points, and modifying prosodic parameters without degrading the sound quality. Diphone-type units [19] and context-dependent phone units [9] have been used to cope with the coarticulation problem. Units of the most frequent phoneme sequences having varying length have also been tried [23].

A new method, which directly stores and concatenates speech waveforms instead of transforming them into source and filter parameters, has recently been proposed, and the naturalness of the resulting sound was notably improved [10][18]. This method uses a pitch synchronous processing approach, in which the speech waveform is decomposed into a sequence of overlapping pitch-length signals. Prosodic modifications take place directly in the time-domain (pitch synchronous overlap-add technique). Although this method requires a large memory, its computational cost is very low. Diphone [18] or context-dependent phone [10] based approach has been combined with this method to cope with the coarticulation problem.

3.5 Controlling synthesized voice quality
The problems for speech synthesis include controlling synthesized voice quality and choosing an individual speaking style, multi-lingual and multi-dialectal synthesis, choice of application-oriented speaking styles, and the addition of emotion. In current commercial speech synthesizers, voice quality can be selected from male, female, and children's voices. No system has been constructed, however, that can precisely select or control the synthesized voice quality. Automatic creation of new voices from a given voice is an interesting and important issue especially for future automatic translation telephony. Research into the mechanism underlying voice quality, including voice individuality, is thus needed in order that a synthesized voice is capable of imitating a desired speaker's voice or to provide any desired voice quality such as harshness or softness.
4. SPEECH CODING

4.1 Overview

Various efforts have been devoted to speech coding, that is, reducing the transmission rate (or memory requirement) for speech without degrading its quality. In the last few years, the demand for speech coding has become very high, especially for digital cellular communication systems (mobile and portable telephones). Progress in reducing the transmission rate is shown in Fig. 3 [17]. The horizontal axis shows the year in which each coding method was developed or approved as a standard.

![Figure 3 - Progress in transmission rate reduction in speech coding.](image)

Most new low-bit rate coders belong to the class of CELP (Code-Excited Linear Prediction) coders. In both LPC vocoder and CELP, the speech waveform is generated by a linear prediction synthesis filter with an excitation source filter. The vocoder models the excitation source by either a pulse sequence of pitch period or a fixed noise sequence. In CELP, however, a codebook is used, and the best excitation vector is selected so that the perceptually weighted distortion between input and the synthesized speech is minimized. In addition, a periodic signal is generated by repeating the excitation signal in the previous frame, and a non-periodic signal is selected from a random excitation codebook. Naturalness and robustness are significantly improved by CELP.

Several years ago, extensive efforts to develop new coding algorithms and error correction techniques as well as the progress of LSI technology enabled the 8kb/s codec (coder and decoder) to be used for digital mobile telephones. In 1990, the VSELP (Vector Sum Excitation Linear Prediction) codec was selected as a full-rate speech codec standard for digital cellular systems both in North America and Japan. VSELP is a CELP coding method invented in Motorola, in which the excitation source is generated by summing several fixed basis vectors.

However, the digital cellular system is expected to run out of radio channels in Japan if the demand continues to grow at the current rate. Therefore, as a result of extensive research, NTT group has developed the PSI-CELP (Pitch Synchronous Innovation CELP) speech coding method and has achieved quality comparable to the current standard (VSELP), while using half the bit rate. PSI-CELP has been chosen as the half-rate codec standard for digital cellular system in Japan [17].

4.2 PSI-CELP speech coder for the digital cellular system

One of the most important features of the PSI-CELP speech coder is that random excitation vectors are given pitch periodicity for voiced speech by a pitch synchronization process. This feature is the origin of the name PSI-CELP. It significantly reduces the distortion and improves the quality for voiced speech, especially for female voices. In addition to the pitch synchronization, a two-channel conjugate structure and fixed codebook for transient speech signals were newly developed. These two features reduce the memory requirement, enhance robustness against channel errors, and improve the quality for the transient part.

5. FUTURE PROSPECTS OF SPEECH TECHNOLOGY

5.1 Ultimate speech recognition and synthesis systems

Ultimate speech recognition systems should be capable of robust, speaker-independent or speaker-adaptive, continuous speech recognition. Speaker recognition techniques are expected to be widely used in the future as
methods of verifying the claimed identity in telephone banking and shopping services, information retrieval services, remote access to computers, credit-card calls, etc.

Ultimate speech synthesis and recognition systems that are really useful and comfortable for users should match or exceed human capabilities. That is, they should be faster, more accurate, more intelligent, more knowledgeable, less expensive, and easier to communicate with than human staff. For this purpose, the ultimate systems must be able to handle conceptual information, the highest level of information in Fig. 2.

It is, however, neither necessary nor useful to try to use speech for every kind of input and output in computerized systems. Although speech is the fastest and easiest means of input and output for a simple exchange of information with computers, it is inferior to other means in conveying complex information. There needs to be an optimal division of roles and cooperation in a multimedia environment that includes images, text, tactile signals, handwriting, etc.

5.2 Articulatory and perceptual constraints

To solve various problems, it is necessary to promote sure and steady research and development by grasping the essence of speech phenomena, instead of developing methods by simply looking at them superficially. Speech technology is related to many scientific and engineering fields, such as physiology and psychology of speech production and perception, acoustics (physics), signal processing, communication and information theory, computer science, pattern recognition, and linguistics; it has an inter-disciplinary nature. It can also be said that speech research exists at the boundary between natural science and engineering. Knowledge and technology from a wide range of areas, including the use of articulatory and perceptual constraints, will be necessary to develop speech technology.

For example, when several phonemes or syllables are continuously spoken, as in the case of usual sentence speech, the tongue, jaw, lips, etc. move asynchronously in parallel, and yet with coupled relationships. Current speech analysis techniques, however, represent speech as a simple time series of spectra. It will become necessary to analyze speech by decomposing it into several hidden factors based on speech production mechanisms [1]. This approach seems to be essential for solving the coarticulation problem, one of the most important problems in both speech synthesis and recognition. It will also be necessary to clarify the process by which human beings understand spoken language, in order to obtain hints for constructing language models for spoken language, which is very different from written language. It is necessary to be able to analyze context and accept ungrammatical sentences.

The human hearing system is far more robust than machine systems - more robust not only against the direct influence of additive noise but also against speech variations (that is, the indirect influence of noise), even if the noise is very inconsistent. Speech recognizers are therefore expected to become more robust when the front end uses models of human hearing. This can be done by imitating the physiological organs or by reproducing psychoacoustic characteristics.

5.3 Speech science vs. statistical approaches

There is no doubt that most recent progress in speech and speaker recognition, even in speech synthesis, came from statistical approaches, such as HMM and stochastic language modeling. These approaches were made possible by recent remarkable progress in computer power. Statistical approaches are usually more reliable and, in many cases, more powerful than knowledge-based approaches, provided that we can obtain a large enough database. However, there is always some limit to the size of the database and we always encounter some discrepancy between the training database and the testing data. Therefore, even in the statistical approaches we need reasonable models.

Although it is not always necessary or efficient for speech synthesis/recognition systems to directly imitate human speech production and perception mechanisms, it will become more important in the near future to build mathematical models based on these mechanisms in order to improve performance [3].

5.4 Telecommunications applications

The most promising application area for speech technology is telecommunications [21]. The fields of communications, computing, and networking are now converging in the form of personal information/communication terminals. In the near future, personal communications services (known as PCS) will become popular, and everybody will have their own portable telephone. Several technologies will play major roles in this communications revolution, but one of the key ones will be speech processing technology. By using advancing speech synthesis/recognition technology, telephone sets will become useful personal terminals for communicating with computer systems.

In speech coding, a speech codec at less than 4k/s is now available, as described above. This will be useful not only for digital cellular systems, but also for various applications in multimedia communications. Low-cost mass-produced LSI chips for cellular systems are expected to be useful for various voice storage systems, voice response system, voice mail, and multiplexed transmission systems for private networks.

5.5 Objective evaluation methods
It is important to establish methods for measuring the quality of speech synthesis/recognition systems. Objective evaluation methods that ensure quantitative comparison of a broad range of techniques are essential to technological development in the speech processing field. Evaluation methods can be classified into the following two categories [4].

- Task evaluation: creating a measure capable of evaluating the complexity and difficulty of tasks
- Technique evaluation: formulating both subjective and objective methods for evaluating techniques and algorithms for speech processing

Task evaluation is very important for speech recognition, since the performances of recognition techniques can be compared only when they are properly normalized by the difficulty of the task. Although several measures for task evaluation have already been proposed, such as word and phoneme perplexity, none of them is good enough at evaluating the difficulty in understanding the meanings of sentences. It may be very difficult to achieve a reliable measure for such purposes, since it involves quantifying all sources of linguistic variability.

Technique evaluation must take the viewpoint of improving the human-machine interface. Ease of human-machine interaction must be properly measured. Recognition systems with the lowest error rates are not always best. There are various tradeoffs among the types of errors, such as substitution, insertion, and deletion. Even if the error rate is relatively high, the systems may be acceptable if the error tendency is natural and matches the principles of human hearing and perception. It is crucial that recognition errors are easy to correct and the system does not repeat the same errors.

To correctly evaluate technologies and to achieve steady progress, it is important to comprehensively evaluate techniques under actual field conditions instead of under a single controlled laboratory condition. Even when recognition systems perform well in laboratory evaluations and during demonstrations to prospective clients, they often do not perform nearly as well in the "real world". This is mainly because the speech that actually has to be recognized varies for many reasons and therefore usually differs from the training speech.

6. CONCLUSION

Speech recognition, synthesis and coding systems are expected to play important roles in an advanced multi-media society with user-friendly human-machine interfaces. This paper reviewed the most important research problems to be solved in order to achieve ultimate recognition, synthesis, and coding systems. The problems include dynamic spectral features, robustness against various voice variation, adaptation/normalization techniques, language modeling, and MCE/GPD approach in speech/speaker recognition, text processing, signal processing and speaking style control in synthesis, synthesis from concepts, use of articulatory and perceptual constraints, and evaluation methods.

Although speech recognition, synthesis and coding research has thus far been done independently for the most part, there will be increasing interaction between these aspects until common problems are being investigated and solved simultaneously.

REFERENCES


DESIGN OF PIANO PERFORMANCES IN STRUCTURED EXPRESSION

Tomoyasu Taguti
Faculty of Science, Konan University, Kobe, Japan

SUMMARY

Music performance for critical listening is a realization of player’s musical interpretation with expressive elements such as dynamics, timing, etc., well organized in a global perspective as well as in every detail of local variations both in time and voice-parts. Computational methods for expressive performance of music provide us with a means for systematic study on the mapping structure of acoustic quantities into the space of subjective impressions. This paper presents the case of piano performance through a dedicated music language for the piano and piano-like musical instruments. The content covers a brief summary of this language[1] (called MUSE for music in structured expression) followed by two examples of piano performance realized with the MUSE system.

A MUSIC LANGUAGE FOR THE PIANO

Physical variables of piano performance are the depression and release of piano keys with individual striking velocities at onsets, and the depths of the damper and shifting pedals. Here, the damper pedal controls the resonance/absorption of vibration of piano strings belonging to all non-depressed keys, and the shifting pedal locates the hammer position to strike whether the full number of unison piano strings (three or two for each) or less. There is a third pedal called the sostenuto pedal, and also there is another physical variable for the contact velocity of damper on release of piano key. But we set them aside for simplicity.

In MUSE, a piano score is represented in a generalized list of note names (88 usual pitch names, and additionally, the rest and two pedal names) which are structured in horizontal/vertical configuration having vertically arbitrary depths, where every note name as well as every sublist is assigned a duration in the unit of beats, either integral or fractional. Five kinds of parameters are provided to musically modify this note list: (1) dynamics that defines the loudness of notes; (2) depth that defines the depth of pedals; (3) agogics that defines the temporal variation of performing speed; (4) articulation that defines the degree of staccato/legato; and (5) shift that defines the temporal displacement in reference to the beat. Here, the first four are specified with envelopes, $e(s)$, which are functions in the metrical time $s$, or beats. Each of these parameters is bound to the entire note list, as well as to any sublists down to the individual note names, where the principle of superposition in each category is applied to evaluate these parameters bound to sublists at different depths.

CONTROL OF TIME

Expressive timing for notes and pedals is specified with the combination of agogics, articulation, and shift parameters. We shall explain here these parameters in some detail.

(Aagogics) This parameter maps the metrical time $s$ to the physical time $t$ through an integral formula of envelope $e(s)$ in such a way that it stretches the duration in the portion $e(s) > 0$
and shrinks where $c(s) < 0$. Let $I = [0, K]$ be an interval in the metrical time axis $s$, and let $e_0(s)$ and $e_1(s)$ be envelopes defined on $I$. Further, let $M$ be a nominal metronomic figure in beats/minute. We consider the following two mappings $s \mapsto t$:

$$t^{(0)}(s) = M^{-1} \times \int_0^s (1 + c_0(\sigma))d\sigma$$

$$t^{(1)}(s) = M^{-1} \times K \times \left[ \int_0^K (1 + e_1(\sigma))d\sigma \right]^{-1} \times \int_0^s (1 + c_1(\sigma))d\sigma.$$

The $t^{(0)}(s)$ maps $I$ to the interval $[0, T']$ ($T' = t^{(0)}(K)$), where $T'$ may be larger or smaller than $T = M^{-1}K$ depending on particular choices of $e_0(s)$. On the other hand, $t^{(1)}(s)$ maps $I$ always to $[0, T]$ regardless of choices of $e_1(s)$. We call the $t^{(0)}$-type mapping an open agogics, and the $t^{(1)}$-type mapping a confined agogics. The open agogics realizes a performance by deviating freely from the uniform tempo such as accelerando and ritardando. The confined agogics realizes the tempo to deviate freely also, but it keeps a strict global tempo over the prescribed metrical timespan $K$. Because of this latter property, the confined agogics is viewed as a mathematical representation of the "give-and-take" principle.

One remark worth to mention is that the expressive timing often depends on a smooth, indistinguishable deviation from the score values of notes found even in very short passages and note groups, such as a single bar in triple time that comprises 6 eighth notes, a trill spanning only one or two beats, a single beat that comprises 4 sixteenth notes, etc. Its origin may be attributed to motor process to a certain degree. The confined agogics is an effective means to cope with these situations also. In this sense, it is more than the classical give-and-take principle that implies a tempo change made intentionally and perceived distinctively.

For a hierarchical use of agogics parameters, the mapping $s \mapsto t$ is as follows. Let $e_0(s)$ be the agogics envelope bound to the top-most level (i.e., depth 0) of a generalized note list, and let $e_d(s)$ be the envelope bound to sublist at depth $d$, where (sub)list at depth $d$ subsumes sublist at depth $d+1$, $d = 0, 1, \ldots$. The formula at arbitrary depth $D$ is constructed on the principle that the open agogics is applied to depth 0, and the confined agogics are applied to all depths below 0 down to $D$, the envelopes being superimposed to depth $D$:

$$t(s) = M^{-1} \times K' \times K''^{-1} \times \int_0^s (1 + \Sigma_{d=0}^D e_d(\sigma))d\sigma$$

where

$$K' = \int_0^K (1 + e_0(\sigma))d\sigma \quad \text{and} \quad K'' = \int_0^K (1 + \Sigma_{d=0}^D e_d(\sigma))d\sigma.$$

Obviously, $t(s)$ is reduced to $t^{(0)}(s)$ in case $D = 0$, or to $t^{(1)}(s)$ in case $D = 1$ and $e_0(s) \equiv 0$.

The performing speed is given by

$$ds/dt = M \times K' \times K'' \times (1 + \Sigma_{d=0}^D e_d(s))^{-1}.$$

The quantity $ds/dt$ is called the local, or instantaneous, metronomic speed.

For note $i$ (in sublist at depth $D$) that is assigned $v$ beats with no staccato or legato, the onset and termination times, $t^m_i$ and $t^m_i$ respectively, are given as $t^m_i = t(s_i)$ and $t^m_i = t(s_i + v)$ if it starts at $s = s_i$ where $0 \leq s_i < K$ and $s_i + v < K$. The note-wise performing speed, $m_i$, is computed as $m_i = v/(t^m_i - t^m_i)$, which is a discrete analog to $ds/dt$ on $[s_i, s_i + v]$. Notice that in the measurement of real performances, the inter-onset interval, $IOI$, usually substitutes for $t^m_i - t^m_i$, that is, the note-wise performing speed is identified as $v/IOI$.

The formula $t(s)$ is general, and its usefulness depends on how to structure the target passages. In most cases, use of two levels (i.e., $D = 1$) seems mandatory to get a convincing response. An example is the use of depth 1 (confined agogics) to create an agogic-induced
metrical rhythm within each bar, and depth 0 (open agogics) to finish with phrasing expression. Reference [2] utilizes this notion of two-level agogics for identification of timing structure in real performances of Chopin's waltz. If there are two voice-parts at depth 1 in the above example and these two have slightly different envelopes, possibly one of them being modified slightly with another envelope at depth 2, then they proceed at slightly different performing speeds, but do have a meeting finish always. This kind of asynchrony will add a special effect in a lyrical performance of melodic passage with accompanying part. Example 2 below utilizes this technique.

(Articulation) This parameter modifies $t_{ag}$ as determined by the agogics parameter. Let $a$ ($a < 1$) be the value of envelope (of the superimposed one, in general) evaluated at $s = s_i$. The formula is $t_{ag} = at(s_i) + (1 - a)t(s_i + v)$. Apparently, $a > 0$ implies staccato, and $a < 0$ superlegato. An envelope bound to a passage, i.e., to a (sub)list, modifies all the constituent notes. For example, a continuous envelope which is large in amplitude and takes values from positive to negative as $s$ proceeds defines a smooth change from acute staccato to deep superlegato over the passage; this technique applied to fast passages will produce a mastery control in timbre change due to a changing degree of detach/overlap of successive tones.

(Shift) This parameter modifies $t_{sh}$ as determined by the agogics parameter. It has two types, relative (in beats), and absolute (in milliseconds), where either may be to the right (positive) or to the left (negative). Let $d$ be the value in beats, and let $d'$ be the value in ms. These can be specified independently as well as hierarchically, and the principle of superposition is applied if it is specified at different depths hierarchically. The resulting formula, in minutes, is $t_{sh} = t(s_i) + M^{-1}d + d'/60000$. It is utilized to define the onsets of grace notes, to adjust the timing of depression of the damper pedal, to break the synchronous onset of notes in the beginning of a phrase, to introduce asynchrony among the notes in a chord, etc.

ENVELOPE

The envelope $e(s)$ is assumed to be a piecewise linear function with or without break points; at break points, it may be either continuous or discontinuous. Allowing discontinuities for $e(s)$ arises from a practical requirement to put a sudden change in parametric values not only in dynamics, depth, and articulation but also in agogics. Further, $e(s)$ is assumed to be right continuous at discontinuous points. This condition conforms to the evaluation rule that the values of expressive parameters (other than agogics) to be assigned to notes and pedals are those evaluated at their onsets.

COMPUTATIONAL PERFORMANCE WITH MUSE

Quantitative analysis of expressive performance have been extensively conducted by many scientists in these decades, and several performance rules have been proposed. Still, many secrets in artistry seem to remain unexplored. This may be partly due to insufficient resolution in physical measurement of real performances for finer analysis, and partly due to lack of adequate models for identifying target musical expressions. Our standpoint is to utilize known rules, possibly with necessary modifications, together with hypothetical rules, part of which are mentioned above, in the form of categorized parameters in structured representation. The following shows two examples. Because of limited space, we shall discuss only the temporal profile of performing speed.

EXAMPLE 1 — PIANO SONATA No.17 in B flat major, K. 570, Mozart

Music by Mozart are characterized by a quick, clear transition of feeling, e.g., calm to vigorous, joyous to depressed, or tense to relaxed, within a short timespan. These feelings require a delicate control over physical variables where the control of timing seems is essential.

Figure 1 presents the note-wise speed in bars 1 ~ 20 of K. 570 realized with MUSE. There are three main rules used: the “slow-fast-slow” (SFS), “long notes longer and short notes
shorter" (LL/SS), and "diminished-half-note to quarter note" (DH-Q) rules. The third one is that the 2:1 construction of two notes in triple time is played as 1.6 ~ 1.8 : 1 in normal.

These rules are specified superimposingly. The following describes their usage. The SFS rule is applied to phrases of bars 1 to 4, bars 5 to 12, bars 13 to 16, and bars 17 to 20, as well as to subphrases (bars 5 to 6, bar 7, bar 8, bars 9 to 10) and small note groups (beat 3 in bars 12, 13, 14, 16, 17, and 18). The LL/SS rule is applied to the timespan of bars 5 through 10 being contrasted to that of bars 1 to 4; and to beat 3 being contrasted to beats 1 and 2 of bars 12, 13, 14, 16, 17, and 18. The DH-Q rule is applied to each of bars 1 to 3, and to bar 11, (though bar 11 is not the case of 2 : 1 construction.)

EXAMPLE 2 — Nocturne No.1 in B flat minor, Op.9 No.1, Chopin

Music by Chopin generally require a large amplitude of deviations in expressive elements. Change in the performing speed and dynamics, pedaling (timbre change), and asynchronism all constitute the essential part of the expressiveness.

Figure 2 presents the note-wise speed in bars 1 ~ 8 of Nocturne No.1 realized with MUSE. Main rules, superimposed in 4 levels at maximum, are SFS, LL/SS, and DH-Q, together with the "slow to fast" (SF) rule that is applied to a sequence of notes with equal note value and equal pitch. A passage with 33 small notes in bars 2 to 3 is covered with 6 SFS's.

CONCLUSION

A descriptive approach for expressive performance of piano music was presented with emphasis on a hierarchical representation of expressiveness parameters. The method of timing control was discussed theoretically and in practice with two demonstrations.

A CONSIDERATION ON THE TIMBRE OF COMPLEX TONES ONLY CONSISTING OF HIGHER HARMONICS

Hisashi Uematsu, Kenji Ozawa, Yōiti Suzuki, Toshio Sone
Research Institute of Electrical Communication, Tohoku University,
2-1-1, Katahira, Aoba-ku, Sendai 980-77, Japan.

SUMMARY
Experiments on the difference limens (DLs) for timbre of complex tones consisting of only high harmonics were carried out by changing their spectral structures. The DLs were then estimated using two models of band-pass filters which partially overlapped each other and were compared with those obtained by experiments. The DLs estimated with roex filters agreed well with the experimental ones.

INTRODUCTION
Timbre is one of the most important and basic psychoacoustic features of sounds, and its perception reflects information processing in the auditory system. However, it is difficult to estimate the timbre of sounds quantitatively from their physical features, though some attempts have been made [1, 2]. To establish a method for estimating a mutual configuration of timbre of sounds from their physical characteristics is interesting and seems to be of value from the engineering point of view; this method could be applied for the design of musical instruments, audio equipment, studios, and halls. We therefore study the perception process of timbre with the intention of establishing such a method.

