BINAURAL HEARING AND SPATIALISATION
Computer simulation of Binaural, Stereo-Dipole, B-format and Ambiophonics impulse responses

A. Farina\textsuperscript{1}, L. Tronchin\textsuperscript{2}

\textsuperscript{1}Industrial Eng. Dept., University of Parma, 43100 Parma, Italy
\textsuperscript{2}DIENCA-CIARM, University of Bologna, 40136 Bologna, Italy

The paper addresses the problem of deriving synthetic impulse responses for auralization and spatialization systems, starting from the results of room-acoustics simulation programs. As it is well known, usually these numerical models are based on geometrical acoustics, and produce as output octave-band energetic impulse responses. These impulse responses have to be converted in wide-band waveform for being used as FIR filters, which can be convolved with dry signals, obtaining synthetic surround sound in various formats. The authors already developed in the past a suitable wide-band conversion algorithm for mono (omnidirectional receiver) impulse response and for binaural impulse responses.

In this paper, all the various possible “surround” formats are taken into account:
- Binaural, for headphone listening
- Stereo Dipole, for transaural listening on one or two pairs of closely-spaced loudspeakers
- B-format, for reproduction through a three-dimensional Ambisonics decoder and loudspeaker array (1\textsuperscript{st} and 2\textsuperscript{nd} order).
- Ambiophonics, which couples a frontal Stereo-Dipole for the direct sound, and an Ambisonics-like surround array for the subsequent reflections and reverberant tail.

The paper explains the mathematics of the conversion from the energetic impulse response to the multichannel waveforms in each of the above formats.

ROOM ACOUSTICS SIMULATION

This paper describes the post processing of the results of a well known room acoustics simulation program [1].

As it is common to many others similar programs based on geometrical acoustics, the computation produces in each receiver an energetic impulse response, computed in several octave bands. In this case, in each receiver the impulse response has a constant time resolution (typically 1 to 10 ms) and 10 octave bands (31.5 to 16000 Hz). The value recorded in each cell of this data structure is the sum of the energy density carried by all the wavefronts arriving to the receiver during the corresponding time slot: the hypothesis is thus that these wavefronts do not interfere, as each of them is carrying a signal which is uncorrelated with the others. It is obvious that this hypothesis has some intrinsic inconsistency for low-order reflections and for low frequencies, whilst instead is reasonable for higher frequencies and in the late part of the reverberant tail.

Furthermore, for the lower order reflections, also the direction of provenience, the amplitude and the exact arrival time of each wavefront are saved in a separate file: this makes it possible to plot the paths of each energy arrival, and to derive realistic pressure impulse responses, in a variety of multi-channel formats corresponding to different types of stereophonic microphones.

In the following it is briefly described how to translate the energetic results of the simulation to the virtual signal, which would be recorded by different types of microphones placed at the receiver positions.

VIRTUAL MICROPHONES

The following types of stereophonic microphones can be virtually placed at the receiver position:
- omnidirectional pressure microphone: for comparison with measured impulse responses
- binaural 2-channels microphone: for auralization by means of headphones.
- 1\textsuperscript{st} order spherical harmonics of pressure field (B-format Soundfield microphone, 4 channels): for auralization on an Ambisonics 1\textsuperscript{st} order decoder.
- 2\textsuperscript{nd} order spherical harmonics of pressure field (Furse-Malham microphone, 9 channels): for auralization on an Ambisonics 2\textsuperscript{nd} order decoder.
- Single Stereo-Dipole “transaural” detector (binaural processed with cross-talk canceling filters): for auralization on a pair of closely-spaced loudspeakers
- Dual Stereo-Dipole “transaural” detector (4-channels, separate frontal and rear soundfields): for auralization on two pairs of closely-spaced loudspeakers
- Ambiophone (pinnless dummy head for the first two channels, plus a 4-channels 1\textsuperscript{st} order B-format for the surround): for auralization on an Ambiophonics playback system.
COMPUTATION OF THE IMPULSE RESPONSES

For each type of microphone, the conversion from a 10-bands energy impulse response to a multichannel, wideband pressure impulse response is done employing the same approach: first the early reflections are processed, taking into account their known arrival direction and exact timing. For each energy arrival, a Dirac’s delta function is generated at the exact arrival time, then it is convolved with the impulse response of an octave-band equalizer which imposes the proper 10-octaves spectrum, and finally it is convolved with the multichannel impulse response of the selected microphone, chosen depending on the direction of arrival.

Then the subsequent reverberant tail is added, based on the whole energetic impulse response data (which do not contain any directional information), assuming valid the hypothesis of a completely isotropic diffuse field, with arrival directions randomly chosen from all the angles. This is done by first generating an independent sample of white noise for each channel of the virtual microphone, and then applying to it a time-dependent octave-band equalization, given by the energy-time-frequency information stored as a result of the room acoustical computation; in more detail, the white noise is first octave-band filtered, then it is amplitude-modulated with the square root of the temporal hystogram of received energy in that frequency band; finally the processed signals of all the octaves are summed back together.

Regarding the impulse response of each directive microphone, it must be noted that for the omnidirectional and for the first and second order Ambisonics these are simply delta functions of proper amplitude and sign, obtained from the director cosines of the incoming ray ($r_x$, $r_y$, $r_z$). The following table contains the formulas which give these gains:

<table>
<thead>
<tr>
<th>Channel</th>
<th>formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>0.707107</td>
</tr>
<tr>
<td>X</td>
<td>$r_x$</td>
</tr>
<tr>
<td>Y</td>
<td>$r_y$</td>
</tr>
<tr>
<td>Z</td>
<td>$r_z$</td>
</tr>
<tr>
<td>R</td>
<td>$1.5 \cdot r_z^2 \cdot 0.5$</td>
</tr>
<tr>
<td>S</td>
<td>$2 \cdot r_x \cdot r_y$</td>
</tr>
<tr>
<td>T</td>
<td>$2 \cdot r_x \cdot r_z^0.5$</td>
</tr>
<tr>
<td>U</td>
<td>$r_x^2 \cdot r_y^2$</td>
</tr>
<tr>
<td>V</td>
<td>$2 \cdot r_x \cdot r_y$</td>
</tr>
</tbody>
</table>

For the binaural and Stereo Dipole method, instead, the impulse responses employed come form the measurements made at MIT on the Kemar dummy head [3]. In the case of loudspeaker reproduction (Stereo Dipole, dual Stereo Dipole) they must be also convolved with the set of proper cross-talk canceling filters, which are derived by inversion with the Kirkeby/Farina approach the pair of binaural impulse responses correspondent to the position of the loudspeakers, as explained in detail in [4].

For Ambiophonics, a similar data-base of theoretical impulse responses was computed analytically, considering an ideal rigid sphere of 180 mm diameter with pressure microphones exactly at the opposite sides. Only the direct sound and the first reflections coming from the frontal hemispace are sent to the sphere microphone, which drives the frontal Stereo Dipole (again by means of cross-talk canceling filters, derived by the impulse responses of the theoretical sphere). The subsequent lateral and rear reflections and the reverberant tail are instead sent to the B-format microphone, which drives the surround loudspeakers. The working principle and some implementation details related to the Ambiophonics reproduction system are detailed in [5].

ACKNOWLEDGMENTS

This work was financially supported by the Ambiophonics Institute (www.ambiophonics.org).