While low harmonics of a complex tone can be heard separately, high harmonics are bundled into bands when they are perceived [3–5]. In this paper, we refer to such perception or analysis in the human auditory system as "band analysis." The main purpose of this study was to investigate the role of this band analysis in the perception of timbre. With this purpose in mind, the difference limens for timbre were measured by changing the spectral composition of complex tones consisting only of high harmonics. This experiment was similar to those conducted in profile analysis [6], though in original profile analysis, the focus seems to be on the discrimination of the shape of the spectral structure.

EXPERIMENTS ON THE DIFFERENCE LIMENS FOR TIMBRE OF HIGH HARMONIC COMPLEX TONES

stimuli
A complex tone consisting of the 10th to 50th harmonics with identical intensity and a fundamental frequency of 200 Hz was used as the standard stimulus. The initial phase angles among the
components were randomly determined and set through experiments to avoid a subject's using temporal information as a cue. The overall level of the stimulus was set at 60 dB SPL. One or two components with a frequency difference of $\Delta f$ Hz in the standard stimulus were enhanced in amplitude and such tones were used as the test stimuli. The difference limens (DLs) for timbre between the standard and the test stimuli with various $\Delta f$ were measured. The center frequencies of the two components of the test stimuli were just or as near as 6.0 kHz for every $\Delta f$. In addition, control experiments were carried out with stimuli in which only one of the two components that were enhanced in the main experiment was enhanced.

**procedure**

Experiments were carried out with the constant method. The standard and the test stimuli were presented monaurally to subjects seated in a soundproof room through a headphone (YAMAHA YHID-3). The temporal pattern of the stimuli is shown in Fig. 1. The subjects were requested to judge whether the timbre of the two stimuli were distinguishable or not.

```
<table>
<thead>
<tr>
<th>standard stimulus</th>
<th>test stimulus</th>
<th>judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0 s</td>
<td>1.0 s</td>
<td>1.5 s</td>
</tr>
<tr>
<td>0.5 s</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

*Fig. 1 Temporal pattern of a pair of stimuli*

The number of levels for the enhancement of the two components in the test stimuli was seven and the number of repetitions was twelve. The DLs were determined with the method of maximum likelihood estimation [7]. Two males and a female with normal hearing in their twenties participated in the experiments.

**results and discussion**

The results of one of the three subjects is shown in Fig. 2. The three subjects showed similar results. In this figure, the abscissa indicates the frequency difference $\Delta f$ between the enhanced components. The point at 0 kHz corresponds to the experimental condition in which only one component of 6.0 kHz was enhanced. The ordinate indicates the DL for timbre between the standard and the test stimuli. The solid line with circles represents results of the main experiment. The dotted and dashed lines indicate the results of the control experiments. The dotted line shows the DLs when only the higher frequency component was enhanced. On the other hand, the dashed line indicates the DLs when only the lower frequency component was enhanced. The error bars show the ±1 s.d..

It can be clearly seen from the results that the DLs for $\Delta f = 200$ Hz and 400 Hz, where 6.0 kHz and 6.2 kHz components and 6.0 kHz and 6.4 kHz components were enhanced respectively, are
smaller than those of the two control experiments, while the DLs for large Δf correspond to one of the results of the two control experiments, i.e., the result with smaller DL. These results indicate that the increment of intensity of one of the two enhanced components was perceived for large Δf, while the total increment of intensity of both enhanced components was perceived for small Δf. From these results, we consider that high harmonics were perceived due to band analysis in the perception process of timbre.

ESTIMATES OF THE DIFFERENCE LIMENS BY USING BAND-PASS-FILTER MODELS

We tried to estimate the DLs for timbre using a band-pass-filter-bank model. Two kinds of filters are considered; one is an ideal rectangular-shaped band-pass filter having a bandwidth equal to the equivalent rectangular bandwidth (ERB) and arranged at intervals of 0.1 on the E scale [8], and the other is a roex filter [8] which is also arranged at intervals of 0.1 on the E scale. We assumed that the subject could perceive the difference in timbre when the amount of the increment of the output level of any band of the filters exceeded a critical value. We also assumed that the critical value was equal to the lower DL of the two in the control experiments.

Figure 3 shows the results of the estimation for the same subject as in Fig. 2 when rectangular-shaped filters were used. The estimated values agree well with the experimental results for large Δf. On the other hand, the estimated values tend to be smaller than the experimental results for small Δf. Figure 4 shows results of the estimation when roex filters were used. The estimated values for small Δf are closer to the experimental results than those estimated with the ideal rectangular-shaped band-pass filters. The discrepancy between the two filters can be explained as follows. When the ideal rectangular-shaped filter was used, the output level of the filter with a center frequency of 6.0 kHz was kept constant so long as the Δf was smaller than the bandwidth of the filter. When roex filters were used, on the other hand, the output level became a little smaller as Δf increased even if the Δf was smaller than the equivalent bandwidth of the filter. Therefore, the estimated values for small Δf are nearer to the experimental results. These results indicate that use of filters with gradual cut-off is valid for interpreting the timbre perception process.

CONCLUSION

In this study we investigated the role of band analysis in the auditory system in the difference limens for timbre of complex tones consisting of only high harmonics. The results indicate that several components of the tones cannot be analyzed separately but are perceived based on band analysis. We then attempted to estimate the difference limens for timbre by using two kinds of
filters that partially overlap. The results of estimation with roex filters seem to agree well with the experimental results. This means that use of filters with gradual cut-off imitating the human auditory system is appropriate for explanation of the experimental results.

ACKNOWLEDGEMENT

This study was partially supported by a Grant-in-Aid for Scientific Research (No. 0670057) by the Ministry of Education, Science and Culture of Japan.

REFERENCE


- 512 -
MEASUREMENT, ESTIMATION, AND MODELING OF WIND INSTRUMENTS USING DSP TECHNIQUES

Vesa Välimäki 1, 2, Matti Karjalainen 1, 2, Jyri Huopanen 1
3 Stanford University, CCRMA, Stanford, CA 94305, USA
2 CARTES, Ahertajankoja 4, FIN-02100 Espoo, Finland
1 Helsinki Univ. Tech., Lab. Acoustics & Audio Signal Processing, Espoo, Finland

SUMMARY

Physical modeling is a modern approach to musical acoustics and model-based sound synthesis of musical instruments. In this paper we apply digital signal processing (DSP) and physical modeling techniques to study the behavior of wind instruments. A DSP-based measurement system is used to obtain impulse responses for wave propagation, reflection, and radiation in the instrument bore, its terminations, and discontinuities. Inverse filtering techniques are then applied to estimate the pressure and volume velocity signals inside the instrument when played normally. The results can be used to estimate the parameters of a synthesis model.

SIGNAL PROCESSING MODELS OF WIND INSTRUMENTS

Digital waveguide modeling is a powerful synthesis technique for musical instruments that is based on a digital filter representation of wave propagation [1]. Related signal processing methods are also well suited to model-based analysis. For linear subsystems both time and frequency-domain processing are available and for nonlinear elements numeric time-domain processing is often the most useful approach. Figure 1 shows a generic wind instrument and its waveguide model (applicable to woodwinds and brass instruments). Signal propagation is assumed to be linear everywhere except in the reed portion.

A digital waveguide model implements a cylindrical bore section as a pair of lossless delay lines, one for each direction, so that the losses and dispersion are commuted to the termination filters (e.g., $R(z)$). This makes the model computationally very efficient [1]. Conical bores may also be modeled effectively [2], [3], and a more complex profile can be approximated as a cascade of cylindrical and conical sections. Finger holes are side branches in the waveguide model and the bell or any remarkable opening forms a combination of a reflection filter $R(z)$ and a radiation filter $T(z)$.

There is an inherent implementation problem in digital models of wave propagation due to the fixed unit delay length, which is determined by the sampling rate. The length of the tube and the positions of finger holes do not in general match a multiple of the discrete unit delay. Fractional delay (FD) approximation using interpolation solves this problem [4], [5]. We have shown various techniques to use FD filters in instrument models, including fractional delay waveguide filters [5], [6] that are able to simulate scattering junctions in arbitrary points between cylindrical and conical tube sections.

Fig. 1. A generic wind instrument and its signal processing model.

- 513 -
Fig. 2. A system for the measurement of wave transmission, reflection, and radiation in a wind instrument where the mouthpiece has been replaced with a prolongation tube.

The complexity of wind instrument modeling is related primarily to the reed portion that behaves highly nonlinearly. Since other parts of the instrument fairly well meet the condition of linearity and time-invariance, the excitation signal from the reed may be analyzed by inverse filtering the radiated signal by an estimated transfer function from the reed to the radiation field.

DSP-BASED MEASUREMENT TECHNIQUES

We have measured transfer functions and impulse responses due to wave propagation in a tube, as well as reflection and radiation at the termination, using DSP techniques in the following way (partly similar to [7]). Let us consider pressure signals in a linear and time-invariant system (see Fig. 2). The mouthpiece or reed portion of the instrument is replaced with a prolongation tube having the same diameter so that the tubes are acoustically matched. An approximation of a unit impulse is generated by a miniature transducer (a Walkman headphone driver) and its response is registered by miniature electret pressure microphones (Sennheiser K211-8) in positions $M_1$, $M_2$, and $M_3$. Averaging, e.g., for 100 repetitions, is used to improve the dynamic range of the results.

A computer program driving 16-bit stereo A/D and D/A converters controls the measurement process. A sampling rate of 22 kHz is used which yields a frequency band up to 10 kHz. Impulse responses are computed by deconvolution in the frequency domain:

$$h(n) = F^{-1}\{F[y(n)] / F[x(n)]\}$$

(1)

where $x(n)$ is the reference signal (e.g., at the position $M_1$) and $y(n)$ is the response signal (e.g., at $M_2$), $F$ is the discrete Fourier transform operator (implemented in practice using the Fast Fourier Transform) and $F^{-1}$ is the inverse discrete Fourier transform operator. Care should be taken to avoid problems due to the circularity of discrete-time operations. Reference signal $x(n)$ should always have a relatively flat spectrum with no zeros.

The microphone pair $M_1$, $M_2$ is used to measure the wave transmission transfer function per unit length in the prolongation tube. A single microphone $M_2$ registers both the excitation traveling to and the signal returning from the instrument bore. When the path from $M_2$ to junction $J$ is compensated off twice (by deconvolution), we obtain the reflection impulse response $h_R(n)$ at junction $J$. Similarly, transmission impulse response $h_T(n)$ from $J$ to $M_2$ is computed from the pair $M_2$, $M_3$ by single compensation for the path from $M_2$ to $J$. Further analysis may be based on the two responses $h_R(n)$ and $h_T(n)$ with Fourier transforms $H_R(\omega)$ and $H_T(\omega)$, respectively.

A major issue with this procedure is to isolate the reference signal $x(n)$ from the response $y(n)$ (see Eq. (1)) in each case since they have been superimposed in the measured signals and may overlap in time. This separation must be done by applying careful time-domain windowing to proper signal portions before the deconvolution. In our measurements the main problem was the spreading of the excitation due to ringing of the transducer at its fundamental mechanical resonance frequency around 200 Hz. This spreading can be countered by using a long enough prolongation tube (for better temporal isolation) or by preprocessing the excitation to the transducer in order to shorten its output, e.g., by using the reciprocal $F^{-1}\{1/F[h(n)]\}$ of its impulse response $h(n)$.

As an example, Fig. 3 depicts the reflection impulse response $h_R(n)$ for a clarinet (note $E_1$, all holes closed) and the corresponding magnitude response $|H_R(\omega)|$. Notice in $h_R(n)$ the series of closed finger hole reflections as well as the main reflection from the open termination.
Fig. 3. a) Reflection function of a clarinet as measured with the system of Fig. 2. b) The corresponding magnitude spectrum.

The measured reflection response $h_R(n)$ may be utilized in estimating and designing the loop delays and the termination reflection filters of a waveguide synthesis model. The main reflection is used to find the total loop delay. In an efficient real-time synthesis system, every finger hole is not modeled but only the first open one. In such a case the closed finger hole reflections cannot be implemented. Only the overall loop transfer function is approximated in the reflection and radiation filters. The transmission response $h_T(n)$ may be utilized to design a digital filter for modeling the sound radiation.

Other interesting properties of the instrument can also be analyzed from $h_R(n)$ and $h_T(n)$. As an example, the acoustic impedance of the instrument bore, as seen from point J, may be derived. A pressure unit impulse $p(n)$ with Fourier transform $P(\omega)$, when driven into the bore, reflects back as $P(\omega)H_R(\omega)$. These are summed for the overall pressure signal at J. The acoustic impedance of the prolongation tube is $Z_0 = \rho c/A$, where $\rho$ is the air density, $c$ is the sound velocity, and $A$ is the cross-sectional area of the tube. The excitation $P(\omega)$ will result in a volume velocity $P(\omega)/Z_0$ initially and $-P(\omega)H_R(\omega)/Z_0$ due to the reflected signal. Thus the impedance of the bore at junction J is the ratio of pressure and volume velocity, i.e.,

$$Z_J(\omega) = Z_0[1 + H_R(\omega)]/[1 - H_R(\omega)]$$

(2)

INVERSE FILTERING TECHNIQUES

Inverse filtering has long been used in speech research (see, e.g., [8]) and it is also applied in string instrument modeling [9] to compute the source excitation from the output when a reasonably good estimate of the source-to-output transfer function is available. The same technique is applied here to woodwind and brass instruments in order to estimate the traveling wave components inside the bore on the basis of the radiated sound.

Let us assume that we have measured the transfer functions $H_R(\omega)$ and $H_T(\omega)$ accurately. The transmission function $H_T(\omega)$ is useful as such since it may simply be applied to inverse filtering in order to compute the pressure wave at junction J traveling to the right, denoted by $P^+_{J}(\omega)$, from the measured pressure wave $P_{3}(\omega)$ at the microphone position M3.

$$P^+_{J}(\omega) = P_{3}(\omega)/H_T(\omega)$$

(3)

We may derive one more signal relation for inverse filtering purposes. The pressure signal $P^-_{J}(\omega)$ at point J traveling to the left can be expressed as

$$P^-_{J}(\omega) = P^-_{J}(\omega)H_R(\omega) = P_{3}(\omega)H_R(\omega)/H_T(\omega)$$

(4)

The derivation shows that we may compute both pressure signal components at the entrance of the tube when we know the radiated sound pressure $p_3(n)$ and the two primary transfer functions. The sum of these components gives the total sound pressure. This may be used to check the validity and accuracy of the method whenever the sum pressure can be measured directly.
The above derivation is valid for the case where the bore has been impedance matched to a driving prolongation tube. One may ask whether the inverse filtering method is valid also during normal playing conditions when the highly nonlinear reed portion feeds the instrument bore. If it is assumed that the instrument is linear and time-invariant from point J to the right, then the inverse filtering is theoretically valid for all driving conditions, since the two primary transfer functions are not dependent on what will be connected to the left of point J. (We must also assume that there are no other signal paths from J to M.) This is not true for, e.g., the flute. This procedure—like deconvolution in general—is critical to the accuracy of the measurements and estimated transfer functions. There are various sources of inaccuracy that may disturb the final results. In addition to the ones mentioned above, one specific problem is how to guarantee that the transfer functions estimated from the instrument without a mouthpiece remain the same during normal playing conditions. Among the possible variations are the air temperature and related sound velocity, as well as changed humidity. The third factor is the possibility of nonlinearities in the bore due to high pressure levels.

CONCLUSIONS

In this paper we have presented a method to measure and analyze the behavior of wind instruments using the physical modeling approach and digital signal processing techniques. Besides being able to efficiently and precisely measure the transfer functions and impedance curves of a wind instrument, it is possible to apply inverse filtering—deconvolution by division in the frequency domain—to analyze the signal components close to the reed portion during normal playing and thus collect information on the nonlinear behavior of the reed. The results of the measurements and estimation techniques can be used to calibrate a parametric synthesis model.

ACKNOWLEDGMENTS

This study was financially supported by the Academy of Finland.

REFERENCES

BRASS WIND INSTRUMENT QUALITY MEASURED AND EVALUATED BY A NEW COMPUTER SYSTEM

Gregor Widholm
Institut für Wiener Klangstil, Hochschule für Musik und darstellende Kunst in Wien.

SUMMARY
For the player of brass wind instruments, instrument quality is mainly defined by three parameters: intonation, responsiveness and timbre. The problem is, that there exists no direct relationship between such "player defined" quality criteria and the data obtained by physical measurements. This paper presents methods how to get results which are in accordance with practical playing experience and highly confirmed by professional players. Data acquisition and processing is done with a new developed software under Windows using a DSP-based PC-workstation.

INTRODUCTION
For the audience the definition of musical instrument quality is commonly reduced to the quality of the radiated sound in the far field. The listener can only valuate the sound of the instrument produced by the player. Quite different is the situation for the player himself, the quality of an instrument includes some more important aspects as there are: RESPONSIVENESS. Means: how easy it is for the player to produce a note with the instrument. Usually these values vary with the dynamic level and the frequency of the played note.

INTONATION. Means: how good do the notes -offered by the instrument- correspond with the equally tempered scale; an important criterium of quality of wind instruments.

TIMBRE. Easy to measure by using FFT algorithms. The assessment however depends on the purpose (classical music, jazz, etc.) and the individual imagination of a "beautiful timbre".

Beside these three main criteria the kind of action of the valves (influences the microstructure of a slur) can be an important criterium of selection by the player. For form’s sake it should be mentioned, that subjective aesthetic aspects like the design and colour of an instrument or the used material can sometimes highly influence the evaluation of quality by the player. Since our approach is the measurement and evaluation of objective quality criteria, such aesthetic aspects are not a subject of our interest, that’s up to each player himself.

RELATIONSHIP BETWEEN PLAYER DEFINED QUALITY AND ACOUSTICAL PARAMETERS
The problem is that we have to deal with quality criteria defined by musicians and expressed in their own terminology. There exists (beside the intonation) no monocausal relationship between such player defined criteria and the data obtained by acoustical
measurements. The complex relationship between musical quality parameters, mechanical parameters of the instrument and the data of acoustical measurements shows the schema below. As it can be seen, the player defined quality criteria can be calculated using the data of input impedance and pulse response measurement. General information about the acoustics of brasses, particularly about the relationship between input impedance data and playable notes of an instrument are given in [1], more details about the measurement arrangement and the conception of the used hard- and software can be found in [2]. To get useful data the measurement of input impedance has to be done in the plain of the player's lips inside the mouthpiece.

As the peaks of such an impedance pattern (Fig. 1) mark the frequencies of the playable notes of the instrument, the deviation due to the equally tempered scale is easy to calculate. Measurements in the frequency domain like impedance measurements give already good informations on the intonation characteristic, but no direct information on parameters which are helpful to calculate values for the responsiveness, like the settling time of the instrument, its response or effects caused by unwanted reflections inside the tube. These transient phenomena are an important part of the acoustical behaviour of a brass instrument and highly influence the „playing feeling“ of the musician. Therefore measurements in the "time domain" have to be done. A combination of both measurements can provide with a sufficient quantity of data as a base material for subsequent processing.

As a realisation with an exactly defined short pulse (Dirac-pulse) causes high efforts of technical resources, we decided to calculate the pulse response of the instrument by processing the data of impedance measurement. An example is given in Fig. 2.

THE PROBLEM OF OBJECTIVE AND SUBJECTIVE QUALITY
Trying to verify the results of calculated intonation error and responsiveness by comparing the calculated values with the subjective assessment by players a new phenomenon occurs, which is well known by instrument makers because of their highly dependence on the valuation of their instruments by professional players: whilst player one is very impressed of the outstanding quality of an instrument, player two locates the same instrument only in the mid- or lower range of a quality scale (needless to say, that nearly all professional players are convinced, that their expert opinion has to
be taken as an objective "factual statement"). Who is right, player one or two? The solution of the problem: both are right! A wind instrument has only one objective quality but may have two or more subjective qualities. It is caused by the fact that the player and the instrument forms (from the physical point of view) a control loop. On the one hand the produced sound primary depends on the vibration characteristics of the individual player’s lips, on the other hand the lip vibrations (=excitation spectrum) are influenced by the impedance characteristics of the instrument. Within this system, the player is represented by its excitation spectrum.

INTONATION

1. Objective intonation error. For the calculation of intonation error due to the equally tempered scale, the frequencies corresponding to the impedance peaks are determined (the peaks 1, 7, 11, 13 are not included into this process). After that the pitch \(a^f = xxx \text{ Hz}\) is determined, where the sum of the deviation of the found frequencies due to the equally tempered scale is a minimum. Finally the deviation of each playable note (impedance peak) from the reference pitch -determined in step two- is calculated and displayed in "cent".

![Figure 3: Objective intonation of a horn. The bars indicate the deviation due to the tempered scale (color print).](image)

2. Subjective intonation error. According to the definition of the excitation signal during the measuring procedure, the calculated intonation values are valid only for a sinusoidal excitation. As the player excites the instrument with a spectrum including more or less harmonics (depending on the individual physiological setup of his lips and the dynamic level of the played note), not only the impedance peak of the fundamental frequency, but all impedance peaks corresponding with the frequencies of the harmonics of the excitation signal have to be taken under consideration. As the values of the amplitudes of the harmonics commonly are different, their contribution to the "over all value" has to be weighted according to their relative amplitude within the excitation spectrum. The result of such a procedure can be different intonation and responsiveness values for one instrument, depending on the musical dynamic level (piano, fortissimo) and the individual player’s lip setup. Numerous tests during the last five years proved the correctness of this method.

RESPONSIVENESS

Musicians commonly differentiate between something like an "over all responsiveness" of an instrument and the responsiveness of each single note. The value of the "over all responsiveness" corresponds with the quotient between the amplitude of the excitation pulse and the response of the bell (Fig.2). The responsiveness of a single note is primarily determined by the amplitude of the corresponding peak of the impedance curve and its ratio to the neighboring maxima and minima (Fig.1). Experiments proved that the influence of the Q-factor on the responsiveness values of given instruments seems to be less important.

- 519 -
As the player is represented through his excitation spectrum, values for individual responsiveness can be obtained by using the above described procedure.

TIMBRE
This is the only point where we needed the help of the player. The sound of the instrument (played by the musician) is analyzed by FFT. At present we work on a method for calculating radiated sound spectra by the use of the transfer function of the instrument and standardized or individual excitation spectra.

VALVE ACTION AND MICROSTRUCTURE OF SLURS
Depending on the kind of valve (perinet, rotary) and its location related to the mouthpiece/bell, the transients of slurs can sound like a glissando or they can be separated by a short noise band. The preference for one of these two valve systems depends on the kind of music to be interpreted and the individual musical taste.

Fig. 4: Difference between Perinet- and Rotary valve. Situation for the player’s lips for a slur from g2 to Bb2. (X-axis: Status of the valve from not engaged=left up to engaged=right, Y-axis: Impedance, Z-Axis: frequency) Performing a slur, the player has to press down the valve button and simultaneously change the tension of his lips. The starting point (g2) in the graph is located left/front, the target point (Bb2) right/back. The graph shows, that in this particular case the rotary valve produces a short noise band ("break down" of the standing wave system).

In such a way, the kind of valve performance can be determined without any player.

USED HARD- AND SOFTWARE
For "BIAS" (= Bras Instrument Analyzing System) a DSP-board is plugged into an ordinary PC. An external subsystem contains the ADC/DAC’s, filters and preamplifiers. The instrument is excited by a "multitone" via a special measuring head during a period of 2 seconds. Microphon 1 is located inside the mouthpiece and picks up the response of the instrument, micro 2 is located -for reference- above the loudspeaker inside the measuring head. Both channel data are sampled down, smoothed and processed by a 16k FFT. The quotient of the magnitude of both spectra is displayed as "impedance curve" on screen (Fig.1). Calculation of pulse response is done automatically, evaluation of intonation, responsiveness, timbre, etc. can be done by a mouse click. The software requires WINDOWS and is written in C and Visual Basic. The frequency resolution is 0.5 Hz, respectively 2 cm for the pulse response. The system is calibrated in acoustic Ohms.

REFERENCES
15th International Congress on Acoustics
Trondheim, Norway 26 - 30 June 1995

PSYCHOACOUSTICAL EVALUATION OF SYNTHESISED GUITAR TONES

Howard A.K. Wright, Bernard E. Richardson
University of Wales College of Cardiff

SUMMARY