REFERENCES

Ambiophonics, Achieving Psychoacoustic Realism in Music Recording and Reproduction

R. Glasgal

Ambiophonics Institute, 4 Piermont Road, Rockleigh, New Jersey, 07647, www.ambiophonics.org

Ambiophonics, a binaural technology, is the most logical successor, to both stereophonics and its similar 5.1 surround sound brethren; where the recording and reproduction of serious live music is concerned. We review how binaural technology, combined with a clear understanding of what the ear requires and what the concert hall provides, yields a practical alternative to the common recording and playback technologies for music, i.e. stereo, 5.1, 6.0, 7.1, or 10.2 surround. We summarize how Ambiophonic technology can consistently generate a "you are there" concert, opera, or pop sound field that the ear-brain system easily accepts as excitingly realistic (not artistic verity but physiological versimilitude), even (as luck would have it) not only from optimized recordings made using an Ambiophone but also from standard two media-channel recordings such as the existing library of LPs, CDs, 6.0 DVD-As, DVD-Vs, or SACDs. Further text, papers, figures, and references are at the above web site.

INTRODUCTION

The goal of Ambiophonics is to deliver to one or two home listeners a realistic replica of the concert hall experience. Ambiophonics, a public domain technology supported by the non-commercial Ambiophonics Institute, combines an exploitation of under-appreciated psychoacoustic principles with the basic rules of good musical performance-space design to create believable concert-hall soundfields in typical home listening rooms. Ambiophonics moves the listener into the same space as the performers by accommodating to the uniqueness of individual pinna characteristics, minimizing ambient interaural correlation at the listening area, abandoning the traditional stereo loudspeaker triangle (including the center speaker in 5.1), generating early reflections and reverberant fields from stored real-hall impulse responses, eliminating comb filtering due to front-loudspeaker crosstalk, and using room correction/treatment technology to insure that the listening room does not impair the illusion.

Ambiophonics can provide exceptionally realistic reproduction of 2-channel music sources because such recordings themselves are not inherently Blumlein stereophonic. That is, they do not contain the crosstalk rays or angle distortions that will be produced by the loudspeakers of the usual stereo triangle and they do not know that their sound sources will not be localized binaurally but rather by using the phantom image hearing mechanism of the antiquated stereo triangle.

The Ambiophonic paradigm includes the Ambiopole, a comb-filtering-free front- speaker pair, real hall impulse convolution, (the process by which even a studio recording can seem to be sounding in one of the worlds great concert halls) Surrstats, (a surround speaker design that facilitates the natural reproduction of hall ambience) room correction to ensure that listening room anomalies do not distort the Ambiophonic sound field, and the Ambiophone, a main microphone arrangement conceived to make new two-channel surround music recordings that are optimum for Ambiophonic playback.

PHYSIOLOGICAL VERSIMILITUDE

In this paper, realism in the reproduction of classical and pop live music and virtual reality compositions is understood to mean the generation of a sound field realistic enough to satisfy any normal ear-brain system that it is in the same space as the performers, that this is a space that could physically exist, that the sound sources in this space are easy to locate and that the ambience is as full bodied as at a live concert.

It is important that localization be as effortless and as natural as in everyday hearing rather than be exaggerated by spot mic's and pan processing. The binaural recording/reproduction ideal states that if you deliver to the entrance of the ear canals precisely what those ear canals would have sensed had they been in a good seat at the live event then physiological realism is guaranteed. I believe that two-channel Ambiophonics does meet this exacting binaural standard for the case where there are no direct sound sources in the rear half of the horizontal plane and no direct sound sources above or below it.

FRONTAL PSYCHOACOUSTIC ELEMENTS

It is well known that human beings localize by using both their pinnae and their head shadow. In this paper we will define the term HRTF to include head/torso shadowing only. Pinna functionality will generally refer to the ability of humans to localize using just one ear. HRTFs are responsible for all interaural differences in phase/time and level and
thus the localization that can be inferred from such interaural time and level differences. By contrast, each pinna is independently sensitive to direction and they provide a higher frequency localization system, dependent on interaural differences only to the extent that one pinna can (maybe) confirm the localization sensed by the other one. Thus we can state the first law of Ambiophonics. In any recording and reproduction chain there shall be only one set of pinnae and those pinnae must be those of the home listener. The second law of Ambiophonics is that there shall be at least one but only one head shadow (HRTF) in a recording/reproduction chain. It is desirable but not essential that this HRTF be that of the home listener.