Over the past ten years at Cardiff we have developed computer models that predict the vibrational behaviour of the guitar including its strings. Modelling of the three-dimensional sound radiation fields created by the vibrating structure has enabled us to synthesise the sound of plucked notes. The input data required by the model relates to the guitar's material properties, construction and shape. By varying any of these parameters we can model small changes in the instrument's physical characteristics and listen to the resulting differences in sound. Moreover we can model many different changes entirely independently.

Previous computational work at Cardiff showed that, to a good approximation, the radiation fields produced by the guitar were either monopoles or dipoles. By building the assumption of monopole and dipole fields into the model and omitting the numerically intensive radiation-field calculations we have been able to considerably reduce the amount of computer time needed to synthesise one note. This has made it feasible to synthesise a large number of different guitar tones so that psychoacoustical evaluation of the sound can be performed.

We have performed listening tests on a large number of guitar tones and have begun to find out the constructional changes to which the ear is most sensitive. This is the first step towards formally linking aspects of construction with an instrument's sound quality.

INTRODUCTION

The tone quality of the classical guitar is largely determined by the vibrational response of the body, comprising top plate, back plate and air cavity. The response of the body's component parts show resonances at a number of different frequencies. However the response of the guitar body is not well described by a simple summation of the responses of the separate components; coupling between the top plate, back plate and air cavity causes significant changes in the response of the complete instrument. To model the instrument accurately these coupling effects must be included.

Significant coupling effects are also introduced by the addition of a string attached at one end to the top plate. The resonant frequencies of the string modes or the body modes can be shifted, in some cases causing an unpleasant, dissonant sound. The interaction of the string and body must be properly accounted for in the model to obtain an accurate sound model of the complete instrument.

By building a numerical model of the instrument we can synthesise the sound of a plucked guitar note based on data relating to the construction of the instrument. The physical parameters relating to construction can be varied independently giving total freedom in investigating links between the sound of the guitar and single physical parameters. This is impossible when dealing with real instruments.

We have developed a model which calculates the response of the instrument from data about the resonances of the body and string. Coupling between all parts is accounted for, and fluid damping, caused by the mass of air which must be displaced by the vibrating instrument, has also been included. From the velocity response, the pressure radiated to a given point in space is calculated. Previous modelling of sound radiation using the Boundary Element Method (BEM) [1] gave detailed three-dimensional radiation fields for each of the vibrational modes of the top plate. These results showed that the modes radiated, to a good approximation, as monopoles or dipoles or a combination of both. In the current model, the time consuming BEM calculations are avoided by calculating the volume of air displaced by the vibrating body; this is used in a simple expression giving the pressure radiation at a distance r from a point source. Dipoles are treated by using two point sources which are separated by a distance d and are π radians out of phase.
By performing a reverse FFT on the complex pressure response, the impulse response of the instrument can be obtained. In reality the plucking force applied to the string will not be a delta function in time. The time duration of the force will cause the lower frequencies in the response to be excited more strongly than the higher frequencies. By taking the FFT of a more realistic plucking force and performing a convolution with the response of the instrument a more realistic sound is obtained.

The aim of this work was to synthesise a number of guitar tones on which listening tests could be performed. Subjects were asked whether they could perceive differences between pairs of tones generated from different sets of input data. A number of different input parameters were varied in this way and the results were used to try to establish the degree of change needed for each parameter before differences in sound were perceived.

**DETAILS OF THE MODEL**

The vibrational response of the body is calculated by modelling the motion of the top plate as that of a harmonic oscillator driving a piston. The back plate and air mass in the soundhole are treated similarly, following the work of Christensen [2], [3]. Each oscillator is defined by an effective mass \( m \), a resonant frequency \( \omega \) and a Q-value \( Q \) and drives a piston which has an effective area \( A \). These oscillators are coupled together via the common pressure changes in the air cavity; movement of one of the pistons causes a pressure change in the cavity which results in a force on the other two pistons. This gives rise to the following three coupling constants

\[
\alpha_{at} = \mu A_a A_t \quad \alpha_{ab} = \mu A_a A_b \quad \alpha_{bt} = \mu A_b A_t
\]

where \( \mu = \frac{\rho}{c} \), \( \rho \) is the density of air, \( c \) is the speed of sound in air and \( V \) is the cavity volume.

A sinusoidal driving force \( F \) of frequency \( \omega \) is applied to the top plate. Solution of the equations of motion of the three oscillators leads to the following expression for the top plate displacement:

\[
z_t = \frac{F(D_b D_a - \alpha_{ab})}{D_1 D_a D_b + 2\alpha_{at} \alpha_{ab} \alpha_{bt} - [D_1 (\alpha_{ab})^2 + D_b (\alpha_{at})^2 + D_a (\alpha_{bt})^2]} \tag{1}
\]

where subscript \( t \) refers to the top plate, \( b \) to the back plate and \( a \) to the air cavity. \( D_t \) is defined as \( D_t = m [\omega_1^2 - \omega^2 + i \gamma_1 \omega] \) where \( \gamma_1 \) is a damping factor. \( D_b \) and \( D_a \) are defined similarly.

The reaction of the vibrating body to the air mass surrounding it causes additional small perturbations in mode frequencies, and more importantly, increased damping of the body vibrations. Pierce [4] gives the fluid reaction force of a piston of radius \( a \) mounted on an infinite expanlar baffle as

\[
\hat{F}_{\text{fluid}} = \rho e \omega a^2 \left[ \frac{ka}{2} + i \frac{ka}{3\pi} \right] \tag{2}
\]

where \( k \) is the wavenumber and \( e \) is the velocity of the piston. Incorporating this into the modelling scheme leads to an expression similar to equation 1 where \( D_t \) is replaced throughout with \( X_t \), which is defined as \( X_t = D_t + \omega \rho e \omega a^2 \left[ \frac{ka}{2} + i \frac{ka}{3\pi} \right] \). \( D_a \) and \( D_b \) are similarly replaced with \( X_a \) and \( X_b \).

When the driving force \( F \) is replaced by the force exerted by a vibrating string important coupling effects between string and top plate must be considered. The string is attached to a rigid support at one end and to the top plate at the other; a sinusoidal driving force is applied at a point on the string. Following the work of Gough [5] the coupling between string and body is accounted for by considering the power in the vibrating string, the power supplied by the driving force and the power transferred to the body. Note that only the vertically polarised transverse string modes are considered. This gives an expression giving the Fourier coefficient \( a_n \) of the \( n \)th string mode in terms of the displacement of the top plate.

Finally we arrive at the following expression for the displacement of the top plate coupled to one string mode, one back plate mode and one air cavity mode and including fluid damping. A summation over all modes of interest gives the total top plate displacement:

\[
z_t = \frac{2 \pi k_n F \sin(ka \pi) (X_t X_a - \alpha_{at})}{D_t \left[ X_t X_a X_t + 2\alpha_{at} \alpha_{ab} \alpha_{bt} - (X_t (\alpha_{at})^2 + X_a (\alpha_{bt})^2 + X_b (\alpha_{bt})^2)] - 27^2 \kappa^2 (X_t X_a - \alpha_{at})^2} \tag{3}
\]

Figure 1 shows two response curves synthesised from the model; one is the body response, the other is the response of the body coupled to a single string. The taller, sharper peaks correspond to the string resonances and other peaks come from the body resonances. In this example a string mode at 660 Hz coincides with a body peak. This string-body coupling means that the string's vibrational energy is transferred rapidly to radiated sound energy which may produce a short lived and unpleasant sound. The string mode at 330 Hz, which is not strongly coupled to the body, has a decay time which is twice that of the 660 Hz mode.

- 522 -
LISTENING TESTS

Listening tests were performed on 19 subjects to find out which changes in physical parameters resulted in the greatest perceived differences in sound. Four synthesised notes were put together into a chord; tests were carried out using chords of E major and D major so that notes from all six strings were used. Chords were presented in pairs—first the control chord, followed by the test chord. The test chord had only one of its input parameters different from the control set of parameters so that the effect of a single physical parameter on the tone quality could be investigated. 30 pairs of chords were presented to each subject and they were asked whether they perceived a difference in sound or whether they perceived no difference. The 30 chord pairs were presented three times to obtain some idea of the consistency of a subjects' response. Included in these 90 chord pairs were six pairs which had identical chords.

Body mode information for the control chord was taken from work done previously at Cardiff using finite element analysis of a 2.6 mm thick spruce guitar top plate. Sixteen variations of these body parameters were used in generating the different test chords; all 16 variations were used to generate the E major chords and the first 12 variations used to generate the D major chords.

The notes synthesised for the chords covered a frequency range of 2.5 kHz i.e. the sampling frequency was 5 kHz. Each note lasted for 3.3 seconds. Body mode information up to 1 kHz and all string modes up to 2.5 kHz were included. The input values for the string mode frequencies were made anharmonic to account for the effects of string stiffness, and the Q values of the string modes were set so that the higher modes decayed more rapidly.

RESULTS

The results of the listening test are shown in Table 1. Note that all parameter changes refer to the fundamental top and back modes or the fundamental air resonance. Each subject's results were used only where identical responses were given to the three repeats of each chord pair. Each subject made consistent responses to around 20 of the 30 chord pairs.

The first point to note is that for most of the physical changes investigated, the trend in results for the D and E chords is similar indicating that many of the changes are not frequency specific. The most notable exception is parameter no. 2, where six subjects perceived a difference in the E chord (against three who did not), but no subjects perceived a difference in the D chord and 11 subjects perceived no difference.

Another general observation is that an increase and decrease of a certain parameter by the same amount does not necessarily lead to similar perceived changes in sound. This is particularly clear in parameters 7 and 8 where the reduction of the top plate mass by a factor of 1.5 leads to most people perceiving a difference in both E and D chords. An increase of the same parameter by the same factor gives results showing that most people perceived no difference.

From Table 1 we find that the parameters which caused the greatest perceived changes include the frequency of the fundamental air resonance ($f_s$) and the frequency, effective mass and area of the top plate fundamental resonance ($f_s$, $m_s$ and $A_s$). The ear seems less sensitive to changes in Q-value of the top plate fundamental mode ($Q_s$), and frequency and effective mass of the back plate fundamental mode ($f_b$ and $m_b$).
<table>
<thead>
<tr>
<th>Parameter Number</th>
<th>Parameter Change</th>
<th>E chords</th>
<th>D chords</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>$f_1$ up 2 tones</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>$f_1$ down 2 tones</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>$f_1$ up 2 tones</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>$f_1$ down 2 tones</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>$Q_1 \times 3$</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>$Q_1 \div 3$</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>$m_1 \times 1.5$</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>$m_1 \div 1.5$</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>$m_1 \times 2$</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>$A_1 \times 1.5$</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>$A_1 \div 1.5$</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>$A_1 \times -1$</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>13</td>
<td>$s_1$ up 2 tones</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>$s_1$ down 2 tones</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>$m_2 \times 2$</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>$m_2 \div 2$</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Results of the Listening Test for 15 subjects: Columns marked No show the number of subjects who perceived no difference with this change; the Yes column marked the number who perceived a difference.

**CONCLUSIONS**

The numerical model outlined in this paper has been shown to be a useful tool in examining links between measurable physical properties of a guitar and the sound of the instrument. The listening tests performed so far have measured responses of subjects to 16 individual parameter changes and the results have highlighted those parameters to which the ear is relatively sensitive.

The changes investigated involve the fundamental top plate, back plate and air resonances only, although input data involving 10 top plate modes was used. The results should be applicable to classical guitars in general since the interaction of the fundamental air resonance and the fundamental top plate mode results in very similar features in the low frequency response of the instrument.

In the ongoing work at Cardiff we are interested in finding out more about the role that the low order modes play in radiating not only low-frequency information, but also high-frequency string harmonics. Previous work by Brooke [1] has shown that when high-frequency string modes are coupled to the body, low-frequency body modes radiate this energy more effectively than body modes at frequencies close to the string modes. Low effective masses and large effective areas for these low order modes will ensure efficient radiation of high-frequency string harmonics.

The role that the body modes play in radiating low-frequency string information is rather different. As a single string mode moves up in frequency, passing through a particular body mode, the sound output will vary considerably. When the string and body mode coincide energy is transferred rapidly from the string to radiated sound resulting in a loud but fast decaying sound. A string mode that coincides with a body mode with a low effective mass will couple strongly and is likely to produce an unpleasant sound.

Further listening tests are planned to investigate the perceived influence of the low-order modes in radiating both low frequency and high frequency string vibrations. We hope to understand how the guitar maker can control effective masses to produce a guitar with good sound radiation at high frequencies yet without compromising the instrument's performance at low frequencies.

**REFERENCES**

A SYNTHESIS MODEL FOR THE FLUTE TRANSIENT
BY ANALYSIS OF REAL SOUND THROUGH
TIME-FREQUENCY METHODS

Sølvi Ystad & Richard Kronland-Martinet
CNRS - Laboratoire de Mécanique et d'Acoustique, 31 chemin
J. Aiguier, 13402 Marseille Cedex. 20 France.

I- Introduction

The flute instrument has been studied by numerous authors and some models based on the
physical characteristics of the instrument have been proposed. These models give a good
description of the relevant phenomena generated by the flute and allow synthesizing of sounds with
good quality. Nevertheless, these models often are too simple to efficiently take into account the
interpretation of the flutist. This is due to the difficulties of modelling the complicated physics
related to the instrument and the way it is played (vortex phenomena especially at the embouchure,
wave propagation in a discontinuous tube, propagation loss and effects generated by the flutist).

In order to design a digital flute imitating as closely as possible a real one and permitting in
addition intimate transformations of the sound by altering the synthesis parameters, we look for a
signal model based on additive synthesis whose parameters are estimated by the analysis of real
sounds. As a first approach towards such a model, we will present some results about the attack
transient and the vibrato obtained with the help of time-frequency techniques and especially the
skeleton of the so-called wavelet transform. Such a technique makes it possible to estimate relevant
parameters such as frequency and amplitude modulation laws corresponding to each spectral
component.

II- some physical aspects of the flute

One of the key aspects characterizing flute-like instrument is the attack transient. Authors as
S. Balibar [1] and Rakowski [2] state that the attack transient is among the longest for wind
instruments, and that the period of growth of a flute tone can be divided in two parts; one where
the proper pitch cannot be detected, followed by a shorter part where the sensation of a pitch
suddenly appears to the listener. It is accompanied by a sudden change in speed of growth of the
harmonics. Spectral analysis of the flute sounds states that there is a cutoff frequency above which
the radiation behaviour of the instrument changes drastically. The expression of the cutoff
frequency of an open tone-hole lattice in a pipe has been given by Benade [3]. For a Boehm
chromatic system the numerical values for this frequency is \( f_c = 2150 \text{Hz} \) [4]. This cutoff frequency
also explains why the harmonic development is considerable for low notes, and why the number of
harmonic components decreases when playing high register notes.

By direct measurements, Fletcher [5] has shown that the pressure fluctuations producing
the vibrato causes a frequency variation of 5 Hz and an amplitude variation of about 10% of the
blowing pressure. Changes in frequency caused by pressure variations are found by considering
the jet admittance \( Y_j \) as a function of blowing pressure. An increase in blowing pressure causes
\( \text{Im} Y_j \) to be negative, giving a small increase in sounding frequency. Conversely, for a decrease in
blowing pressure, the frequency will fall. This explains the variation in frequency when playing a
vibrato, since the player uses pressure fluctuations to obtain this effect.

In the next part of the paper, we shall analyse these phenomena in order to propose a
model.

- 525 -
III. analysis of real sounds in their musical context

III.1 time-frequency techniques

Both the time behaviour and the frequency behaviour of the flute sound are relevant for hearing. From a signal processing point of view, this leads to describe the sound with the help of time-frequency analysis methods such as Gabor or Wavelet transforms. We shall briefly describe how the wavelet transform has been used for the estimation of amplitude and frequency modulation laws corresponding to real flute sounds.

The wavelet transform \( S(\tau,a) \) of a square-integrable signal \( s(t) \) is obtained by its scalar product with respect to a dilated and translated version of a basic wavelet \( g(t) \), also named analyzing wavelet [6]:

\[
S(\tau,a) = \left( \frac{1}{\sqrt{a}} g \left( \frac{t-\tau}{a} \right) \right) s(t) \, dt = M(b,a) \exp(i\Phi(\tau,a))
\]

Such a representation spreads out the information on a time-scale half-plane and shows how the energy and the phase are distributed for each component.

If one considers the signal as "asymptotic", i.e that it can be written:

\[
s(t) = A(t) \exp(i\phi(t)) \quad \text{with} \quad \left| \frac{\partial A}{\partial \tau} A^{-1} \right| \ll \frac{\partial \phi}{\partial \tau},
\]

one can show that the restriction of the transform along the trajectories (also called "skeleton" of the transform) defined by the sets of points \( t(\tau,a) \), satisfying the equation:

\[
\frac{\partial \Phi(\tau,a)}{\partial \tau} = \frac{\phi_\tau(0)}{a} \quad \text{(where} \phi_\tau(t) \text{is the phase of the wavelet} g(t))
\]

contains all the signal information [7]. Moreover, the modulus of the skeleton is related to the amplitude modulation law of the components of the sound, and the derivative of the phase with respect to the parameter \( \tau \) is related to the frequency modulation laws. This method has been used to estimate precisely both the amplitude and the frequency modulation laws discussed below. As an example, figure 1 shows respectively the amplitude and the frequency modulation laws corresponding to the third component of a 8 seconds A4 played with a strong vibrato and recorded in an anechoic room.

\[\text{Figure 1: Amplitude and frequency modulation laws of the 3rd component of A4.}\]

III.2 discussion of the results

In order to modelize the transient part of the sound (the attack), one looks for a correlation between the behaviour of the amplitude modulation law of each component and the total energy of the sound (which is related to the jet pressure). For that purpose, we have plotted the local energy

- 526 -
(with respect to time) of each harmonic of a sound versus the local energy of the whole sound. Figure 2 presents such a graph. One can see that the mean steepness of the curves is different as well as the deviation corresponding to the hysteresis effect (this deviation is due to the different laws driving the component energy during the attack and the decay of the sound). By plotting the mean steepness of these curves versus the corresponding harmonic rank (figure 3), one can roughly fit the corresponding curves with a first order polynomial. One can thus link the amplitude of the n-th harmonic $A_n$ to the amplitude $A_T$ of the sound by the law:

$$A_n = K_n + (0.75 + 0.3(n-1)) A_T$$

where $K_n$ is a constant depending on $n$.

Preliminary measurements have shown that $K_n$ increases roughly linearly with respect to $n$. Such a result would correspond to a delay between two successive components proportional to $1/n$. This is in accordance with the measurements made directly on the modulations laws of the harmonics (figure 4 left).

In our basic model, we have not taken into account the existence of an hysteresis effect, which roughly becomes more important as the rank of the harmonic component increases (figure 3 right). This shows that more energy is needed to excite higher modes than to maintain them, an effect that could be modelled by adjusting the $K_n$ in accordance to the "width" of the hysteresis, which implies the use of different $K_n$-values during the attack and the decay of the sound.

Figure 2: local energy of the harmonic (with respect to time) versus energy density of the whole sound. left: for the first harmonic of C3, right: for the second harmonic of C3.

Figure 3: left: mean steepness of the curves described in fig.2 versus the corresponding harmonic rank. right: mean width of the hysteresis effect versus the harmonic rank.
Another important feature of the flute sound concerns the vibrato effect. The measurement of the vibrato depth $\Delta \omega$ for each component have shown that the ratio $\Delta \omega/n$ is constant (figure 4:right) and that the vibrato is in phase with an amplitude variation (tremolo) (figure 1). This is related to the fact that the vibrato is due to jet pressure fluctuations [5]. One can thus pilot the frequency modulation by measuring the air jet pressure. Independently of the note played, the frequency of the vibrato is about 5Hz.

In the same way, we have tried to find out whether such a correlation between the amplitude of the sound and the frequency modulation exists during the very beginning of the sound. For the time being, we cannot conclude on this because, even if such correlations can be seen on the graphs, we have the impression that they are an artefact due to the existence of two close frequencies, as if the sound contained two harmonic sounds with a small difference in their fundamental frequency (less than 100Hz).

IV- conclusion

The analysis of real flute sounds has shown that the components involved during the transient attack of the sound follow laws related to the jet pressure. A rough model has been proposed, as well as future improvements taking into account the hysteresis effect. The vibrato has been correlated to the jet pressure variations, and we have shown that its depth is related to the frequency component. We have finally discussed the possibility of a superposition of 2 tones in the transient attack, leading to a correlation between the amplitude and frequency modulation laws.

Acknowledgements: This work has been supported by Norges Forskningsråd

References

- 528 -
AEROACOUSTICS, ATMOSPHERIC SOUND AND INFRASOUND
LINEARITY MEASURES FOR IN-DUCT ACOUSTIC ONE-PORT SOURCES

Hans Bodén
The Marcus Wallenberg Laboratory for Sound and Vibration Research,
Department of Vehicle Engineering, KTH, S-100044 Stockholm, Sweden

SUMMARY

Three different linearity coefficients for determining if an acoustic one-port source under test is linear are presented. Their sensitivity to random noise and ability to detect non-linearities are investigated by simulations and measurements.

INTRODUCTION

Methods to calculate the acoustic field generated in duct systems by fluid machines, e.g., pumps, fans, internal combustion engines, are of great interest for the design of mufflers and silencers and to gain a better understanding for the sound generating mechanisms in the machines. If a linear time-invariant model is applicable for the machine it can be described as an acoustic one-port source. In the frequency domain an acoustic one-port can be completely described by a source strength and a source impedance (or a reflection coefficient), see equation (1).

\[ p_r \cdot Z = p \cdot Z_s + p \cdot Z \quad (1) \]

where \( p_r \) is the source pressure, \( Z_s \) is the source impedance, \( p \) is the acoustic pressure at the outlet of the source and \( Z \) is the acoustic impedance of the rest of the system seen from the source. The source data is often determined experimentally. The measurement methods used to determine the source data of acoustic one-ports can be divided into methods "with an external source", e.g., the two-microphone method [1,2] and methods "without an external source", i.e. multi-load methods [3].

Many fluid machines such as compressors and IC-engines generate high sound pressure levels or high flow velocities, and it is not certain that linear models can be used to describe them. It is therefore useful in the experimental situation to have a method to determine whether a linear model can be used. Such a linearity test has been suggested for the methods with external source in Ref. 2. In this paper general linearity tests for all measurement methods are presented.

LINEARITY TESTS

If we assume that we have a problem with \( m \) complex unknowns and make \( n \) measurements the over-determined equation system for determining the unknowns \( x \) can be written in the following way

\[ A \cdot x = b \quad (2) \]
where \( A \) is an \( n \times m \) matrix, \( x \) is a \( m \times 1 \) vector and \( b \) is a \( n \times 1 \) vector. The number of measurements, \( n \) in equation (2), has to be larger than \( m \) for the linearity coefficients to be meaningful. If \( n \) equals \( m \) the linearity coefficients will always be equal to one. A linearity coefficient which is similar to the coherence function can be defined as

\[
\gamma^2 = x^{-1} \cdot x = b^{-1} \cdot A \cdot A^{-1} \cdot b,
\]

where \( y^{-1} \) is interpreted as the pseudo-inverse of \( y \). This linearity coefficient will have a value in the interval \( 0 \leq \gamma^2 \leq 1 \), where the upper limit represents a perfect linear relationship.

One alternative approach to define a linearity measure is to calculate all solutions for \( x \) using all possible combinations consisting of \( m \) of the \( n \) equations. The mean (\( \mu \)) and standard deviation (\( \sigma \)) of all the solutions can then be calculated for each component of \( x \). To get dimensionless quantities the standard deviation can be divided by the mean. This will give a result which is equal to zero when there is a perfect linear relationship and larger than zero otherwise. A linearity coefficient which has a value between zero and one and is equal to one when there is a perfect linear relationship can be defined in the following way.

\[
\gamma^2 = \frac{\mu^2}{\mu^2 + \sigma^2}.
\]

A third possibility is to use a dimensionless residual

\[
\varepsilon = \frac{(A \cdot \hat{x} - b)^* \cdot (A \cdot \hat{x} - b)}{(A \cdot \hat{x})^* \cdot (A \cdot \hat{x})},
\]

where \( \hat{x} \) is the estimate of \( x \) and \( (\cdot)^* \) stands for transposition and complex conjugation. A linearity coefficient which is equal to one when there is a perfect linear fit for the data and which goes to zero when there is a bad fit can be defined in the following way

\[
\gamma^2 = \frac{1}{1 + \varepsilon}.
\]

The suitability of the different linearity coefficients for detecting non-linearities will in the following sections be tested both by computer simulations and by measurements. All three linearity coefficients will also be affected by random noise disturbances. To separate the effects of noise and non-linearities in a measurement the number of averages and, if possible, the level of the excitation should be varied. The effect of random noise should decrease with an increased number of averages and the effect of non-linearities should increase if the level of excitation is increased.

NUMERICAL SIMULATIONS

To check how well the linearity coefficients are able to detect different degrees of non-linearities some numerical simulations have been made. Simulations made by adding random noise to simulated as well as measured data showed that all three linearity coefficients was affected in a similar way by noise. This is important since it means that any large differences in the calculated linearity coefficients is not caused by measurement noise and must therefore be caused by some
non-random type of disturbance, i.e., non-linearities. To investigate how the three linearity coefficients react to a non-linear system some simple tests have been made. In Figure 1 one interesting result is shown, where $A$ was a vector with six components with a large $(A = (1,10^3,10^6,10^9,10^{12},10^{15}))$ variation in the data and $b = A + A^\kappa$, where $\kappa$ is a constant which was varied. Two of the three linearity coefficients ($\gamma^2_r$ and $\gamma^3_r$) does still give identical results, as they did for random noise disturbances, when applied to non-linear data. It can also be seen that they can give a result close to unity even if the system is strongly non-linear and the variation in the data is large. It can therefore not be recommended to rely only on $\gamma^2_r$ and $\gamma^3_r$ in a measurement situation, but since the linearity coefficients use the same input data it is simple to calculate all three. If the result is equal to one then we know that there is a linear relationship between the measured data and it is noise free. How close to unity the linearity coefficients should be must be determined for each application.

**EXPERIMENTAL RESULTS**

**Impedance measurement** To test the linearity coefficient in an experiment where the measurement errors should be small and the system should be linear a standard impedance measurement was made. The data was measured using the two-microphone method and one of the measurement objects was a straight duct filled with absorbing material at the end. Figure 2 shows the calculated linearity coefficients. As expected all three linearity coefficients are very close to unity.

**Axial flow fan** The data presented in this section were originally used for determining the acoustic two-port source data for a fan [4] and has been reanalysed in order to calculate the one-port data and the linearity coefficients. In [4] the source data was determined using an external source and the two-microphone method. To get over-determination five different loads were used. An example of the results is shown in Figure 3. All previous experience [2,4] indicates that the axial flow fan is a linear acoustic source. The decrease in the linearity coefficients at certain frequencies is therefore probably caused by insufficient suppression of flow noise.

**Internal combustion engine** Data from earlier studies [5,6] using the two-load method has been reanalysed to calculate the linearity coefficients. Data from both the exhaust and the intake side has been analysed. For the exhaust side measurements were made for 20 loads and a number of different speeds and loads. Figure 4 shows the measured linearity coefficients which clearly indicate some non-linearity. This could be expected from earlier analysis of this data [5,6].

**CONCLUSIONS**

Three different linearity coefficients for determining if an acoustic one-port source under test is linear has been suggested. It has been shown that all three linearity coefficients have the same sensitivity to random noise. By simulations and measurements it has been shown that non-linearities can be detected. One of the three linearity coefficients seems to be a better choice than the other two since they can give misleading results for strongly non-linear systems.

**REFERENCES**


**Figure 1** Linearity coefficients for the system $A = b^x$; $\gamma_c^2$; $\gamma_e^2$; $\gamma_s^2$.

**Figure 2** Linearity coefficients for a standard impedance measurement, $\gamma_c^2$; $\gamma_s^2$.

**Figure 3** Linearity coefficients for a source impedance measurement on an axial fan; $\gamma_e^2$; $\gamma_c^2$.

**Figure 4** Linearity coefficients for a source data measurement on an IC-engine, $\gamma_e^2$; $\gamma_s^2$ for $Z_s$; $\gamma_e^2$ for $p_s$. 
THE EFFECT OF FLOW VARIATIONS ON ACOUSTIC RADIATION FROM A PLATE IN A VORTEX WAKE

M. K. Bull, A. M. Blazewicz, Y. Li, J. M. Pickles
Department of Mechanical Engineering, University of Adelaide,
South Australia 5005, Australia.

SUMMARY

Measurements have been made of the effect of flow velocity on the acoustic power radiation generated by flow over a tandem array of two thick plates. They indicate that this radiation, which is dominated by the flow in the gap between the plates, undergoes a dramatic increase in level when the flow in the gap changes from a trapped-vortex regime to a vortex-street regime.

INTRODUCTION

The investigation reported forms part of a more general study of the mechanisms of noise generation by bluff bodies immersed in vortex wakes. It relates particularly to two-dimensional flow over two thick plates in tandem, one in the vortex wake of the other, but has wider application - for example in the generation of aerodynamic noise by the teeth of circular saws.

Previous work [1] has shown that for such arrays the fluctuating forces on the downstream plate and the acoustic radiation have a dominant, essentially discrete-frequency, component at the frequency associated with vortex motion in the gap between the two plates. It has been concluded that the force fluctuations, produced by interaction between the vortex motion in the gap and the downstream plate, constitute the source of the acoustic radiation, which is essentially dipole in nature.

It has also been shown [2] that for a leading plate with a faired leading edge and a downstream plate of rectangular cross-section, the flow in the gap between the plates depends critically on the ratio \( G = g/t \) of the gap length \( g \) to the plate thickness \( t \) and on the Reynolds number of the flow \( Re_f = U g / \nu \) (where \( U \) is the flow velocity and \( \nu \) the fluid viscosity): in general, for \( G < 2 \), the gap flow consists of two large side-by-side counter-rotating vortices extending over the whole streamwise length of the gap (the "trapped-vortex" regime), while for \( G > 3 \) vortices are shed alternately from the top and bottom corners of the trailing edge of the upstream plate, forming a vortex street within the gap (the "vortex-street" regime). At intermediate values of \( G \) various forms of transitional flow, between the extremes of vortex-street flow at high Reynolds numbers and trapped-vortex flow at low Reynolds numbers, occur [3]. Regime boundaries in the \( G - Re_f \) plane are given in [2].

The main aim of the present work is to investigate the effects of changes in the flow regime in the gap on acoustic power radiation at the vortex frequency.

EXPERIMENTAL DETAILS

The test plates both had a thickness of 3.85 mm and a chord of 25.0 mm. The leading edge of the upstream plate was semi-elliptical (with major axis/minor axis = 5), and the trailing edge was square. They were mounted horizontally in tandem in an air jet issuing from a rectangular nozzle 30 mm x 67
The jet was confined between vertical side-walls 67 mm apart, 100 mm high, and extending 200 mm downstream of the jet exit, but was otherwise open to the surroundings. The test plates were mechanically isolated from the test rig: each passed through the side-walls confining the jet, and was supported outside the flow; the clearance space between the plate and the side-wall was sealed with soft rubber.

Air was supplied to the nozzle from a compressed-air reservoir which was allowed to run down during the experiment; the flow velocity thereby decreased continuously from about 220 m/s to about 60 m/s and the Reynolds number from \( Re_l = 56 \times 10^3 \) to \( 15 \times 10^3 \).

The rig was installed in a reverberant acoustic chamber (6.84 m x 5.57 m x 4.3 m high). Spatially-averaged spectral densities \( 0 \) of sound pressures were obtained from a traversing microphone. Known reverberation times of the acoustic chamber as a function of frequency allowed conversion to the corresponding spectral densities \( 0 \) of radiated acoustic power.

Measurements were made at gap settings in the range \( 2 \leq G \leq 6 \), although results are not presented for all gaps tested.

RESULTS AND DISCUSSION

As the velocity falls in the course of a test, the spectrum of the acoustic pressure \( 0 \) in the reverberation chamber exhibits a large, sharp, essentially discrete-frequency, peak at a frequency which decreases with the flow velocity and corresponds to the Strouhal number \( St_V = 0/2U \) of vortex motion in the gap between the plates, previously determined [1,2] by hot-wire measurements of velocity fluctuations in the gap. The variation with flow velocity of the corresponding spectral density of radiated acoustic power \( 0 \) at the frequency of vortex motion in the gap, for several different values of the gap parameter \( G \), is shown in figure 1.

For all gaps, \( 0 \) shows a series of distinct peaks and troughs over the velocity range. As the gap between the plates is increased, they become rather less pronounced and the peak-to-trough amplitude decreases. The peaks occur at the same frequencies, but not the same velocities, at all values of \( G \), as can be seen from figure 2 where \( 0 \) is plotted against the vortex frequency (which through \( St_V \) is proportional to flow velocity at any particular value of \( G \)).

The cut-off frequencies of higher-order acoustic modes in the circular pipe (internal diameter 72.54 mm) which supplies air to the test rig (calculated from [4]) are also indicated in figure 2. In most cases, but not all, the frequencies of the \( 0 \) peaks can be identified with these frequencies. This indicates that radiation at the vortex frequency is enhanced by resonance, as a result of acoustic feedback, when the vortex frequency coincides with a higher-order-mode cut-off frequency in the air-supply pipe; and measurements of the spectrum of wall-pressure fluctuations in the pipe confirm that resonance does occur under these conditions. Behaviour of this kind, involving acoustic resonance in the supply system, has been previously observed - see, for example, Rockwell [5].

With emphasis on the results for the larger gaps where the effects of resonance appear quite mild and do not appear to obscure the general form of the dependence of \( 0 \) on flow velocity, the index for a relation of the form \( 0 = U^{(n-1)} \) (and correspondingly between radiated acoustic power \( W \) and flow velocity of the form \( W \propto U^n \) has been determined as \( n = 5.5 \). As can be seen from figure 1, this gives a good representation of the general trend of the selected data presented. The present results are therefore generally consistent with previous conclusions that (in the absence of feedback coupling) the dominant component of acoustic radiation from a two-plate array is characteristic of a dipole (or modified dipole) source (identified as force fluctuations on the downstream plate produced by the interaction between this plate and the vortex motion in the gap between the plates).
Figure 1. Effect of $G$ on the variation of spectral density of acoustic power radiation with flow velocity.

Figure 2. Identification of $\phi_n$ peaks with higher-order acoustic modes in the air-supply pipe.
Direct determinations of the flow regime between the plates in the present experimental rig have not yet been made; but the boundaries between trapped-vortex and vortex-street flow determined by low-speed wind-tunnel tests [2] suggest that for flow velocity variation at fixed \( G \), within the Reynolds-number range of the present tests, transition from one regime to the other might be expected for \( G < 2.5 \), or perhaps trapped-vortex flow only for \( G \) values near two. Vortex-street flow over the whole range is indicated for all \( G > 2.5 \).

The experimental results are somewhat at variance with these indications. For \( G = 2 \) (figure 1), variation of \( \Phi_a \) essentially as \( U^{5.5} \) for \( U > 100 \text{ m/s} \) is interpreted to mean that in this velocity range the flow regime is the same throughout; however, the extremely rapid decrease in radiated acoustic power with decreasing flow velocity for \( U < 100 \text{ m/s} \) appears to correspond to transition between regimes. The Reynolds number here at the onset of transition is \( Re_t = 26 \times 10^3 \) compared with a value well in excess of \( 30 \times 10^3 \) indicated by the data in [2]. For \( G = 2.2 \) and all higher values, no such sudden decrease is observed within the present velocity range. While these results indicate that the Reynolds numbers at the transition boundary for the present experimental conditions are substantially lower than those in [2], the rapid drop in \( \Phi_a \) is in accord with the sharp drop in surface pressures (and forces) on the downstream plate previously found [2] to accompany transition. More comprehensive investigation of the dependence of radiated acoustic power on flow regime clearly requires an extended velocity range, as well as direct examination of the flow regime within the gap.

CONCLUSIONS

From the present experiments the following conclusions can be drawn.

1. A previous conclusion that the dominant component of acoustic radiation generated by a rectangular-section plate in the vortex wake of a faired upstream plate is dipole (or modified dipole) in character is supported, in that the radiated power is proportional to about the 5.5th power of the velocity of flow over the plates.

2. In the transitional flow regime in the gap between the plates, in which trapped-vortex flow changes to vortex-street flow, extremely rapid variation of acoustic power radiation with flow velocity occurs. This is indicative of dramatic changes in the pattern of flow over the plates as flow velocity changes, rather than the character of the acoustic source in any given flow regime.

ACKNOWLEDGEMENT

The work reported is supported by grants from the Australian Research Council.

REFERENCES


CAN HIGH SPEED FLOW BE SILENCED?

J. E. Ffowcs Williams
University of Cambridge

SUMMARY

Sound is an unsteady process and only unsteady flows make sound. The ultimate noise control objective, to silence the flows, is actually a plan to avoid all unsteady effects. Is that prospect within the realms of reality or merely idle speculation? Useful flows must have a beginning and an end; they must start and stop, so absolute steadiness is a nonsensical idea. Quasi-steady, lacking components in the audio frequency range, is not so inconceivable.

The paper dwells on the reasons why flows designed to be steady, break down into modes of varying complexity eventually becoming chaotic. Their early breakdown is linear, instabilities growing from nothing, and these instabilities might well be subject to the control techniques of anti-sound. Flows artificially maintained to be different from their natural noisy state, is the prospect that emerges from these thoughts.

STEADY PROPULSION

Jet propulsion is noisy and the need to control jet noise has been a constant restraint on the development of high speed aircraft. Today's commercial airliners are very much quieter than earlier types because the speed of their propulsive jets has been lowered, jet speed being the dominant factor setting both noise and efficiency levels. Propulsive thrust depends on the jet flow being faster than the aircraft so there are obvious limits on the flight speed for which jet-speed-restraining noise control strategies are useful. Reasonably quiet supersonic airliners are unlikely without a radically new approach to the problem and that immediately raises the fundamental question of whether better solutions can exist. It would be surprising if the answer to that question were a definite no, notwithstanding the fact that the intense research efforts of the past have only brought about a modest improvement in technology.

Jet propulsion is smooth, steady thrust being the desired product of an aircraft's engine. Yet if the thrust were absolutely steady one can question whether there would be any noise at all. It is a fact that unsteadiness of the exhaust flow is a crucial feature of the modern low specific thrust engines whose 'low' jet speed keeps 'pure' jet noise under control. Unsteady nozzle flows even matter at Concorde conditions, measurements leaving no doubt that much of the noise is excess to that of its high speed jet. Unsteady exhaust is an efficient acoustic source, not unlike the action of a powerful loudspeaker, and the smoothing of internal flows incident on the nozzle is an obvious noise control measure. Some of the internal unsteadiness is the noise of the combustion process and the sound of turbomachinery; that core noise element can be attacked by lining the interior duct surfaces with sound absorbent
material. Some of the unsteadiness is caused by acceleration of inhomogeneous material through the propulsion nozzle, the uneven mixture of burnt and unburnt gases which, being of different density, respond differently to nozzle pressure gradients. The 'chuffing' as pockets of air emerge from the nozzle of a garden hose is a familiar example of how noisy this process can be, but it is easily controlled. Just how smoothly the exhaust of an engine could be made to emerge from the propulsive nozzle is unknown, the engineering effort needed to achieve optimal smoothness having yet to be mounted despite the significant role these nozzle-based sources play in practical jets. Internal fluctuations would undoubtedly have been eliminated were they the only source of noise and noise was a crucial constraint.

But the noise and confusion in an engine's exhaust is not entirely determined by conditions upstream of the nozzle. Some of it is caused by the unsteady pressure field of the turbulent jet varying the nozzle back-pressure and therefore its discharge rate. To avoid unsteady nozzle conditions completely, the jet must either be steadied or the effects of its unsteadiness kept away from the nozzle. In principle actively controlled secondary sources could be made to nullify unsteady nozzle discharge from whatever cause and indeed anti-noise technology is already developed to the state that it provides an impressive demonstrable solution to the noise of static engine installations. Consider then the hypothetical but desirable example of a steady jet emerging from a propulsive nozzle and emerging absolutely smoothly into the outside world. Being completely free of nozzle-based sources such a jet would be abnormally quiet. That prospect should drive a serious attempt to produce smooth nozzle flows just as soon as the jet noise problem is sufficiently critical. But there is another reason for seriously considering that hypothetical flow and that is the possibility that the jet itself might never develop an unsteady structure were its starting conditions controlled to be steady. What kind of a jet could develop from a completely steady nozzle flow?

PURE JET NOISE

High Reynolds number jets are turbulent, their flow being very inhomogeneous and unsteady even when they come into still surroundings through a smooth opening from a large reservoir at high pressure. Their turbulence makes noise and the term 'pure' jet noise refers to that remaining after the removal of contributions attributed to identifiable jet features. Turbulence at the interface of different materials or modified by standing compressive waves makes 'impure' noise. It was on pure jet noise that the understanding of the way turbulence makes sound was most tested but, despite the fact that the subject is extensively documented, it has only limited practical value; the pure high speed jet is far too noisy.  

As the jet mixes with its environment its turbulent eddies evolve and travel downstream at the interface of fast and slow flow. Aeronautical jets are often supersonic so most jet noise sources also travel at speeds comparable with the sonic speed. The Doppler factor matters a lot, jet turbulence radiating higher frequency sound towards the jet axis than it sends upstream. The sound, coming from the highly sheared flow, between hot and cold gas, is amplified refracted and scattered as it emerges from the jet. The size and shape of the propulsion nozzle also has a strong bearing on both the jet and its noise and it is obvious from all the different possible jet arrangements that the process of controlling the jet's acoustic output is bound to be a complicated one. The subject is still effectively empirical, the enormous ad hoc experimental effort of the past being necessarily focused on the most practical range of operational conditions.

Whether it is possible, in principle, to control the jet's noise without controlling also the jet turbulence that causes it all is doubtful. Schemes to shield the
jet from sensitive regions by interposing acoustically opaque gas streams or by exploiting the aircraft's shadowing properties, though definitely helpful, have yet to provide enough shielding to offer a practical solution. Controlling turbulence is similarly a most unlikely prospect, unpredictability being one of turbulence's most distinguishing characteristics. But there are clear signs that among the chaos and confusion we call turbulence that a little well chosen stimulation can change things. Seeding of the jet pipe flow with the right tonal sound can make a big difference to the broadband noise of the 'pure' jet, and it has been discovered that under those conditions there is a great deal of order in the structure of the eddies that were previously so random and chaotic and unmanageable. Sensitive jets soon became an attractive research topic so much is known about them now. It is the basic instability of the flow that takes the jet's kinetic energy to fuel the growth of eddies that degenerate into chaos. Instabilities are essentially unsteady with the jet being most sensitive to distinctive instability modes. When they are stimulated artificially, by seeding the nozzle flow for example, those modes grow most rapidly to alter the previous disordered arrangement; they are observable till they blow up under the chaotic influence of strong non-linear effects. At one speed extreme, the flickering flames so carefully described by Tyndall reveal the same flow-controlling properties of sound as did Poldervaarts' supersonic jets at another extreme. Both those flows were transformed by tiny external changes. The possibility that the right external change might bring about quieter practical jets is not therefore a new idea; what is new is that technological progress in controlling delicate external conditions now justifies our taking a new look at the problem.

ANTI SOUND

Sound shares with instabilities the fact that both are linear over much of their practical range. Superposition schemes are key to their understanding. The superposition of sound on to its phase-inverted replica, the essence of anti-sound, is an essentially linear idea. Instabilities are even more linear, being exponentially small in their formative infant state, and those are similarly subject to active control by superposition of a destructively interfering element. Instabilities grow of course, soon becoming powerful and non-linear - but only as long as they have not been destroyed in their linear state. The control will only be as good as the degree to which the anti-sound interference is perfect, which can never be absolute so continual refinement of the control signal is essential. Time is required to observe the controller's effect and refine its output, a time during which any error grows exponentially, putting severe accuracy demands on the control system, but demands that can be met by today's technology in important applications. Compressor surges that put the mechanical integrity of gas turbines at risk have been avoided by controlled acoustic seeding in this way.

The key to this kind of flow control is to find some sensitivity to external stimulation and to feedback into the flow an intelligent stimulation, through actuators responding in a controlled way to instantaneous flow conditions. When operating, the signal going through the controller, to induce the instability-cancelling flow perturbation, is infinitesimal; so is the flow disturbance. The fact that the natural uncontrolled flow contains highly energetic disturbances is no indication at all of the controller's energy requirements. We have seen then the first essential requirement for controllability in jets; some are definitely responsive to external stimulus. The next step is to find the most appropriate stimulus, with a control scheme that feeds back into the flow a stimulus intelligently arranged to stabilise it.

What these thoughts are leading to is this. The turbulence that exists in jets because the basic flow is unstable might not be there if we could devise an active control device to stabilise the basic flow. That flow would be silent. Of course the
matter is absurdly oversimplified by this single degree of freedom analogy and it is very much with tongue in cheek that the control prospect is raised at all. But since noise concerns only the large scale unsteady behaviour, it is at least worth a though as to whether the acoustically significant features of the jet, the unsteady jet development, is more susceptible to feedback control than it is to manipulation of the steady flow parameters.