**AMBIPOLES**

Let us assume that we have two super directional, ideal, point source loudspeakers and place them head spaced directly in front of a listener. Then the sound from the left loudspeaker only impinges on the left ear and the sound from the right loudspeaker goes only to the right ear at least over the frequency range of interaural and pinna interest. In this configuration, little sound goes past the nose to reach the wrong ear and, for central stage sources, the incident angles to both pinnae are correct. If the signals to the speakers have been recorded either as described below or are reasonably unprocessed, then each ear hears what the recording microphone (or the mix) heard and phantom imaging of the central stage is not needed. There is also no possibility of comb filtering since we postulate that the speakers are focused on the ears in the comb-filtering region. Even if there were some crosstalk, the path length difference is so small with the speakers so close together, that the combing onset moves up in frequency and is much less audible. I call such speaker arrangements Ambiopoles. Ambiopoles that come close to the ideal can be created in software. Software Ambiopoles work by generating delayed, opposite polarity crosstalk cancelling signals in an infinite series.

**AMBIOPHONES**

Before we end our consideration of the realistic reproduction of the front stage, we must consider whether there is a recording method that is ideal for Ambiophonic reproduction via the Ambiopole and the ambience surround speakers described below. Indeed there is. In an ideal recording and reproduction chain, we have indicated that the rule is that there should be only one set of pinna (your own) and one HRTF. If one takes a head shaped sphere and places two omni directional microphones on the surface where the ears would be one has introduced a reasonable replica of a head shadow into the front stage recording chain. Remember that this is necessary because in the reproduction of the front stage left and right channels the Ambiopole speakers are directly in front and so there is no head shadow for stage sounds coming from the sides. A head-size sphere or head shaped microphone without pinna can perform this function. If you simply place this microphone fifth row center you will record and then reproduce the same perspective that a listener at that seat would have had. Normally, one cannot do this because most microphones would pick up more ambience than direct sound and this ambience coming from the front speakers would make the musicians seem to be performing in a sewer. Thus the Ambiphone is shielded to the rear, the extreme sides, and overhead by sound absorbent material.

**THE AMBIVOLOVER**

If one looks at the literature as a whole, one must conclude that any preconceived notion as to the number, placement, level, and reflection parameters used to fabricate a hall in the home listening room will prove incorrect. The proper method is not to assume but to measure the impulse responses of the worlds great halls and then allow the domestic concert hall denizens to select that hall that complements the music they are about to listen to and that suits the room and the number of speakers they can afford and have room to site.

The impulse responses of halls can now be easily measured before, during or after a recording session and included with the CD or stored on CD-ROM or placed on a web site. The process of applying the hall impulse response to the left and right stage signals to recover the early reflections and late reverberation tails and feed these signals to the appropriate, directionally HRTF correct loudspeakers is called convolution. We call the computing device that does this in real time an Ambivoixer. In actuality, an Ambivoixer is an all purpose device that can convolve the left and right channels to calculate the signals for the front Ambiophone, to calculate any speaker or room corrections needed, and to compute all the hall reflections for as many surround speakers as you like and have the processing power to support.

**CONCLUSION**

Ambiophonics has been shown to be able to reproduce new or existing two channel recordings with exceptional realism. Comparative listening tests performed by the University of Parma confirm the superiority of Ambiophonics for concert-hall music over Ambisonics, stereo, or multi-channel surround.
Monaural and binaural processing for automatic estimation of room acoustics perceptual attributes

A. Baskind and O. Warusfel

IRCAM – Room Acoustics team, 1, place Igor Stravinsky, 75004 Paris, France
e-mails : Alexis.Baskind@ircam.fr, Olivier.Warusfel@ircam.fr

Monaural and binaural auditory modeling is more and more used in practical applications, such as pitch estimation, source localization, or "cocktail-party" processors. In these applications, the room effect is often regarded as a disturbance, and rarely as the main issue of the analysis. The present work consists in finding methods related to monaural and binaural processing to estimate the room acoustics perceptual attributes of auditory scenes from the knowledge of monophonic or multichannel recordings, without extra information, such as a geometrical description or a set of impulse responses. Beyond the interest for a better understanding of human spatial hearing, foreseen applications are automatic labeling, reproduction of a room effect and tools for mixing dry sources in a pre-recorded soundscape.