UNSTABLE SHEAR LAYERS

Thin shear layers are linearly unstable to a wide variety of perturbations and even the simplest two dimensional problems involve distributed parameter controls and ideas more general than needed so far in anti sound. But it is worth working the simplest models just to get a feel for the leverage one can expect to get by managing the development of any controllable instability modes. This subject area is wide open and new and before embarking on an experimental programme it would be good to have definitive statements on analytically tractable theoretical models to build up one's intuition. The long wave response of the linear semi-infinite two dimensional vortex sheet might well be the simplest starting point, and given that flows inability to support upstream travelling modes, controllability of 'nozzle' or starting conditions would probably imply controllability everywhere. That is the kind of conjecture that might be proved to give some confidence that controllability can exist. Similarly, an experimental investigation could start on a shear flow already known to be susceptible to unsteady boundary changes, the vortex layer that excites a whistle or an organ pipe for example. Both those vortex layers are known to respond to the acoustic action of the resonator and they are also known to be controllable by antisound.

CONCLUSIONS

The underlying theme of this paper has been to suggest that steady and therefore silent high speed flows exist in principle but are essentially unstable and unnatural. Feedback control changes stability and it might be possible to devise actively controlled stable flow systems which would also be silent and their controllers energetically neutral. Even if absolute control is impossible, it might be easier to reduce the noise of jets whose unsteady characteristics are different because of the action of active controllers modifying the jet's stability. Because the high speed jet noise problem is important, the case is put for turning attention to active control schemes and seeking in them a way to produce propulsive jets of quite different dynamic characteristics, which might be acceptably quiet.

REFERENCES

MEASUREMENT OF SOURCE DATA FOR FLUID MACHINES - INCLUDING HIGHER ORDER MODE EFFECTS

Jüri Lavrentjev, Mats Åbom
Technical Acoustics, Dept. of Vehicle Engineering, Royal Institute of Technology, S-100 44 Stockholm, Sweden

SUMMARY

Sound generation from fluid machines such as pumps, fans, etc., is an important problem - because it can create high noise levels in connected duct systems which can radiate via walls and openings. For the plane wave region a number of measurement methods to determine the acoustic character of in-duct sources (source data), have been published. However, many fluid machines generate their dominating acoustic power above this region and measurement methods to determine source data - including higher order mode effects - are therefore needed. The aim of this paper is to suggest a method for this type of measurements.

THEORY

We will here for convenience discuss a machine with only one acoustically active opening which is connected to a duct. This is actually no limitation and the results we present can be directly extended to a machine with any number of openings. We now assume that the active opening on the machine can be regarded as a linear and time-invariant acoustic system. In the plane wave region the opening can then be completely described as an acoustic 1-port source; characterised by a source strength and a source reflection coefficient (the source data). If \( N \) modes can propagate in the duct coupled to the active opening the acoustic state can be completely described by using \( N \) state variables (an \( N \)-port source). If we use the acoustic pressure amplitudes as these state variables, then in the frequency domain the source (see Figure 1) can be described by the equation

\[
p_{1s} = R^s p_{1s} + p_s^s, \tag{1}
\]

where \( p_{1s} \) and \( p_{1s} \) are \([N \times 1]\) vectors which contain left- and right-going pressure wave amplitudes for \( N \) modes, \( R^s \) is the \([N \times N]\) source reflection matrix and \( p_s^s \) is the \([N \times 1]\) source strength vector. The elements of the source reflection matrix \( R^s \) represents the reflection coefficients of \( N \) modes seen from the reference cross-section towards the source-side. The elements of the source strength vector \( p_s^s \), can be interpreted as the pressure of \( N \) modes generated by the source-side when the system is terminated reflection-free, i.e., when \( p_{1s} = 0 \).

The easiest way to find the unknowns \( R^s \) and \( p_s^s \), is to use a two-step procedure [1]. First, we use an external source (e.g., a loudspeaker) which is uncorrelated to the source and is located
to the left of the reference cross-section. Using this external source we can eliminate $p^e_s$ via correlation techniques from equation (1). By varying the position of the external source a number

\begin{figure}
\centering
\includegraphics[width=0.8\textwidth]{figure1.png}
\caption{Experimental characterization of an one-port flow noise source}
\end{figure}

($\geq N$) of different incident fields ($p_\alpha$) can be generated and the source reflection matrix $R^s$ can be determined from measurements of $p_1$, and $p_{10}$. Secondly, when $R^s$ is known we can by removing the external source and applying a known load determine the source strength $p^e_s$ via equation (1).

In order to determine the source reflection matrix, we have to obtain first the state vectors $p_+, p_-$. For this we perform a spatial sampling of the sound field by measuring the field at $2N$ independent spatial positions. These positions can be divided equally between the two cross-sections 1 and 2. We can formulate the result of such measurements as [2]

\begin{align}
    p_+ &= M \, p_{1+} + M \, p_{1-}, \\
    p_- &= M \, p_{2+} + M \, p_{2-},
\end{align}  \tag{2a} \tag{2b}

where $p_+$ and $p_-$ are $[N \times 1]$ vectors containing the measured acoustic pressures at $N$ measurement points at the cross-sections 1 and 2, respectively, and $M$ is a modal $[N \times N]$ matrix containing eigenfunctions for the $N$ modes. The $n$th element of the modal matrix $M$ is given by $(M)_{mn} = \psi_n(r_m)$, where $r_m$ is the 2-D position vector (over the duct cross-section) of measurement point $m$ and $\psi$ is a duct eigenfunction. Now, using $p_{2+} = T_+ p_+$, and $p_2- = T_- p_-$ and transfer functions instead of directly measured pressures in equation 2, we can rewrite it as

\begin{equation}
    \begin{bmatrix}
        H_{11} \\
        H_{21}
    \end{bmatrix}
    =
    \begin{bmatrix}
        M & M \\
        M \, T_- & M \, T_+
    \end{bmatrix}
    \begin{bmatrix}
        H_{1+} \\
        H_{1-}
    \end{bmatrix}, \tag{3}
\end{equation}

or $H = M_H H_z$, where the diagonal matrices $T_z$ contain exponential propagation factors and transform the state vectors from section 1 to 2, $H=p/e$ and $e$ is the electrical signal driving the external loudspeaker. Different acoustic states ($N' \geq N$) can easily be created by positioning the external source at different positions along the axis and at different angles around the duct perimeter. Then we get from equation (3)

\begin{equation}
    [H_1 H_2 \ldots H_{N'}] = M_e \left[(H_1)_{1}(H_2)_{2}(H_3)_{3} \ldots (H_{N'})_{N'}\right], \tag{4}
\end{equation}

or $H^E = M_e H^E_z$. From equation (4) $H^E_z$ can be solved giving the two submatrices $H^E_z$ and $H^E_z$,
from which the source reflection matrix can be calculated using a matrix pseudo-inverse

$$ R^s = H^e (H^e)^{-1}. $$

(5)

Now we can determine the source strength vector from equation (1). In order to rewrite this equation in measurable quantities we use the load reflection matrix $R^l$, which can be determined in a separate measurement by using a similar procedure as used for the source reflection matrix (but with the external source to the right). If we now apply equation (1) to cross-section 1, use equation (2a) and also take into account that $\mathbf{p}_{1-} = R^l \mathbf{p}_{1+}$ we can obtain

$$ \mathbf{p}_{1+}^s = (E - R^s R^l)^{-1} (M_e + M_e R^l)^{-1} \mathbf{p}_{1} = C_{1} \mathbf{p}_{1}. $$

(6)

where $E$ is the unit matrix. To get a formulation valid for both random and periodic types of signals we introduce the source power-spectrum matrix [2] $G^s = \mathbf{p}_{1+}^s (\mathbf{p}_{1+}^s)^*$. In order to suppress the flow noise we could apply equation (6) at two different cross-sections (1 and 3); and then transfer the result measured at 3 back to section 1 using: $\mathbf{p}_{1+}^s = T^{-1}_3 C_3 \mathbf{p}_{3}$. The source strength matrix can now be calculated as

$$ G^s = \mathbf{p}_{1+}^s (\mathbf{p}_{1+}^s)^* = (T^{-1}_3 C_3 \mathbf{p}_3) (C_1 \mathbf{p}_1)^* = T^{-1}_3 C_3 (\mathbf{p}_3 \mathbf{p}_1)^* C_1 = T^{-1}_3 C_3 G^s_3 C_1, $$

(7)

where the superscript $c$ denotes transpose and complex conjugate. When the distance between cross-sections 1 and 3 is chosen large enough, the flow noise disturbances at these two cross-sections will be uncorrelated which will result in an efficient flow noise suppression.

RESULTS AND CONCLUSIONS

To verify and test the method described in this paper measurements were first done using a loudspeaker source and with no flow [2]. The method was then applied to study an axial fan (6 blades, 1430 rpm) mounted at the ended of a duct ($D=0.4$ m), see Figure 1. The axial separation between section 1 and 2 was $s=0.126$ m, which allowed good measurements results in the frequency range from 100 Hz up to the cut-on of the first radial mode 1100 Hz [2]. The cut-on frequencies for the first two circumferential modes were $f_{10}=520$ Hz and $f_{20}=863$ Hz. Since these two circumferential modes are doubly degenerate; $N=1+2+2=5$ for our set-up. The spatial sampling points were chosen on the duct wall and were evenly distributed along the perimeter, see Figure 1. The measured source data was transferred to the end of the fan section (point c in Figure 1).

In Figure 2 some of the results are presented. Figure 2a shows a comparison between directly measured auto-spectra at section 1 and predicted using the measured source data. For this test a new acoustic load different from the load used in the measurement of the source data was applied. In Figure 2b the source strength components for the different modes are presented. All modes of the same order are summarized and the resulting curves are presented. We can see that for frequencies above the first higher mode cut-on, the higher order modes dominate over the plane wave. In Figure 2c and 2d the source reflection coefficients for different modes are presented.

In conclusion - the suggested measurement procedure has been verified via a simple test case and also successfully applied to an axial fan.
Figure 2. Measurement results for the test case in Fig.1: (a) auto-spectrum at mic.3, —— measured, —— predicted; (b) source strength (f_0= blade-passing frequency=143 Hz); (c), (d) diagonal elements of the source reflection matrix, (c) magnitude, (d) phase. (— —) plane wave, (— —) first higher order mode, (— —) second higher order mode.

REFERENCES

DIRECTIVITY OF JET NOISE

Ricardo E. Musafir
Acoust. & Vib. Lab./Dept. Mech. Eng./COPPE and Hydraulics Dept./EE,
Universidade Federal do Rio de Janeiro, Brazil.

SUMMARY

Expressions for the high and low frequency limits of the far field directivity of a statistically axisymmetric point quadrupole embedded in a cylindrical plug flow are derived. These expressions display the main effects of flow in the generation and propagation of sound, including aspects that are frequently neglected, providing a model that may be useful to explain observed features of the sound field of round jets. A brief discussion of these features regarding the field of excited jets is presented.

INTRODUCTION

The prediction of jet noise directivity received a major attention mainly up to the end of the 70's, when a large number of papers focused on the sound field of the various quadrupole components when embedded in a flow of limited transversal extent, with or without shear [1-5]. It is thus surprising that comparatively little information concerning the overall directivity of a point quadrupole with suitable statistical properties, considering the effect of flow, is available. As a consequence, with the renewed interest in jet noise, recent experiments or numerical sound field simulations and up being analysed by models in which flow effects are not considered (like Ribner's pioneer shear noise/self noise model [6]) or which are inadequate to the situation under discussion.

Although Mani [1] emphasized that a main effect of the presence of flow around the source was that each quadrupole component would be differently affected by it, destroying the simple and attractive picture given by Lighthill's analogy, in which the effect of flow (restricted to source convection) appears as an overall multiplicative factor, the finding of Tester and Morfey [2] that, in the high frequency limit, the effect of flow in the field of a statistically isotropic point quadrupole can also be described by a similar factor, has probably contributed to the neglect of flow effects other than refraction.

The present paper aims to discuss the basis of a model that displays the main effects of flow, providing closed form solutions for a statistically axisymmetric point quadrupole located in the centerline of a cylindrical plug flow. The use of the simple plug flow model (first employed by Mani [1] and further discussed by Dowling et al [3]) instead of the full Lilley's equation was justified in [7]. The use of a single point source to infer jet noise properties has been employed by a number of authors [1,2,6,8]. This procedure is expected to work as long as the jet can be seen as an ensemble of compact independent emitters. Otherwise, the model can be used to suggest guidelines to the analysis. The expressions obtained, applying to isothermal round jets, are a by-product of the analysis of far field sound pressure correlations [8]. Some of the results were given previously, without derivation, in [9].

MODELLING OF SOURCE CORRELATIONS

The governing equation for the pressure field of a point quadrupole \( T_q(t) \) located at the origin of the coordinate system and on the centerline of a plug flow of velocity \( v = U \hat{e}_1 \) (\( U \) represents the velocity at the source, not the nominal jet exit velocity, \( U_j \)) and mean properties identical to those of the surrounding medium, is written, for points inside the flow, as

\[
\left( c_0^{-2} \frac{D^2}{Dt^2} - \nabla^2 \right) p(x,t) = T_q(1) \cdot \nabla \delta(x) = T_q(1) \frac{\partial^2 \delta(x)}{\partial x_1 \partial x},
\]

where \( D/ Dt = \partial t/ \partial t + U \partial / \partial x_1 \) and \( c_0 \) is the speed of sound.
The far field solution outside the flow can be represented by

\[ 4\pi|k|c_0^2 p(x, t) = F_y \partial^2 T_y(t^* - \tau) / \partial \tau^2 \]  

(2)

where \( t^* \) is the appropriate retarded time and the \( F_y \), which represent the flow transmission coefficients, depend on \( M = U/c_0 \) and on the coordinates of \( x | x | = (\cos \theta, \sin \theta \cos \phi, \sin \theta \sin \phi) \) and also on the spectra of the corresponding \( T_y \), being thus given by inverse Fourier integrals, which reduce to simple expressions, for finite \( M \), in the high and low frequency limits or for \( M = 0 \), when \( F_y = n_n \).

The far field directivity \( D(\theta, M) \) will therefore depend on the fourth derivative of the mean value of \( T_y(t^*) T_y(t^* + \tau) F_y F_y \) with respect to \( \tau \), evaluated at \( \tau = 0 \), that will be noted simply as \( <T_y T_y> F_y F_y \).

The modelling of the source correlations \( <T_y T_y> \), of which few attempts are found [6,8], is greatly simplified by the hypothesis that the source is statistically axysymmetric (SA). This is readily seen if one starts with the source basic directivity \( D_0(\theta) = D(\theta, 0) = <T_y^2> = n_n n_n <T_y T_y> \). 

There are up to six independent \( T_y \) - what reflects the fact that the field of a general point quadrupole in an homogeneous medium at rest can be described by five spherical harmonics (SH) of order two plus a SH of order zero (i.e., an omnidirectional term). There are thus 36 possible different correlations but, since the source is SA, \( D \) must be independent of \( \phi \), so that only 6 distinct nonzero terms occur: four auto-correlations, and two cross-correlations. It is a simple matter to show that, for an SA quadrupole, \( D \) will not depend on \( <T_y T_y> \), since we see from [1] that \( F_y T_y \) will be always proportional to \( \sin 2\phi \) and we can advantageously set \( \phi = 0 \). So, we are left with the necessity of modelling up to five parameters (which can, of course, be frequency dependent) to obtain the mean square field, and four for the normalized \( D(\theta, M) \), which can be written, recalling that \( F_y T_y \) will be also proportional to \( \sin \phi \), as

\[ <T_y^2> D(\theta, M) = F_{11}^2 <T_y^2> + 2F_{11} (F_{15} + F_{33}) <T_y T_{12}> + 4F_{11}^2 <T_y^2 > + 4F_{11} (F_{15} + F_{33}) <T_y T_{12}> + 4F_{11}^2 <T_y^2 > + 4F_{11}^2 <T_y T_{12}> + 4F_{11}^2 <T_y T_{12}> \]  

(3)

In the most general situation, where the \( F_y \) depend on the order \( m \) of the azimuthal modes, \( \cos m \phi \), emitted by each \( T_y \) in the absence of flow, the need for 3 (or 4) parameters may be explained as follows: the mean square pressure field will be obtained by the summation of 3 terms, each corresponding to one azimuthal mode \( (m = 0, 1, 2) \) - what is readily seen from the two-point far field correlation analysis [8]. Modes 1 and 2 are related only to the square of the corresponding SH of order two that is even in \( \phi \), while the axysymmetric term relates to up to 3 parameters, one associated with the square of the corresponding SH of order two, a second with the SH of order zero (omnidirectional) and a third with the coupling of these two spherical harmonics.

An interesting situation occurs when the \( F_y \) can be written as \( g \delta \) for some vector \( g \). Then, we may set \( F_{11} = 0 \) and since \( F_{11} F_{11} \) will be equal to \( (F_{11})^2 \), it follows that \( D \) can be written as

\[ D(\theta, M) = |g|^4 + B F_{11}^2 + 4 A F_{11}^2 \]  

(4)

where \( B = <T_y^2>/ <T_y^2> \) - 1 and \( A = <T_y^2> / <T_y^2>/ <T_y^2> <T_y^2> \) - 1/2, are the two parameters that are needed to describe \( D \). They govern the anisotropy of the source basic directivity which, for an SA source, can always be written as

\[ D_0(\theta) = 1 + B \cos^4 \theta + A \sin^2 2\theta \]  

(5)

When \( T_y = Q(\tau) \delta_y \), i.e., \( T \) stands for an isotropic quadrupole, the resulting directivity \( D = (F_y \delta_y) \) may be rather particular, since it will be identical to that of the volume displacement point monopole corresponding to a right hand side of (1) given by \( c_0^2 D_0 \delta_y (x) / \partial x^2 \), being thus quite different from what would be anticipated from the directivity of a single quadrupole component.

For the modelling of jet noise, a situation somewhat opposed to the above is particularly relevant: When \( \partial^2 \tau (T) / \partial \tau^2 = 0 \), so that there is no SH of order zero in the description of the sound field for \( M = 0 \). Then, only 3 parameters will be needed, each corresponding to one azimuthal mode, their number reducing to 2 for \( D \) as well as for the normalized cross-correlation field. As a consequence, the normalized values of the source cross-correlation terms \( <T_y T_y> \) and \( <T_{22} T_{22}> \) can be modelled from the knowledge of \( B \) in \( D_0(\theta) \) (which is related to a ratio of strengths of longitudinal components and thus, to \( \tau \{ T \} \)). From the restriction on the trace, it is straightforward to obtain that the needed terms, which will be noted as \( r_{ij} = <T_{ii} T_{jj}> / <T_{22}^2> \), i ≠ j and no summation implied, will be given by
\[ r_{12} = -(1 + B)/2, \quad r_{23} = -(1 - B)/2 \] (6)

In this case, where the field for \( M=0 \) is described by spherical harmonics of second order only, the sound generation process may be associated with isovolumetric distortion, compressibility, as long as the source field is concerned, playing no part in it.

The use of this hypothesis for jet noise at moderate \( M \) has been frequently advocated. The facts that compressibility effects appear in (6) as an \( O(M^2) \) correction [8] and that it was successfully employed in both qualitative [8] and quantitative [10] predictions of normalized far field correlation data, give strong support to its use.

The normalized properties of an 'incompressible' SA point quadrupole are seen to depend on \( A \) and \( B \) (and on their frequency counterparts). Equation (6) and the definitions of \( A \) and \( B \) impose \( A \geq -(B+3)/4 \) and \(-1 \leq B \leq 3 \); \( B=1 \) implies \( < T_{32} >= 0 \), while \( B=3 \) corresponds to an axisymmetric sound field \( (r_{23} = 1, < T_{33} ^2 >= 0) \). For jet noise a good initial guess is, since \( D_i \) is roughly uniform, to set \( A=0 \) and expect \( B \) to fall between 0 and 2. The simulations in [8] where conducted with \( B=0 \) (for jet exit Mach number \( M_e=0.4 \)) and \( B=1 \) (\( M_e=0.75 \)).

ASYMPTOTIC EXPRESSIONS

High Frequency Limit. In the HF limit, outside the so called 'cone of relative silence', defined by \( \varepsilon^2 = (1 - M \cos \theta)^2 - \cos^2 \theta \simeq 0 \), the \( F_{ij} \), obtained by Goldstein [5], are given by

\[ F_{ij} = \nu_i \nu_j (1 - M \cos \theta)^{-1}, \quad \nu = (\cos \theta, \varepsilon \cos \phi, \varepsilon \sin \phi). \] (7)

In this case, \( D \) is obtained by inserting the \( F_{ij} \) into (4). Source movement is accounted for by multiplying the result by the traditional quadrupole convection factor, \( C^2 = (1 - M \cos \theta)^{-1} \), where \( M_e \) is the convection Mach number (for narrow-band directivity, the corresponding factor is \( C_e^2 \)). One obtains then, with \( C = (1 - M \cos \theta) \),

\[ D(\theta, M) = C_e^{-2} [C^2 + (B - 4A)C^{-2} \cos^4 \theta + 4A \sin^2 \theta] \] (8)

which, for \( A=B=0 \), can be written as \( C_e^{-2} C^2 \sim C^2 \), since \( M \sim M_e \). It is interesting to note that this \( C^2 \) dependence was first found by Goldstein and Howes [11], who made no restrictions on \( D_0 \), although the hypotheses of statistical isotropy is implicit in their derivation. It was posteriorly obtained by Tester and Morsey [2] for an uniform \( D_0 \) and, although the full equation (8) can be derived from their reasoning, in a subsequent paper [1], they described departures from statistical isotropy as \( D = C^2 (1 + B \cos \theta) \), what is inconsistent with (8). Goldstein [5], on the other hand, associated the \( C^2 \) factor with instantaneous isotropy, an unnecessary assumption.

The existence of terms with the factor \( C_e^{-2} \) was first identified by Balsa [3]. The modifications in the \( F_{ij} \) for off-axis source positioning and for \( \varepsilon^2 < 0 \) are discussed in [5].

Low Frequency Limit. The LF \( F_{ij} \) are given explicitly by Dowling et al [3], and can be written as

\[ F_{ii} = \cos^2 \theta (1 - M \cos \theta)^2; \quad F_{ij} = n_i n_j (1 + (1 - M \cos \theta)^2)^{-1}, \quad i \neq j; \]

\[ F_{ii} = \frac{1}{2} \frac{\varepsilon^2}{(1 - M \cos \theta)^2} + \frac{(-1)^i \sin \theta \cos \phi}{1 + (1 - M \cos \theta)^2} + \frac{1}{2} \frac{\cos^2 \theta}{2(1 - M \cos \theta)^2} + \frac{(-1)^i \sin^2 \theta \cos \phi}{1 + (1 - M \cos \theta)^2}, i = 2 \text{ or } 3. \] (9)

\( D \) now is to be obtained from (3), being dependent on \( r_{12} \) and \( r_{23} \). With the incompressibility hypothesis (6) and considering source convection effects, one obtains

\[ D = C_e^{-2} \left[ \frac{B + 1}{4} (9C^{-4} \cos^4 \theta - 6C^{-2} \cos^2 \theta + 1) + (3 - B + (16A + 5B + 9) \cos^2 \theta \frac{\sin^2 \theta}{(1 + C^2)^2} \right] \] (10)

for which a possible approximation is \( D = C_e^{-2} (1 + \beta C^{-4} \cos^4 \theta) \) where \( \beta \) cannot be taken as zero.

Another interesting situation is obtained if the source is SA (\( A=B=0 \), \( r_{12} = r_{23} \)) but \( r = r_b \) is treated as a free parameter. \( D \) is then given by

- 549 -
\[ D(\theta, M) = C_2^2 \left\{ \frac{1 + r}{2} + (1 - r) \left( - \frac{C_4}{2} \cos^4 \theta - \frac{C_2}{2} \cos^2 \theta + 2 \frac{\sin^4 \theta + \sin^2 2\theta}{(1 + C_2^2)^2} \right) \right\} \] (11)

This expression justifies the success of Mani's modelling [1], who supposed a statistically isotropic source with \( r = 0 \) (an approximation to Ribner's \( r = 1/8 \)) instead of the 'incompressible' value \( r = 0.5 \). It is evident that as long as \( r \) is not too close to unity, so that the source cannot be treated as instantaneously isotropic, the term \( C_4 \cos^4 \theta \) will dominate for low \( \theta \).

Equation (10), or approximations derived from it, may be relevant in many situations. Although the mixing noise of a real jet engine is mostly 'high frequency', the same is not true for most model jets, for which, due to the reduced diameter, a significant part of that field falls in LF limit. For very small jets (cleaning jets, for instance) increasing jet exit velocity will not alter this, since it will simply shift the spectrum to the ultrasonic range, the audible sound remaining 'low frequency'.

A situation where this equation may be particularly useful is that of excited jets. In most experiments the excitation frequency corresponds to a large wavelength when compared with jet diameter, so that its subharmonics (associated with formation, pairing and development of coherent structures) are definitely LF. The directivity of the narrow-band peak(s) evidenced by excitation shows, in general, a marked dip, which can be explained with the use of (10): Since a vortex ring can be modelled by a diagonal point quadrupole with one principal axis aligned with the direction of movement, the corresponding \( T_{12} \) and \( T_{13} \) components will be always zero, so that \( A \) assumes its minimum value, \( - B(2)/3 \), which causes the dip. The current expectation, based on the axisymmetric case \( B = 3 \) with \( M = 0 \), is that it would fall around \( \theta = 55^\circ \), although it generally happens for larger values of \( \theta \). We see from (10) that the location of the dip depends on both \( B \) and \( M \). In the experiments analyzed in [10] \((M = 0.64)\), good far field correlation data agreement was obtained for \( B = 0.7 \); the location of the dip was well predicted but not the directivity, which was much stronger than that obtained from (10) with \( M = 0 \) (the coherent structures usually appear as non-moving sources). This suggests that the sources are axially non-compact what, in turn, indicates that, in that experiment, the main sound generation mechanism is not vortex pairing, but the downstream development. It is interesting to note that in the experiments reported in [13] \((M = 0.21)\), the dip is seen at \( \theta = 75^\circ \), what cannot be explained by the present model. This may be due to the fact that, when two or more pairing positions occur, interference between the field of the vortices produces effects not explained by this simple point source model.

CONCLUSIONS

The expressions presented explicitate the effects of flow embedding in the directivity of a statistically axisymmetric, but otherwise general, point quadrupole in the high and low frequency limits. For the high frequencies, the appropriate restrictions on source directivity that permits the expression of these effects by an overall multiplicative factor were discussed. For the low frequencies, the increase with flow Mach number in the difference in the flow transmission factors applying to each quadrupole component is stronger, privileging the axisymmetric mode of emission. Consideration of this effect is essential for understanding the sound field of excited jets. Extension of the present point source model to include strong non-compactness effects may provide a valuable tool to this end.

REFERENCES

MODÉLISATION DE SILENCEUX PASSIFS EN PRÉSENCE D’ÉCOULEMENT

O. Planade, J. Ducourneau, J. P. Guilhot
Laboratoire d’Acoustique de Métrologie et d’Instrumentation, Toulouse, France

RÉSUMÉ

Cette étude s’intéresse à la prévision de l’efficacité acoustique de silencieux passifs. Ce type de silencieux est actuellement utilisé dans les systèmes de conditionnement d’air pour avion. Les faibles dimensions de tels systèmes s’accompagnent généralement d’un accroissement de performances aéronautiques des ventilateurs et par conséquent, du niveau sonore induit.

La modélisation de ces silencieux implique la prévision de l’impédance acoustique en incidence normale du traitement acoustique, constitué généralement par des matériaux recouverts d’une tôle perforée. Dans le cas de matériaux poreux plusieurs modèles peuvent être envisagés. Pour des matériaux à pores fermés, l’étude s’appuie sur une approche mécanique basée sur la propagation dans les solides élastiques.

Le modèle de calcul de l’atténuation d’un silencieux de section circulaire ou annulaire est développé selon la théorie modale. Le point important de ce calcul réside dans la résolution des conditions aux limites, insolubles analytiquement.

L’étude expérimentale est réalisée sur un silencieux prototype, et l’atténuation est comparée à la mesure de perte par insertion.

INTRODUCTION

La modélisation du comportement acoustique d’un silencieux passif repose sur la connaissance de ses données géométriques, de la nature de l’écoulement, et de l’impédance acoustique de la structure absorbante utilisée.

Dans un premier temps, nous rappelons les différents modèles de prévision de l’impédance de surface concernant les tôles perforées et la propagation acoustique dans les matériaux poreux. Ensuite, nous étudions le comportement acoustique de ces mousses considérées comme des solides élastiques.

Dans une seconde partie, la méthode paramétrique de calcul de l’atténuation de silencieux passifs est présentée avec quelques rappels sur la propagation d’ondes acoustiques dans des silencieux de section annulaire et circulaire en présence d’écoulement.

MODELISATION DE L’IMPÉDANCE NORMALE DE SURFACE

Milieu poreux :

Toutes les configurations utilisées dans des silencieux passifs sont généralement représentées par des tôles perforées (épaisseur h) placées à une distance e de la coque. La différence entre les deux structures réside dans la nature de l’espace séparant cette partie de la coque. Il est possible de diviser cet espace autour de la tôle en cylindres indépendants de section carré : tableau de résonateurs. Pour des structures continues, l’espace peut être saturé par une couche d’air ou par un matériau poreux. Dans l’aéronautique, les fabricants utilisent des mousses rigides possédant des pores fermés pouvant supporter des températures élevées.
Nous utilisons la théorie d'Ingard [1] pour prévoir l'impédance caractéristique \( Z_c \) et le nombre d'onde \( k_c \) dans le matériau. Les paramètres caractéristiques du matériau sont : la porosité \( P \), la résistance à l'écoulement \( r \), la tortuosité \( m \), ...

Des équations de continuité et de la dynamique, nous déduisons l'impédance normale de surface donnée par

\[
Z = -j \, Z_c \, \cot(k_c e) \quad \text{où} \quad k_c = \frac{\sqrt{Pq}}{c} \sqrt{\left[m - j \frac{r}{\rho \omega} \right]}
\]

Le modèle de Delaney-Basley [4] peut être également utilisé; il donne l'impédance caractéristique et le nombre d'onde du matériaux poreux en utilisant des expressions complexes.

Structure continue : matériau et tôle perforée

Nous pouvons diviser l'espace autour de la tôle perforée en cellules élémentaires isolées, sans déformation du champ acoustique. Ces cellules ont une section carré de coté \( D \), et chacune d'elles débouchent sur une ouverture circulaire de la tôle perforée.

À l'aide du nombre d'onde du matériau \( k_c \) et de son impédance caractéristique \( Z_c \), nous pouvons écrire la pression acoustique moyenne sur l'ouverture d'une cellule :

\[
Z(B) = -j \sigma \cot(k_c e) + jZ_c k_c \sum_{(m,n) \neq (0,0)} \frac{2 \nu_{m,n} R^2 \left[ 2 \pi \frac{R}{D} \sqrt{m^2 + n^2} \right]}{\pi^2 (m^2 + n^2)^{3/2}} = -j \sigma \cot(k_c e) + jZ_c k_c \delta
\]

L'impédance acoustique normale de la structure est alors :

\[
Z(A) = \frac{Z(B)}{\sigma} + j \left( \frac{1}{\sigma} (d + \delta) \rho \omega \right) = \frac{r \delta}{\sigma} + j \left[ k_c (d + (1 + m) \delta) - \sigma Z_c \cot(k_c e) \right]
\]

avec \( \delta \): longueur corrigée et \( \sigma \): taux de perforation

Mousses rigides et matériaux élastiques :

Le comportement acoustique de ces mousses repose sur la théorie de Biot [7]. Nous considérons que seule la phase solide existe compte tenu des pores fermés, et que seule la vitesse de cette phase \( c \) est prise en compte dans le calcul. L'élasticité de la phase solide existe à cause de la porosité occlusée dans les différents bulbes. Les coefficients du tenseur de contrainte sont uniquement dépendants des coefficients de Lamé \( \lambda \) et \( \mu \). L'expression de l'onde de cisaillement et de l'onde de compression se propageant dans la structure permettent d'écrire l'impédance acoustique normale :

\[
Z = -j \sqrt{\rho (\lambda + 2\mu)} \cot \left( \omega e \sqrt{\frac{\rho}{(\lambda + 2\mu)}} \right) \quad \text{où} \quad (\lambda + 2\mu) = f(E)
\]

Le module d'élasticité complexe \( E \) peut être calculé à partir de la mesure du second coefficient de Lamé \( \mu \), du module d'Young et du coefficient de Poisson. Nous avons préféré exciter l'échantillon en compression avec des ondes planes de façon à récupérer l'impédance normale pour différentes fréquences. Les différentes valeurs obtenues avec cette méthode ne sont pas comparables avec la méthode classique. Ces valeurs, incluses dans le modèle théorique, nous montrent que la prévision de l'impédance normale est validée par l'expérience.

Résultats :

Nous comparons les résultats obtenus avec l'expérience réalisée au Tube de Kundt : pour une mousse rigide et pour différentes combinaisons entre les deux matériaux avec la tôle perforée.
ATTÉNUTION THÉORIQUE DE SILENCIEUX PASSIFS

Les modèles empiriques concernant les silencieux industriels s'appliquent efficacement pour des conduits de grandes dimensions, et pour des fréquences relativement basses. Les fréquences et dimensions rencontrées ici nous orientent vers une approche modale du problème. L'effet de la dissipation acoustique de la structure absorbante apparaît dans les conditions aux limites, que l'on résout par une méthode paramétrique. Les silencieux passifs considérés ici, se divisent en deux catégories : silencieux de section circulaire et annulaire.

Propagation des ondes acoustiques dans un silencieux :

L'équation de propagation d'une onde acoustique dans un conduit en présence d'écoulement uniforme s'écrit :\[ \Delta p - \frac{1}{c^2} \left( \frac{\partial}{\partial t} + M \frac{\partial}{\partial z} \right) p = 0 \]

La solution générale est \[ p(r, \theta, z, t) = \sum_{m,n} \left[ A_{m,n} E_m(K_{m,n} r) \cos(m\theta) \exp(j(ut - k_z z)) \right] \]

où \[ E_m(K_{m,n} r) = J_m(K_{m,n} r) + C_{m,n} Y_m(K_{m,n} r) \]

Le nombre d'onde radial \( K_{m,n} \) et le nombre d'onde axial \( k_z \) satisfont l'équation de dispersion :\[ k_z^2 + K_{m,n}^2 = (k - M k_z)^2 \]

Le nombre d'onde axial étant complexe, sa partie imaginaire induit un terme d'atténuation pour le mode \( (m,n) \) considéré. La détermination du nombre d'onde radial implique la résolution de la condition aux limites sur la surface interne du silencieux. Cette condition exprime la continuité de la composante normale de la vitesse acoustique et de la pression acoustique, ce qui s'écrit :\[ x \, E_m'(x) = \eta \left( 1 - M^2 \frac{k_z}{k} \right) E_m(x) \quad \text{avec} \quad x = K_{m,n} b \quad \text{et} \quad \eta = j k b \frac{\rho_0 c}{S \, Z} \]

où \( b, S, \rho_0, c, \) et \( Z \) représentent respectivement le rayon interne du silencieux, sa section, la densité de l'air, la célérité du son dans l'air, et l'impédance acoustique de la structure absorbante.

Les solutions analytiques de cette équation n'existent pas dans le cas général. Dans cet objectif, nous utilisons une méthode paramétrique, développée par Eversman [9]. Le principe consiste à partir des solutions connues de cette équation, dans le cas de conduits cylindriques dont la paroi interne est parfaitement rigide. Par intégration numérique de l'équation différentielle suivante, équivalente à la condition aux limites précédente, nous obtenons les solutions cherchées.

\[ \frac{dy}{d\alpha} = -2 \eta y \left( \frac{y}{(\eta \, \alpha)^2 - m^2} \right) \quad \text{où} \quad y = (K_{m,n} b)^2 \]

Le calcul de l'atténuation du silencieux a été développé sur un logiciel de calcul scientifique. Le programme ainsi réalisé nous permet d'identifier les modes propagatifs, d'évaluer l'atténuation de chaque mode, et alors de déduire l'atténuation globale du silencieux.

Dans le cas de silencieux de section annulaire, il est nécessaire de résoudre une deuxième équation différentielle relative à la condition aux limites pour le traitement acoustique placé sur l'axe de symétrie du silencieux.

Essais expérimentaux :

Afin de valider les résultats théoriques, nous avons réalisé un banc d'essai pour la mesure de l'atténuation de silencieux. La détermination de la perte par insertion du silencieux testé est basée sur la méthode de substitution, conformément à la norme ISO 7235. Dans une première étape, on note la pression acoustique moyenne \(<L_p>\) en utilisant le silencieux. Dans la seconde étape, la même mesure est réalisée en remplaçant le silencieux par un conduit rigide de même
dimensions. Le niveau de pression acoustique est alors $<L_{pH}>$. 
La perte par insertion est $D = <L_{pH} > - <L_{pI}>$
Les configurations de mesure sont définies sur les points suivants :
- caractéristiques de la structure absorbante considérée : porosité, résistance au passage de l'air, facteur de structure, ...
- géométrie interne du silencieux : longueur, diamètre, section circulaire ou annulaire
- vitesse du fluide
Les pertes par insertion sont comparées aux atténuations théoriques données par le modèle de calcul. La figure suivante représente un exemple de résultat obtenu pour un silencieux circulaire en absence d'écoulement.

**CONCLUSION**

Premièrement, notre étude concernant les matériaux poreux s'accorde avec les modèles déjà existant (Delaney-Basley, JF Allard et d'autres...), mais la prévision selon le modèle d'Ingard est la plus simple.

L'impédance acoustique pour des mousse à pores occluse est encore mal maitrisée. Pour le moment, l'étude du comportement acoustique de ces matériaux par une approche mécanique des solides élastiques montre que les modèles sont en accord avec l'expérience.

La vérification expérimentale nécessite de faire d'autres configurations. Nos perspectives sont d'anticiper une configuration bien définie pour, par exemple, satisfaire la valeur de l'impédance recherchée pour une atténuation optimale du silencieux.

En ce qui concerne la prévision de l'atténuation de silencieux, les résultats théoriques sont validés par les essais expérimentaux dans une gamme de faible vitesse. Toutefois, les résultats diffèrent lorsque celle-ci augmente. L'influence de l'écoulement sur le comportement de la structure absorbante n'est pas prise en compte dans le modèle. D'un point de vue expérimental, la nature même de l'écoulement est également mal contrôlée.

**RÉFÉRENCES**

ON THE POSSIBILITY OF APPLYING A HARD-WALLED CIRCULAR DUCT AS A REFERENCE SOUND SOURCE—THEORY AND MEASUREMENT

A. Snakowska, H. Idczak*
Institute of Physics, Pedagogical University, 34-310 Rzeszów, rejtana 16a, Poland

INTRODUCTION

From analysis of theoretical and experimental data of the acoustic field radiated from a hard-walled circular duct, a possibility of applying such a duct as a reference sound source has arisen. The aim of this paper is to present considerations leading to it.

The theory of propagation of acoustic waves in a semi-infinite cylindrical duct predicts that only some wave modes can propagate without damping [1]. The number of admissible modes depends on the nondimensionalized wavenumber, $ka$, being a product of the wavenumber $k$ and the duct radius $a$.

When the incident wave consists of more than one Bessel mode, the analysis of the field complicates, because quantities by means of which we usually describe the field (the pressure and intensity directivity functions, the power output, etc.) depend on moduli and phases of the individual modes, which could not be predicted from the theory or measured.

To deal with these difficulties we proposed a model of the phenomena based on two assumptions [2]

— the total energy is shared in equal parts between all excited modes,
— phases are independent random variable of a uniform distribution in $[0, 2\pi]$.

The carried out experiments, reported in [2,3] have verified the presented model and let us suppose that the power radiated might be evaluated from the intensity (or pressure) measured in only one field point.

THE FAR FIELD OF A CYLINDRICAL DUCT FOR THE RANDOM PHASE

For an axisymmetric excitation with frequency $\omega$, the velocity potential in the far field outside the duct has, in the spherical coordinates $R, \theta, \varphi$ the form

$$\Phi(R, t) = \sum_{i=0}^{N} A_i d_i(\theta) \frac{\exp \left(ikR - \omega t\right)}{R}, \quad (1)$$

where $A_i$ is the complex amplitude, $d_i(\theta)$ is the directivity function and $N$ labels the highest mode admissible in the duct. Detailed investigations leading to final formulae of $d_i(\theta)$ are enclosed in [4]. Note that each radial mode in the duct contributes to the spherical wave outside the duct with a coefficient $d_i(\theta)$.

For the potential given by (1) the sound pressure can be expressed as

$$p(R, t) = \sum_{i=0}^{N} P_i(R) \exp \left(\phi_i - \omega t\right), \quad (2)$$

- 555 -
where \( P_l(R, \theta) = \omega \rho_0 R^{-1} |A_l| d_l(\theta) \) is the real amplitude of the pressure, while \( \phi_l \) denotes the phase. For the purpose of this paper we are interested in the root mean square pressure \( p_{rms}^2 = \langle Re^2 p(R, t) \rangle_t \), where \( \langle \cdot \rangle_t \) denotes time-average.

If we consider the phases \( \phi_l \) to be mutually independent random variables with the uniform distribution in \((0, 2\pi)\), then the expected value \( E(\cdot) \), the variance \( \text{var}(\cdot) \) and the normalized standard deviation \( \varepsilon(\cdot) \) are equal to [5]

\[
E(p_{rms}^2) = \frac{1}{2} E\left(\sum_{l=0}^{N} H_l \cos \phi_l \right)^2 + \left(\sum_{l=0}^{N} H_l \sin \phi_l \right)^2 = \frac{1}{2} \sum_{l=0}^{N} P_l^2,
\]

\[
\text{var}(p_{rms}^2) = E(p_{rms}^4) - E^2(p_{rms}^2) = \frac{1}{2} \sum_{l=0}^{N} \sum_{m=0}^{N} (P_l P_m)^2,
\]

\[
\varepsilon(p_{rms}^2) = \sqrt{\text{var}(p_{rms}^2)} / E(p_{rms}^2) = \left[ \frac{1}{2} \sum_{l=0}^{N} \sum_{m=0}^{N} (P_l P_m)^2 \right]^{1/2}.
\]

The second assumption, on equal energy distribution between the modes excited in the duct, results in calculating the relative modulus of amplitudes \( |A_l|^2 : |A_m|^2 = \gamma_l : \gamma_m \) [2].

The total power radiated from the outlet can be calculated by means of the intensity \( I \) or the square of \( p_{rms}^2 \). For axisymmetrical excitation, according to (1) and (3) we obtain

\[
P = 2\pi R^2 \int_{0}^{\pi} I(R, \theta) \sin \theta d\theta = \frac{2\pi R^2}{\varrho c} \int_{0}^{\pi} p_{rms}^2(R, \theta) \sin \theta d\theta,
\]

thus, the expected value and the variance are expressed as follows

\[
E(P) = \frac{2\pi R^2}{\varrho c} \int_{0}^{\pi} E(p_{rms}^2) \sin \theta d\theta = \frac{2\pi R^2}{\varrho c} \sum_{l=0}^{N} \int_{0}^{\pi} P_l^2 \sin \theta d\theta,
\]

\[
\text{var}(P) = \left( \frac{2\pi R^2}{\varrho c} \right)^2 \int_{0}^{\pi} \text{var}(p_{rms}^2) \sin \theta d\theta = \left( \frac{2\pi R^2}{\varrho c} \right)^2 \sum_{l=0}^{N} \sum_{m=0}^{N} \int_{0}^{\pi} (P_l P_m)^2 \sin \theta d\theta,
\]

and so, the normalized standard deviation can be calculated as

\[
\varepsilon_P = \sqrt{\text{var}(P)} / E(P).
\]

THE METHOD UNCERTAINTY AND THE EXPERIMENTAL ERROR

Next step is to analyze relations between real and measured values of the power output in the light of some uncertainty introduced by the theoretical model (coming from the assumption on the random phase) and measuring errors. Brief repetition presented below is based on results included in [3].

The real value of power can be estimated by an interval

\[
P = E(P)(1 \pm \varepsilon_P).
\]
The expected value of the power can be expressed as

$$E(P) = 2\pi R^2 E(I(R, \pi)) \int_0^\pi \frac{E(p_{ms}^2(R, \theta))}{E(p_{ms}^2(R, \pi))} \sin \theta \, d\theta.$$  (10)

Note that we cannot derive the value of $E(I(R, \pi))$ from the theory, because applying the principle of equipartition of energy we can only calculate relative amplitudes of modes. Thus, to calculate the absolute value of the power output we have to replace in (10) theoretical value $E(I(R, \pi))$ by the measured one $\tilde{I}(R, \pi)$. Consequently, the measurement error $\delta_I$ appears in theoretical formulæ, because we substitute

$$E(I(R, \pi)) = \tilde{I}(R, \pi)(1 \pm \delta_I),$$  (11)

and thus

$$E(P) = E(P')(1 \pm \delta_I),$$  (12)

where $E(P')$ is calculated along to Eq. (10), in which, however, $E(I(R, \pi))$ have been replaced by $\tilde{I}(R, \pi)$.

From (9) and (12) we obtain

$$(1 - \varepsilon_P)(1 - \delta_I) \leq \frac{P}{E(P')} \leq (1 + \varepsilon_P)(1 + \delta_I)$$  (13)

We are used to expressing the above discussed quantities in dB. The acoustic level $L_{E(P')}$ of estimated power $E(P')$ and the uncertainty in evaluating the level of power output, $\Delta L_P = |10 \log P/E(P')|$, resulting from the applied model, fulfill the relations

$$L_{E(P')} = 10 \log \frac{E(P')}{P_o} = L_{I(\pi)} + 10 \log(2\pi R^2) + 10 \log K,$$  (14)

$$\Delta L_P \geq -10 \log[(1 - \varepsilon_P)(1 - \delta_I)].$$  (15)

where $P_o = 10^{-12}$ [W] and $K$ denotes integrand in (10).

The above presented formulæ constitute a base for introducing the idea of applying the circular duct as a reference sound source.

THE CIRCULAR DUCT AS A REFERENCE SOUND SOURCE

In [3] we compared theoretical and experimental data of the power output measured by means of three different methods (intensity on the duct cross-section, intensity and pressure in the far field outside the duct) for three values of $ka$. The relation $\Delta L_P \leq L_{I(\pi)}, L_{I(\pi)}$ being the error of experiments, has been fulfilled for each value of $ka$ and for all measuring methods.

That made us consider a possibility of applying the cylindrical duct as a reference source.

The idea is as follows: we measure the level of sound intensity in one point $(R, \pi)$ on the axis and with the help of Eq. (14) and Fig. 1 evaluate the power output.

Knowing the error in measuring the sound intensity $\delta_I$ (usually specific for a considered measuring system), relation (15) and Fig. 2, presenting the normalized standard deviation of power $\varepsilon_P$, allows us to establish the error of power estimation $\Delta L_P$.

Another possibility is to determine the range of $ka$ for which it is possible to measure the power level with the error not exceeding a fixed value. For a chosen $\Delta L_P$ and known $\delta_I$ the standard deviation can be calculated from Eq. (15). Having done that, the next step is to determine, with the help of Fig. 2, the range of nondimensionalized wavenumber, for which
the inequality is fulfilled. If the wavenumber (or frequency) is fixed, then we can derive the radius \( a \) of the duct, by means of which we can carry out experiments with fixed error in the power level.

![Figure 1: Values of \( K \) versus \( ka \).](image1)

![Figure 2: Standard deviation \( \varepsilon_p \).](image2)

**CONCLUSIONS**

The presented model can be useful in carrying thorough analysis of the acoustic field inside and outside the duct, by means of such quantities as the pressure and the intensity directivity functions, the power-gain function, the power output etc. The derived mathematical formulae allow to estimate not only these quantities but also the error of estimation.

With the help of Figs 1 and 2, the power output and the measuring error can be estimated basing on the measuring data in only one point on the axis. It is also possible to establish conditions (by choosing adequate duct radius) under which the error of evaluating the power output does not exceed some fixed value.

The modal character of the phenomena inside the duct shows also in the far field outside. On both figures, the values vary significantly in the neighborhood of the cut-on frequencies, when a consecutive mode is excited (\( ka = 3.83, 7.02, 10.17, \ldots \)).

For \( ka \leq 3.83 \), the normalized standard deviation \( \varepsilon_p \) is equal to 0 (see Fig. 1), because in this range of \( ka \) only the plane wave can propagate in the semi-infinite circular wave-guide, and thus there are no random phases and no uncertainty coming from theoretical model.

**REFERENCES**


ACOUSTIC PROBING OF THE ATMOSPHERE

Igor P. Tchountchouzov
Institute of the Atmospheric Physics, Pyzhevskii 3, 109017 Moscow, Russia

SUMMARY

The tendencies in using of the acoustic tomography methods in the atmosphere outlined during the last years [1-4]. There was a number of circumstances causing these tendencies:

the successes in practical use of acoustic tomography in the ocean [5, 6];
the results of the inclined acoustic probing by the pulse sources of both the atmospheric boundary layer [3, 7-8] and upper atmosphere [9-11];
the intensive development for the last ten years of the wave theory of sound propagation in moving inhomogeneous media [12-15].

One of the problems, which arises in acoustic probing of the atmosphere, is to take into account the influence of the mean wind velocity stratification and its space-time fluctuations upon the probing pulse parameters, such as the travel time and the time duration of the signal, which are the entrance parameters of the inverse problem. In this work we’ll summarize some results of the theoretical study of acoustic pulse propagation in the moving stratified atmosphere and experimental results of the inclined probing of the atmosphere. Among them we’ll emphasize the following:

1. The wind velocity stratification leads to a certain dependence of the signal shape, travel time and the number of arrivals on the azimuth of the reception point. This dependence was studied both theoretically and in the experiment [3, 7].

2. There were measured the time fluctuations of the pulse travel time and the time interval between different arrivals, which as was found are mainly caused by the wind velocity fluctuations.

3. There was found the influence of the thin layered structure of the wind velocity and temperature upon the shape and time duration of the acoustic signal, reflected from both the boundary and stratospheric layers [9, 10]. This influence results in arising of the multi path rays and in the time delay of the reflected signal.
PULSE PROPAGATION IN THE MOVING STRATIFIED LAYER

The inclined method of the acoustic probing of the atmosphere, which uses the explosions of different power as acoustic pulse sources, has been essentially developed in the last years [3, 7-11]. While probing by explosions with the power of about 0, 26-kt three-nitrate-fuel oil slurry (TNFO) [9] or of 20-30t TNFO [10] the three main arrivals were registered by receivers placed on the ground at distances of 200-300km from the source. It was found from the travel time, that the first arrival propagates in the sound channel near the ground, the second arrival arises due to pulse reflection from the stratospheric layer concluded between altitudes of 30km and 50km, and the third arrival is reflected from the thermospheric layer, laying at the altitudes more than 90 km.

The arrivals as was found differ from each other not only by a travel time, but also by shape and time duration. The thermospheric arrival has its own shape of U-wave with time duration much more than that of the initial signal near the source. Such waveform is the result of the \( \pi /2 \) phase shift of the spectral components of N-wave after touching with the caustic. The N-wave in its turn was formed due to nonlinear steepening with height of the initial pulse.

The stratospheric signal has quite another character: it consists from the irregular set of arrivals, which is supposed to be the pulse reflections from the thin discrete layers [9] with different vertical gradients of the wind velocity and temperature.

Thus the shape and duration of a received signal carry the integral information about the wind velocity and temperature inhomogeneities, crossing the propagation path of the signal. To extract this information it's necessary to know the relation between the parameters of the received signal and the wind velocity and temperature profiles in the sound channel, i.e. to solve the direct problem of pulse propagation in the moving stratified atmosphere. This problem can not be solved by any high-frequency approach, because in the case of a broadband frequency spectrum of the pulse (like an explosion) the main part of the wave energy is radiated at low frequencies, for which the wave length is comparable with the vertical scale of the wind velocity and temperature inhomogeneities. Therefore we have used the normal mode method [3, 12, 13, 15], which takes into account the movement of the atmosphere and is valid for the arbitrary wavelengths. By this method the solution was obtained, which describes the wave field of a point pulse source [3] in the layer above the ground surface \( z=0 \) (both in the case of rigid and finite impedance) with the arbitrary wind velocity \( V(z) \) and temperature \( T(z) \) profiles. The wave field was calculated in cases of exponential layer

\[
V(z) = V_\infty \left[ 1 - \exp(-z/h) \right], \quad T(z) = T_\infty \left[ 1 + \Delta \left[ 1 - \exp(-z/h) \right] \right],
\]

and Epstein layer

\[
V(z) = V_\infty \left[ 1 - \cosh^2(z/2h) \right], \quad T(z) = T_\infty \left[ 1 + \Delta \left[ 1 - \cosh^2(z/2h) \right] \right],
\]

where \( h \) is the vertical scale of layer inhomogeneity, \( V_\infty \equiv V(\infty), \quad T_\infty \equiv T(\infty), \quad \Delta \equiv [T(\infty) - T(0)]/T(0). \)

Assuming that \( \beta = V_\infty /c_0 < \Delta << 1 \), where \( c_0 \) is the sound velocity at \( z=0 \) and neglecting in the Helmholtz equation by a small terms of order \( \beta^2 \) we found it's solution.

At large horizontal distance from the source \( (kr > 1) \), \( k = \omega/c_0 \), \( \omega \) is the central frequency of the pulse spectrum) the pulse wave field \( P(r, \phi, t) \approx p_0/R_0 \) (where \( p \) is the acoustic pressure, \( p_0 \) is the pulse peak

- 560 -
pressure near the source at distance \( R_s \leq c \) is a function of a nondimensional time \( \omega (t - r/c(\omega)) = \omega t \), distance \( r/L \) and effective thickness of the layer \( M(\omega) = 2e^{r^2/k} \), where \( \omega \) is the azimuth of the detection point relative to the wind velocity direction, \( c = 2 \beta \cos \varphi + \Delta, L(\omega) = 4 \pi c \omega / k \) is the minimum horizontal period of mode's interference, \( c_c = c_c(1 + r/2) \).

If \( c \leq 2 \), then the waveguide is formed, and if, on the contrary, \( c < 0 \), then the antiguide propagation takes place.