INTRODUCTION

Most of research in auditory scene analysis is focused on the different sources that compose the scene, providing methods to estimate their own low-level cues, such as pitch or localization, or to provide a more high-level description in an attempt to isolate and recognize each of them. From this point of view, room effect is needed to be as weak as possible : not only it is not the topic of interest, but it also disrupts the estimation. Of course, many room acoustics characterization techniques have been developed, for monaural or binaural cues [1], but they lean on impulse responses, not on sound scene recordings.

Our goal is to find ways to estimate such cues from complex recordings, without extra knowledge neither on the room nor on the original (dry) sound. There are several manners of describing the spatial cues of a sound scene; we decided to focus on perceptual cues, mainly because one of the main goals of this research is to provide tools to characterize the room effect of a sound scene, in an attempt to reproduce it, or to mix a dry source in the scene; from this point of view, most of the information provided by a signal description (i.e. a complete impulse response recording) is useless.

ARCHITECTURE OVERVIEW

The architecture of the system (see Fig. 1) is inspired from binaural modeling: front-end peripheral filtering imitates the behavior of the basilar membrane; the physiological relevance of coincidence processors is more questionable, but they succeed in explaining many perceptual phenomena, like localization, or the effect of left-right correlation on apparent source width. Thus the system is composed of several steps :

- **linear and passive peripheral filtering** : the filter bank used here is based on Patterson’s “Gammatone” model, and uses 8th-order constant-Q IIR filters distributed according to a logarithmic frequency scale.
- **monaural analysis and cue detection** : the aim of this stage is double : estimating monaural cues (such as reverberancy, diffuse-field spectral color, perceptual distance) linked to a time-frequency description of the room effect [2,3], and enhance this room effect as much as possible, in order to make the work of the next stage easier.
- **binaural coincidence processor** : this stage receives inputs from left and right, and looks for correlations between them. It returns running coincidence patterns for each frequency channel, as a function of time lag and, possibly, gain between both sides.
- **pattern evaluation** : the evaluation consists in finding extrema in patterns that are common to all frequency channels, in order to localize reflections in time and space, and thus estimate binaural cues (such as localization, spatial impression, envelopment).

**Binaural coincidence processor**

The core of this stage is short-term correlation, which has been proposed by Jeffress [4] to measure similarity between one side and a delayed version of the other side (see Fig. 2). An alternative algorithm would be Durlach Equalization-Cancellation model [5] which consists in finding the delay and gain to apply to one side to minimize the difference between both sides. Those models are equivalent, but the latter offers a extra degree of freedom (gain) and seems to be more precise.
Whatever model we choose, it provides, when it is used with binaural impulse responses, a spatial map of the direct sound and the reflections, which allows to localize the source, and to estimate envelopment and spatial impression. However, when it is applied to a complex binaural recording, the reflections are not as well isolated, and the patterns become unreadable. That is why the previous stage tries to provide an impulse response as clean as possible.

\[ y_i = h * x_i \]

\[ y_N = h * x_N \]

with \( \hat{e} = \frac{1}{N} \sum_{k=1}^{N} \hat{y}_k \) \]

In practice, several problems occur: first of all, the all-pass part of the impulse response is much harder to estimate than the minimum-phase part; this has to be overcome, since the all-pass part contains the most relevant information on early reflections; moreover, the relative level of each frequency channel cannot be properly estimated by this only method, and some synchronization problems appear when resynthesizing a broad-band impulse response; finally, it needs an efficient prior time segmentation, which can be driven thanks to a separate onset detector. This low-level analysis module has also to be completed with monaural cues detectors, such as a reverberation time estimation device.