In the case of \( M \gg 1 \), when the ray approach is valid, the initial pulse "splits" at large enough distance from the source into the set of arrivals A, B, C and the "tail" (fig. 1); the resultant duration of the signal is limited by the time interval \( t = (c_c - c_c^{-1}) = \omega / 2c_c \). The arrivals are formed in certain time moments, corresponding to a condition of the constructive interference of a large number of modes with their own frequencies \( \omega (\varphi), n = 0, 1, 2, \ldots \), laying within the spectral width of the initial pulse. These time moments depend on distance \( r \) and parameters of profile \( e(\varphi) \) and \( h \), but they don't depend on initial pulse duration \( \tau = \omega^{-1} \). So the small fluctuations of \( V \) or \( T \) cause the corresponding fluctuations of the arrival's time moments. The measuring of the small fluctuations of the arrival's time moments at the set of receivers can be used for solving the linear inverse problem [5] in order to retrieve the small fluctuations of \( V \) and \( T \).

When \( M \) decreases (as function of \( \phi \) or \( \omega \)), then the signal transforms into the continuous wave packet, being a result of the interference of a finite number of modes. Each of mode is the wave packet with its own frequency \( \omega (t, r, \phi) \) and horizontal wavenumber \( k_x(\omega) \), satisfying the equation \( \partial_k / \partial \omega = 1/r \). In this case the characteristic frequency of the wave packet is close to the central frequency of the initial pulse spectrum.

At last, when \( M < 1 \), then the only first mode propagates in the waveguide and its evolution with distance over the rigid surface is well described by the linear KDV-equation for the cylindrical waves [16].

THE INCLINED PULSE PROBING OF THE ATMOSPHERIC BOUNDARY LAYER.

The considered above transformation of the pulse wave field as function of \( M \) was observed in the experiments. This experiment had the purpose to investigate the influence of the time variations of the wind velocity and temperature profiles in the boundary layer upon the travel time, waveform and duration of signals, being received on the ground surface at different azimuths and distances (20m-4.5km) from the pulse source. We used a detonation source [7], which radiated the acoustic pulses with the shape similar to that of the explosions. The period

![Fig. 2. Waveform transformation (n=2 km) with changes in the state of the boundary atmospheric layer. (a), (b) Unstable convective stratification. (c) State of stable stratification.](image)

![Fig. 3. Time variation of the wind velocity \( V(t) \) and its direction \( \varphi(t) \), causing the pulse waveform transformation in Fig. 2.](image)

![Fig. 4. Azimuthal anisotropy of the signal shape recorded in two directions \( \varphi_1 \) and \( \varphi_2 \) at the distance 2 km from the source S.](image)
of pulse radiation was about 20s in our case (we could vary it from 1s to 1min).

The continuous control of the wind velocity profiles up to 400m above the ground was carried out by Doppler sodar and by anemometers, placed at the mast with the height of 56m. The temperature was measured by thermometers placed on the mast or with the use of radioacoustic sounding. The experiments showed that the wind velocity stratification in the boundary layer influences upon the pulse wave field more strongly than the temperature stratification. The azimuthal anisotropy of the signal parameters was found, such as a number of separated arrivals and signal duration (fig. 4).

During the formation of the temperature inversion and wind shear we observed the process of the pulse "splitting" (fig. 2—3).

It was found, that the small time variations of the wind velocity profiles with magnitudes in the range of 0.5-2m/s cause the variations of the pulse travel time and the interval between different arrivals (received at distances of 2.7km and 4.5km) of about 10-20ms.

The travel time and duration of the signal, being the entrance parameters of the inverse problem, well followed on the small fluctuations of the wind velocity and pressure (measured by the microbarographs), caused by the internal waves with periods of about 5-20min. Thus the possibility arises in presence of a net of receivers placed at different distances and azimuths from the source to retrieve by the linear tomography methods [17] the small space-time fluctuations of the wind velocity and temperature in the boundary layer. The continuous acoustic control of the wind velocity and temperature field within the finite space volumes opens a new perspective for both the study of the space dynamics of the atmosphere (internal waves, meteofronts, eddy structures and others) and for solving the problem of the pollution transport in the environment.

REFERENCES

2. B.M. Kopyrov, "The sound refraction in the stratified atmospheric surface layer." Izvestiya, Atmospheric and Oceanic Physics, 1989, 25, 8, 819-825.
SOUND TRANSMISSION AT A SUDDEN AREA EXPANSION IN A FLOW DUCT

Mats Åbom, Börje Nilsson
Technical Acoustics Dept. of Vehicle Engineering, Royal Inst. of Technology.
S-100 44 Stockholm.

SUMMARY

The objective of the present work is to show how the theory of Nilsson and Brander [4] can be modified and based on a relaxed Kutta-condition. The model obtained in this way will be the most general presented for the problem of sound transmission at a sudden area expansion. In particular the model is not restricted to low frequencies since it is based on a complete field description with all modes included.

INTRODUCTION

The understanding of linear sound propagation in subsonic flow ducts is still incomplete, a number of partly unsolved problems exist. To this category belongs the question of interaction effects between an acoustic field and a flow field. In the subsonic regime this interaction is normally small since the dominating spatial and temporal scales of the two fields are very different. However, it has been known for some time that in regions with flow separation (i.e., unstable shear layers) a strong interaction can occur [1]. This strong interaction can have an important influence on the reflection and transmission of sound in such regions. A suitable starting point for analysing this effect is a sudden area expansion in a duct with a stationary mean flow, see figure 1. A number of investigations of this problem has been published, here we shall briefly discuss the most important contributions.

---

Figure 1. Flow situation at a sudden area expansion in a duct. Region I: Forming of jet, negligible flow expansion; Region II: Mixing region, flow expansion and generation of turbulence; Region III: Fully expanded and settled flow.
EARLIER WORKS

Ronneberger [1] suggested a theory valid for the plane wave region, i.e., for duct diameters much larger than the acoustic wavelength. The theory is based on the assumption that the change of the sound field between region I and III can be obtained via a quasi-steady perturbation of the flow field. To account for the irreversibility of the process (the turbulence production) entropy waves are assumed in region III. From these assumptions and by applying conservation of energy, momentum and mass to a control volume extending from region I to III, the reflection and transmission properties could be analysed. However, the result only fits the experimental data at very low frequencies and Mach-numbers. It can also be mentioned that Alfredson and Davies have presented a model [2] which is essentially equivalent with Ronneberger’s. Cummings [3] pointed out that one problem with the work of Ronneberger; is that he assumes that the entropy waves occur in region III where in reality they occur in region II and have decayed in region III. By introducing a control volume which only includes region I and then applying an analysis similar to that used by Ronneberger, Cummings derived an expression for the reflection coefficient. Cummings theory gives a better fit to experimental data than Ronneberger’s but there are still discrepancies. Nilsson and Brander [4] have presented a general theory for sound propagation at a sudden area expansion in a cylindrical duct. This theory can be applied to the problem in question although it assumes an ideal flow situation where the jet never expands so that region I extends indefinitely. This theory is not restricted to the plane wave region and the origin of the entropy waves is present in the form of a growing acoustic mode (hydrodynamic mode) generated at the inlet edge of the expansion. The excitation of this growing acoustic mode is closely related to the so-called Kutta-condition, which in ideal fluid theory determines the flow behavior at a sharp edge. Nilsson and Brander only present a very limited comparison between their theory and measurements. Ronneberger [5] has presented a "one-dimensional" version of Nilsson’s and Brander’s theory. This theory is only valid for plane waves, it includes a simplified Kutta-condition but excludes all higher order mode effects. From comparison with measurements Ronneberger concludes that; his "one-dimensional" version of Nilsson’s and Brander’s theory gives a better agreement than earlier theories. However, still significant discrepancies occur for certain cases which Ronneberger attributes to: i) effects of non-propagating higher order modes; ii) uncertainty about the correct form of the Kutta-condition. Regarding the last point Ronneberger [5] points out that the classical Kutta-condition only gives a good description for small Strouhal-numbers \((St = f \delta / U)\), where \(f\) is frequency, \(\delta\) is boundary layer thickness upstream of the inlet edge, and \(U\) is mean flow velocity. Ronneberger [5] also suggests a modified Kutta-condition which gives an improved fit to the data for high Strouhal-numbers.

THEORY

The starting point for the analysis is the situation depicted in Figure 2. The fluid is assumed to be ideal with a uniform flow distribution (plug flow) over the inlet duct cross-section. A small or a moderate Mach-number is assumed and the speed of sound and density is the same everywhere in the fluid. Linear acoustics is also assumed and the ducts are considered to be rigid. The acoustic pressure for harmonic sound fields \(\exp(-)\) in regions A to C satisfies Helmholtz equation (with appropriate boundary conditions) and can be written as mode sums. The detailed expressions can be found in references [6]. In this reference a short discussion of the causality of the modes is also given and its shown that all modes behave as expected (i.e. decay in their direction of propagation), with the exception of one mode in region B which is growing in its
direction of propagation. This mode corresponds to a so-called hydrodynamic mode and is related to the Kelvin-Helmholtz instability of the free jet in region B.

Figure 2. Two concentric circular ducts. The inner semi-infinite duct with radius $a$ carries a uniform mean flow with Mach-number $M$. In the outer infinite duct with radius $b$ the fluid is non-moving.

If we now assume a mode $n$ incident from region A we can write the total acoustic pressure as

$$ p = p_m + A_n^s \phi_{nn}^s(p) \exp(i\alpha_{nn}^s x), $$

(1)

where $p_m$ is the scattered pressure. By taking a spatial Fourier transform

$$ P(\alpha, p) = \int \exp(-i\alpha x) p(x, p) dx $$

of all the governing equations including their boundary conditions it is possible to derive the following Wiener-Hopf equation [6]

$$ G(\alpha) \Psi_-(\alpha) + J_+(\alpha) + \frac{iA_n^s \phi_{nn}^s(\alpha)}{\alpha - \alpha_{nn}^-} = 0, $$

(2)

where $G$ is a function of both $\alpha$ and the frequency $\omega$ and represents the dispersion equation for region B. The index +/- denotes the region of regularity of the spatial Fourier transforms, i.e., upper/lower $\alpha$-plane. Here only the essential step in the relaxation of the Kutta-condition will be presented and for details of the rest of the solution procedure, see reference [6]. When the function: $\Psi_+(\alpha) = \frac{\partial P(\alpha, a^+)}{\partial p}$ is known the total sound field can be calculated and by also studying other incident waves the complete scattering matrix can be calculated [6]. When this matrix is known we can for instance analyse the case in Figure 1. By splitting $G$ in two parts $G = G_+G_-$ equation (2) gives, after we have followed the standard steps in the Wiener-Hopf procedure,

$$ G_+ \Psi_+ + \frac{K_+}{\alpha - \alpha_{nn}^+} = \text{polynomial in } \alpha, $$

(3)

where $K_+$ is a constant. With a strict Kutta-condition the polynomial becomes identically zero which is the case treated in reference [4]. If we relax this condition then in the general case we have equation (3). The next step after a strict Kutta-condition will correspond to putting the right hand side equal to a constant $C_+$ (i.e., independent of $\alpha$). Of course this $C_+$ can be a
function of the frequency or rather the Strouhal-number (St). If we solve $\Psi_-$ from equation (3) we obtain

$$\Psi_- = \frac{K_-}{G_-} \left\{ \frac{-1}{\alpha - \alpha_{\lambda_n}^*} + C_s' \right\}. \tag{4}$$

In this equation we have a pole in $G_-$ corresponding to the hydrodynamic mode ($\alpha_{\lambda_n}^*$). The value of $C_s$ will determine to what degree we excite this hydrodynamic mode and by choosing $C_s$ (relaxing the Kutta-condition) we can even "turn off" this mode. This is achieved by choosing $C_s$ so that the expression in the parenthesis in equation (4) is zero for $\alpha = \alpha_{\lambda_n}^*$. This gives:

$$C_s' = \frac{1}{(\alpha_{\lambda_n}^* - \alpha_s^*)};$$

to obtain a partly excited hydrodynamic mode we simply multiply this with a dimensionless function of $St$. This function should be positive valued and only assume values in the interval 0 to 1. The exact form of the function must be determined by measurements or some more complex theory.

For examples of results calculated using the modified theory presented here please refer to reference [6].

![Diagram](image)

**Figure 3.** One example of a calculated plane wave reflection coefficient (wave incident from A) at a sudden area expansion of the type depicted in Figure 2. Mach-number 0.3, $b/a$=1.54.

**REFERENCES**

Author index

Abe Kazutaka vol II 349
Abe Masato vol IV 3
Abel Sharon M. vol III 3, vol III 195
Adachi Seji vol III 385
Ahnert Wolfgang vol II 353, 357
Akinaga Masahiro vol I 325
Apfel Robert E. vol IV 291
Arai Shuichi vol III 7
Arai Yasuhiko vol III 11
Asselineau M. vol II 361
Attenborough Keith vol II 3, 365
Azevedo Rui G. vol IV 7
Azn C. A. Nierop vol II 223

Baggeroer Arthur B. vol I 211
Barron Michael vol II 369, 373
Behler Gottfried vol IV 185
Behne Dawn M. vol III 15
Berengier Michel vol II 7
Berg Tone vol I 215
Berge Svein vol IV 11
Berglund Birgitta vol III 199, vol II 11, 17
Berhout A.J. vol II 377
Berntson Alf vol II 381
Bertsch Matthias vol III 393
Bessac François vol I 471
Bilsen Frans A. vol III 203
Björn Ole-Herman vol IV 15
Bobrovnietskii Yu. I. vol I 479
Blaże A. vol I 0
Blauert Jens vol III 207
Bodén Hans vol III 531
Bohn Ocke-Schwen vol III 19
Bonilha M. W. vol I 483
Borg Carsten vol II 385
Botte Marie Claire vol III 211
Bottelidooren Dick vol II 21

Bradley J. S. vol II 389
Bratt J. F. vol I 487
Brezeale M. A. vol I 73
Brekke Arild vol II 25
Brinks J.R. vol II 393
Brouard B. vol II 397, vol II 401
Briel Per V. vol I 3, 491, vol II 29, 405, vol IV 3
Bruneau Anne-Marie vol IV 189
Bruneau Michel vol IV 199, 197
Bull M. K. vol III 535
Busch-Vishniac Ilene J. vol IV 201

Camiz P. vol III 215
Carcattera Antonio vol I 495, 499
Carter Norman L. vol II 33
Carvalho António Pedro O. vol II 409
Cato Douglas H. vol I 219
Chaigne Antoine vol III 397
Champelovier Patricia vol II 37
Cheatham M.A. vol III 221
Cheng Nai Ping vol IV 19
Chen Tong vol III 401
Cholewa Wojciech vol I 503
Chu Wing T. vol IV 25
Cocchi Alessandro vol II 413, 417
Collins Michael D. vol I 223
Condomines Ana Barjau vol III 389
Cox Trevor J. vol II 421
Crum Lawrence A. vol IV 295
Cuschieri J. M. vol I 13

Dąbrowski Zbigniew vol II 41, 45
Daigle Gilles A. vol II 49
Dalenhäck Bengt-Inge vol II 425, 429
Dalmasso V. vol I 227
D'Antonio Peter vol I 17
Daumal i Domènec Francesc  vol III 405
Davies P.O.A.L.  vol III 23
de Billy M.  vol I 21
de Graaff D. F.  vol II 57
de Jong Ronald G.  vol II 61, 65
Delius Michael  vol III 319
Delsanto P. P.  vol I 329
Depollier C.  vol II 433
Dérogis Philippe  vol III 409
Desmet C.  vol IV 333
Díaz César  vol III 27
Diebold G. J.  vol IV 337
Dinghua Guan  vol III 31
Dhainaut M.  vol II 69
Don C. G.  vol II 73
Dowson Susan P.  vol IV 29
Dobrucki Andrzej  vol IV 205
Dutilleux Pierre  vol III 413
Dutono Titon  vol III 35
D'Spain Gerald L.  vol I 231

Ebbini Emad S.  vol III 323
Efimtsov Boris M.  vol I 507
Elbaum Charles  vol I 341
Emerit Marc  vol II 437
Engan Helge E.  vol III 327
Evans J. M.  vol I 345
Eversman Walter  vol I 25

Fahy F. J.  vol I 29, vol II 77
Farina Angelo  vol II 81, vol II 441, 445, 449
Fausti Patrizio  vol III 453
Fehlmann R.-F.  vol III 417, 423
Fessina Michael I.  vol II 85
Fields James M.  vol II 89
Filippi Paul J. T.  vol II 93
Fletcher Neville H.  vol IV 33
Flottorp Gordon  vol III 225
Frederiksen Erling  vol IV 209

Fredò Claes R.  vol I 33
Freitas Diamantino R. S.  vol IV 37
Friberg Anders  vol III 431
Fricke Fergus  vol II 457
Frivik Svein A.  vol I 235
Fröysa Kjell-evin  vol I 349
Fujii Kensaku  vol IV 41
Fujiwara Kyoji  vol II 97
Furuc Yoshihiro  vol II 461
Fukudome Kimitoshi  vol IV 45
Furui Sadaoki  vol III 497
Fyne K. R.  vol II 101

Gabrielli P.  vol I 239
Gade A. C.  vol II 465
Gade Svend  vol IV 49
Galembo Alexander  vol III 435
Gallego-Juárez Juan A.  vol I 41
Garnier B.  vol IV 55, vol I 45
Gatehouse R. W.  vol III 229
Gaunaud Guillermo C.  vol I 49
Genuit Klaus  vol II 469, vol III 233
Gerard Christophe  vol III 45
Gerdes F.  vol II 105
Gerfault L.  vol I 353
Giguère Christian  vol III 237
Ginn Kevin Bernard  vol IV 61, 65
Gjestland Truls  vol II 109
Goertz Anselm  vol IV 69
Gonneau E.  vol IV 73
Groothoff Beno  vol III 241
Gusev Vitaly  vol I 53
Guzas D.  vol I 57
Gramatica Bruno  vol II 49
Griesinger David  vol II 473, 477
Griffiths J.W.R.  vol I 243
Grillon Vincent  vol II 481
Gummer Anthony W.  vol III 245
Grim Michael  vol IV 77
Gustafsson Jan-Inge  vol II 485
Guyader J. L.  vol II 117
Grönberg Fredrik  vol I 511
<table>
<thead>
<tr>
<th>Name</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habault D. M. L.</td>
<td>vol I 61</td>
</tr>
<tr>
<td>Hagen Ole</td>
<td>vol II 121</td>
</tr>
<tr>
<td>Hansen Uwe J.</td>
<td>vol IV 303</td>
</tr>
<tr>
<td>Hanson Carl E.</td>
<td>vol II 125</td>
</tr>
<tr>
<td>Hanson Roger J.</td>
<td>vol IV 299</td>
</tr>
<tr>
<td>Hartmann Bruce</td>
<td>vol I 65</td>
</tr>
<tr>
<td>Hashimoto Osamu</td>
<td>vol III 57</td>
</tr>
<tr>
<td>Heimdal Andreas</td>
<td>vol III 331</td>
</tr>
<tr>
<td>Hermansky Hynek</td>
<td>vol III 61</td>
</tr>
<tr>
<td>Harmegnies Berard</td>
<td>vol III 53</td>
</tr>
<tr>
<td>Hemilä Simo</td>
<td>vol III 249</td>
</tr>
<tr>
<td>Henry M.</td>
<td>vol II 489</td>
</tr>
<tr>
<td>Hernandez Ange</td>
<td>vol III 335</td>
</tr>
<tr>
<td>Herráez Marta</td>
<td>vol IV 85</td>
</tr>
<tr>
<td>Herzog P.</td>
<td>vol IV 91</td>
</tr>
<tr>
<td>Hettergott Alexandra</td>
<td>vol III 439</td>
</tr>
<tr>
<td>Hobæk Halvor</td>
<td>vol I 69</td>
</tr>
<tr>
<td>Hojan E.</td>
<td>vol III 253</td>
</tr>
<tr>
<td>Holger David K.</td>
<td>vol II 129</td>
</tr>
<tr>
<td>Hongo Satoshi</td>
<td>vol I 357</td>
</tr>
<tr>
<td>Hornowski Tomasz</td>
<td>vol I 77</td>
</tr>
<tr>
<td>Horoshenkov K. V.</td>
<td>vol II 493</td>
</tr>
<tr>
<td>Hosokawa Atsushi</td>
<td>vol I 361</td>
</tr>
<tr>
<td>Hthersall D. C.</td>
<td>vol II 133</td>
</tr>
<tr>
<td>Hovem Jens M.</td>
<td>vol I 247</td>
</tr>
<tr>
<td>Howarth M. J.</td>
<td>vol II 497</td>
</tr>
<tr>
<td>Holm Sverre</td>
<td>vol III 339</td>
</tr>
<tr>
<td>Houtsma A. J. M.</td>
<td>vol IV 307</td>
</tr>
<tr>
<td>Huan Hou Chao</td>
<td>vol IV 95</td>
</tr>
<tr>
<td>Huopaniemi Jyri</td>
<td>vol III 447</td>
</tr>
<tr>
<td>Hunt Melvyn J.</td>
<td>vol III 65</td>
</tr>
<tr>
<td>Ichiro Yamada</td>
<td>vol II 339</td>
</tr>
<tr>
<td>Idczak H.</td>
<td>vol II 137</td>
</tr>
<tr>
<td>Iida Kazuhiro</td>
<td>vol III 443</td>
</tr>
<tr>
<td>Ikeda Masahiro</td>
<td>vol I 81</td>
</tr>
<tr>
<td>Ikeda Takashi</td>
<td>vol III 69</td>
</tr>
<tr>
<td>Imamura Tohru</td>
<td>vol I 365</td>
</tr>
<tr>
<td>Ishida Koji</td>
<td>vol II 501</td>
</tr>
<tr>
<td>Ise Shiro</td>
<td>vol I 85</td>
</tr>
<tr>
<td>Juliul Pop</td>
<td>vol I 369</td>
</tr>
<tr>
<td>Jacobsen Finn</td>
<td>vol IV 99</td>
</tr>
<tr>
<td>Jaffe Dr. Christopher</td>
<td>vol II 509</td>
</tr>
<tr>
<td>James M.</td>
<td>vol IV 81</td>
</tr>
<tr>
<td>Jaroch Andrzej</td>
<td>vol II 145</td>
</tr>
<tr>
<td>Jensen Finn B.</td>
<td>vol I 253</td>
</tr>
<tr>
<td>Jimenez-Dianderas Carlos R.</td>
<td>vol II 513</td>
</tr>
<tr>
<td>Jin Guoliang</td>
<td>vol I 257</td>
</tr>
<tr>
<td>Job R. F. Scames</td>
<td>vol II 287, 291, 295</td>
</tr>
<tr>
<td>Johnson Paul A.</td>
<td>vol I 89</td>
</tr>
<tr>
<td>Johansson Örjan</td>
<td>vol IV 103</td>
</tr>
<tr>
<td>Juliie Jean-Pascal</td>
<td>vol II 517</td>
</tr>
<tr>
<td>Kajikawa Yoshinobu</td>
<td>vol IV 213</td>
</tr>
<tr>
<td>Kahle Eckhard</td>
<td>vol II 521</td>
</tr>
<tr>
<td>Kammeringer Ewald</td>
<td>vol II 513</td>
</tr>
<tr>
<td>Kaneko Seiji</td>
<td>vol IV 217</td>
</tr>
<tr>
<td>Kang Kyeongok</td>
<td>vol IV 221</td>
</tr>
<tr>
<td>Karjalainen Matti</td>
<td>vol III 451</td>
</tr>
<tr>
<td>Kawanishi Yasuhito</td>
<td>vol II 529</td>
</tr>
<tr>
<td>Kido Ken’iti</td>
<td>vol III 73</td>
</tr>
<tr>
<td>Kihlman Tor</td>
<td>vol IV 311</td>
</tr>
<tr>
<td>Kikuchi Toshiaki</td>
<td>vol IV 225</td>
</tr>
<tr>
<td>Kim Dae-won</td>
<td>vol III 77</td>
</tr>
<tr>
<td>Kleiner M.</td>
<td>vol II 533</td>
</tr>
<tr>
<td>Kojima Toshiro</td>
<td>vol I 373</td>
</tr>
<tr>
<td>Kohlrauch Armin</td>
<td>vol III 455</td>
</tr>
<tr>
<td>Kolaini A. R.</td>
<td>vol I 261</td>
</tr>
<tr>
<td>Kollmeier Birger</td>
<td>vol III 81</td>
</tr>
<tr>
<td>Kolomenskii Al. A.</td>
<td>vol I 93, 97</td>
</tr>
<tr>
<td>Korman Murray S.</td>
<td>vol I 101</td>
</tr>
<tr>
<td>Kosek Miloslav</td>
<td>vol I 105</td>
</tr>
<tr>
<td>Krämer Peter S.</td>
<td>vol II 537, 541</td>
</tr>
<tr>
<td>Krebber Winfried</td>
<td>vol II 85</td>
</tr>
<tr>
<td>Kringlebott Magne</td>
<td>vol III 257</td>
</tr>
<tr>
<td>Kropp Wolfgang</td>
<td>vol I 109, vol II 157</td>
</tr>
<tr>
<td>Kryklov Victor V.</td>
<td>vol II 161</td>
</tr>
<tr>
<td>Kuperman W. A.</td>
<td>vol I 113</td>
</tr>
<tr>
<td>Kuramoto Kazuoki</td>
<td>vol III 261</td>
</tr>
<tr>
<td>Kurra Selma</td>
<td>vol II 165</td>
</tr>
<tr>
<td>Kuroiwa K.</td>
<td>vol IV 107</td>
</tr>
<tr>
<td>Kuttruff Heinrich</td>
<td>vol II 545</td>
</tr>
<tr>
<td>Name</td>
<td>Volume/Issue</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Kuwahara Shin'ya</td>
<td>II 553</td>
</tr>
<tr>
<td>Kuwano Sonoko</td>
<td>II 169</td>
</tr>
<tr>
<td>Kaarmann Hans T.</td>
<td>III 343</td>
</tr>
<tr>
<td>Käes A.</td>
<td>I 117</td>
</tr>
<tr>
<td>Lai J. C. S.</td>
<td>II 173</td>
</tr>
<tr>
<td>Laissus Jean Jacques</td>
<td>I 515</td>
</tr>
<tr>
<td>Lai Xiaoming</td>
<td>III 347</td>
</tr>
<tr>
<td>Lauterborn Werner</td>
<td>I 121</td>
</tr>
<tr>
<td>Larsson Conny</td>
<td>II 177</td>
</tr>
<tr>
<td>Lavrentiev Jyri</td>
<td>III 543</td>
</tr>
<tr>
<td>Lawrence D. E. P.</td>
<td>I 265</td>
</tr>
<tr>
<td>Leander Johan Leo</td>
<td>I 269</td>
</tr>
<tr>
<td>Leclaire P.</td>
<td>I 125, 377</td>
</tr>
<tr>
<td>Lehner Hilmare</td>
<td>III 265</td>
</tr>
<tr>
<td>Lee Yung P.</td>
<td>I 273</td>
</tr>
<tr>
<td>Letourneaux Fabien</td>
<td>II 181</td>
</tr>
<tr>
<td>Letowski Tomasz</td>
<td>III 269</td>
</tr>
<tr>
<td>Lyamshev Leonid M.</td>
<td>I 129</td>
</tr>
<tr>
<td>Li Ma</td>
<td>I 277</td>
</tr>
<tr>
<td>Liasjö Kåre H.</td>
<td>II 185</td>
</tr>
<tr>
<td>Lie Torgrim</td>
<td>III 351</td>
</tr>
<tr>
<td>Lin Fa</td>
<td>IV 229</td>
</tr>
<tr>
<td>Linjama Jukka</td>
<td>IV 111</td>
</tr>
<tr>
<td>Lotsberg O.</td>
<td>III 355</td>
</tr>
<tr>
<td>Lundberg Karl-Ola</td>
<td>I 519</td>
</tr>
<tr>
<td>Lund Per</td>
<td>I 381</td>
</tr>
<tr>
<td>Lynge Claus</td>
<td>II 557</td>
</tr>
<tr>
<td>Lyon Richard H.</td>
<td>II 189</td>
</tr>
<tr>
<td>Løkberg Ole J.</td>
<td>IV 115</td>
</tr>
<tr>
<td>Madehus Christian</td>
<td>II 195</td>
</tr>
<tr>
<td>Maeda Toshiyuki</td>
<td>IV 119</td>
</tr>
<tr>
<td>Majewski Wojciech</td>
<td>III 273</td>
</tr>
<tr>
<td>Malbos Francois</td>
<td>III 89</td>
</tr>
<tr>
<td>Malecki Ignacy</td>
<td>I 523</td>
</tr>
<tr>
<td>Markus Stefan</td>
<td>I 527</td>
</tr>
<tr>
<td>Marshall Arthur Harold</td>
<td>I 561, 565</td>
</tr>
<tr>
<td>Marquis C.</td>
<td>I 531</td>
</tr>
<tr>
<td>Masujima Tsutomu</td>
<td>I 385</td>
</tr>
<tr>
<td>Matsumoto Hiroshi</td>
<td>III 93</td>
</tr>
<tr>
<td>Matsumura Masafumi</td>
<td>III 97</td>
</tr>
<tr>
<td>McCoy John J.</td>
<td>I 133</td>
</tr>
<tr>
<td>McGowan Richard S.</td>
<td>III 101</td>
</tr>
<tr>
<td>Melon Manuel</td>
<td>I 389</td>
</tr>
<tr>
<td>Metsaveer Jaan</td>
<td>I 137</td>
</tr>
<tr>
<td>Mellert Volker</td>
<td>II 569</td>
</tr>
<tr>
<td>Meyer J.</td>
<td>III 463, II 573</td>
</tr>
<tr>
<td>Meyer-Falcke Andreas</td>
<td>II 199</td>
</tr>
<tr>
<td>Meynial X.</td>
<td>IV 233</td>
</tr>
<tr>
<td>Mikhalevsky Peter N.</td>
<td>I 281</td>
</tr>
<tr>
<td>Ming-Kun Zheng</td>
<td>II 705</td>
</tr>
<tr>
<td>Minami Yasuhiro</td>
<td>III 105</td>
</tr>
<tr>
<td>Miura Kenzo</td>
<td>IV 237</td>
</tr>
<tr>
<td>Miśkiewicz Andrzej</td>
<td>II 203</td>
</tr>
<tr>
<td>Molin Nils Erik</td>
<td>IV 123</td>
</tr>
<tr>
<td>Momertz Eckard</td>
<td>II 577, IV 127</td>
</tr>
<tr>
<td>Monay Gilbert</td>
<td>II 207</td>
</tr>
<tr>
<td>Moreno Antonio</td>
<td>II 211</td>
</tr>
<tr>
<td>Moreno Pedro J.</td>
<td>III 109</td>
</tr>
<tr>
<td>Morimoto Masayuki</td>
<td>II 581, 585</td>
</tr>
<tr>
<td>Motooka Seiichi</td>
<td>I 285</td>
</tr>
<tr>
<td>Musafir Richards E.</td>
<td>III 547</td>
</tr>
<tr>
<td>Munteanu Ligia</td>
<td>I 141</td>
</tr>
<tr>
<td>Nagle Steven F.</td>
<td>IV 241</td>
</tr>
<tr>
<td>Nakai Takayoshi</td>
<td>III 113</td>
</tr>
<tr>
<td>Nakamura Toshiaki</td>
<td>I 289</td>
</tr>
<tr>
<td>Namba Seiichiro</td>
<td>II 215</td>
</tr>
<tr>
<td>Narasimhan Sunil</td>
<td>I 293</td>
</tr>
<tr>
<td>Naylor Graham</td>
<td>II 589</td>
</tr>
<tr>
<td>Neveryov A. N.</td>
<td>I 393</td>
</tr>
<tr>
<td>Nickel Michael</td>
<td>III 359</td>
</tr>
<tr>
<td>Nicolas Jean</td>
<td>II 219</td>
</tr>
<tr>
<td>Niclsen Johan L.</td>
<td>II 593, 597</td>
</tr>
<tr>
<td>Niewiarowicz Marek</td>
<td>IV 245</td>
</tr>
<tr>
<td>Nihalani Paroo</td>
<td>III 117</td>
</tr>
<tr>
<td>Niimi Yasuhisa</td>
<td>III 121</td>
</tr>
<tr>
<td>Nishimura Ryuichi</td>
<td>III 125</td>
</tr>
<tr>
<td>Nomura Yasuo</td>
<td>IV 249</td>
</tr>
<tr>
<td>Nummela Sirpa</td>
<td>III 277</td>
</tr>
<tr>
<td>Nosek Jaroslav</td>
<td>I 397</td>
</tr>
</tbody>
</table>
Tachibana Hideki vol II 669, vol IV 167
Taguti Tomoyasu vol III 505
Takahashi D. vol II 673
Takane Shouichi vol I 177
Takano Takehiro vol I 181
Takeshima Hisashi vol III 297, 301
Tanaka Manabu vol I 185
Tchountchouzov Igor P. vol III 559
Tenenbaum Roberto A. vol II 677
Ternström Sten vol II 681, vol IV 319
Teuber Sabine vol II 685
Thoen J. vol I 189
Thomas R. L. vol I 453
Thomson Shirley J. vol II 307
Tokhi M. O. vol II 311, 315
Tomikawa Yoshiro vol IV 277
Tonning Andreas vol I 437
Torp Hans vol III 371
Toyoda Taro vol I 441
Tsujino Jojmaru vol I 445

Watanabe Akira vol III 185
Watanabe Kazumi vol IV 281
Watanabe Yuji vol I 459
Warusfel Olivier vol II 693, 697
Weiqing Zhu vol I 313, 317
Widholm Gregor vol III 517
Wollherr Horst vol II 335
Wong George S. K. vol I 205

Xiang Ning vol IV 175
Xia Zhang Yu vol III 375
Yamamoto Ken  vol IV 285
Yamamoto Teruji  vol III 309
Yamanouchi K.  vol I 463
Ylitalo Juha  vol III 379
Yoshinari Horinouchi  vol II 701
Ystad Sølv  vol III 525

Zhang Shu-yi  vol I 467
Zhao Yunxin  vol III 189
Zhang Renhe  vol I 309
Zhu Ye  vol I 321
Zindeluk Moysés  vol II 343
Zóltogórszki Bronislaw  vol IV 179

Öztürk Cüneyt  vol II 227, 231

Åbom Mats  vol III 563