**CONCLUSION**

The system proposed is focused on an original feature in auditory scene analysis, which is the room effect. The overall architecture is inspired from usual binaural models, but less traditional signal processing techniques, such as cepstral processing and minimum-phase / all-pass decomposition, were included. At present, the evaluation stage, which estimates high-level cues from low-level information, has not been implemented. For this purpose, we need a paradigm for room responses, which will also be useful in recovering lost information from the all-pass component of the signal.

**REFERENCES**

The Equalisation of a Microphone Array for Surround Sound Recording

M. A. Poletti

Industrial Research Limited, PO Box 31-310, Lower Hutt, New Zealand

A novel microphone array has been developed for recording second-order and higher spatial harmonics of sound fields. This paper introduces two methods for equalising the array to produce a constant response magnitude with frequency.

INTRODUCTION

A new method of recording sound fields has been recently developed [1,2]. The recording system comprises a circular array of N directional microphones facing outwards from the center of the array. The outputs of the microphones are sampled in time, and at each sample period, the FFT of the N output samples is taken. This process calculates the azimuthal phase modes of the sound field, which may be combined to form \( \sin(m\theta) \) and \( \cos(m\theta) \) amplitude modes. If the number of array elements is sufficient, and the array radius \( r \) small enough, the array responses are close to the continuous Fourier integral responses

\[
z_{m,\nu}(t) = A e^{i\alpha k_0 t} \left[ J_m(k_0 r) - J'_m(k_0 r) \right] e^{j m \phi} = A e^{i\alpha k_0 t} \left[ J_m(k_0 r) - J'_m(k_0 r) \right] e^{j m \phi}
\]

where \( k_0 = \frac{2\pi}{c} f_0 \), \( A \) is the plane wave amplitude, \( J_m(z) \) is the \( m \)th Bessel function of integer order of the first kind, and the polar response of each microphone element – pointing in direction \( \theta_n = \frac{2\pi (n-1)}{N} \) – is

\[
p_n(\theta) = \alpha + (1-\alpha) \cos(\theta - \theta_n)
\]

The array response must be equalised in order to produce a flat response with frequency for the \( m \)th phase mode. For \( \alpha = 0.5 \) (cardioid) the ideal response magnitude has no spectral zeroes and can be equalised without producing singularities.

Two equalisation methods have been developed based on the ideal array spatial impulse response which is the inverse Fourier transform of equation 1 [1]. The impulse response is of finite duration, corresponding to the time a plane wave delta function takes to propagate over the array [1].

POLE-ZERO MODEL EQUALISATION

In the first approach, the impulse response was sampled and pole-zero modelled using the Stieglin-McBride algorithm in MATLAB. The inverse filter was then determined by converting poles to zeros and minimum phase zeros to poles. Maximum phase zeros were converted to minimum phase zeros and inverted (in practice these zeros were close to the unit circle). The transfer functions were then obtained using an FFT and compared with the theoretical response in equation 1. Seven zeros and six poles were found to produce good results for 1st to 3rd order responses.

The impulse response and transfer function for \( m=2 \) are shown in figure 1. The model is quite accurate, although it diverges slightly at high frequencies.

![Figure 1](image)

FIGURE 1: a: Sampled impulse response (solid) and modelled response (dashed), \( m=2 \), b: Transfer function of sampled impulse response (solid), of modelled impulse response (dashed), and theoretical transfer functions (dash dotted)

The model produces a single zero near zero Hz which is required to model the first order rise with frequency that the transfer function in figure 1b demonstrates. The zero was maximum phase with radius 1.001, and was converted to a minimum phase zero for the equaliser.

A single pole shelf filter was included with shelf frequency 13 kHz and 1.5 dB boost to improve the high frequency response.
The accuracy of the equaliser was determined by adding the equaliser response in dB and the ideal array response in equation 1. The sum response is shown in figure 2.

The maximum deviation is less than 0.3 dB across the frequency range.

**LEAST SQUARES-BASED EQUALISATION**

An alternative approach to the equalisation of the microphone array is based on the power series expansion of the two Bessel function terms in equation 1. Combining these, the expansion of the array response is

$$H(kr) = \frac{1}{2} \sqrt{2} \left[ \frac{(-1)^m}{\sqrt{m+1}} \right] \sum_{n=0}^{m-1} (1 + \frac{a_k}{2}) \left[ \frac{k_0}{2} \right]^{2(n+1)}$$

The $m$th order response thus has an $(m-1)$th order differentiator term $(jk_0r/2)^{m-1}$ which produces a $20(m-1)$ dB-per-decade rise in frequency at low frequencies. At high frequencies the sum modifies this behaviour and the response peaks and then reduces.

An inverse equaliser may be constructed which contains an $(m-1)$th order integrator which compensates for the differentiator. A least squares inverse FIR filter may then be designed to compensate for the remaining frequency response function. The integrator may also be made to roll off at low frequency to avoid high gains. In this case it becomes an $(m-1)$th order low pass filter.

The inverse least squares filter for $m=2$, using 14 taps including a modelling delay of 4 taps, and the combined least squares and integrator response, are shown in figure 3.

A one pole integrator with $2^{nd}$ order Fir compensator was used, with an extra shelf filter (1.4 dB boost above 4 kHz) to improve high frequency accuracy. The sum response is shown below.

**REFERENCES**

Individualization of HRTF by Spectral Warping

V. Larcher\textsuperscript{a}, J. M. Jot\textsuperscript{b} and O. Warusfel\textsuperscript{c}

\textsuperscript{a}GENESIS, CEEI, Domaine du petit arbois,13545 Aix-en-Provence cedex 04, France
\textsuperscript{b}CREATIVE ATC, 1600 Green Hills Road, P.O. box 660015, Scotts Valley, CA 95067, USA
\textsuperscript{c}IRCAM, 1 place Igor Stravinsky, 75004 Paris, France

Head-Related Transfer Functions (HRTF) vary upon individuals due to the specificity of one's head/ear/torso shapes. Perceptual artefacts entailed by the use of non-individual HRTF can be significantly reduced by scaling in frequency the non individual HRTF to best fit the listener's ones. This method, which was proposed by Middlebrooks ([1], [2]), consists in translating one head's spectra on the log-frequency axis to match the features of the second head's HRTF. As an extension, we first show that the efficiency of the scaling is improved when applied independently to two frequency bands above 1kHz. In order to automatically set the scaling factors, correlation to the dimension of morphological features is studied. In order to further simplify the evaluation of the scaling factors, it is shown that a satisfying estimate can be obtained that relies on a limited number of HRTF. They correspond to the "principal directions" given by a statistical analysis of HRTF.

IMPLEMENTATION OF BINAURAL SYNTHESIS

The “brute force” implementation of binaural synthesis involves fractional delays which represent interaural time delay (ITD), in series with minimum phase filters corresponding to the left and right HRTF for the desired position. When multiple sources have to be spatialized, a multi-channel implementation can be used to reduce the computing cost. It is based on the linear decomposition of minimum-phase HRTF into spatial functions \( C_i(\theta) \) and reconstruction filters \( L_i(f) \):

\[
\text{HRTF} = \sum_i C_i(\theta) \cdot L_i(f) \quad (1)
\]

Several authors have proposed and compared methods to derive optimal \( C_i(\theta) \) and \( L_i(f) \), among which statistical analysis of minimum-phase HRTF ([3]). More specifically, by applying Independent Component Analysis (ICA), we have shown that \( C_i(\theta) \) can be defined as non-individual patterns with high-compacity support ([4], [5]). They point towards “principal directions”, and lead to \( L_i(f) \) close to HRTF measured at these directions.

MEASURE OF INTER-INDIVIDUAL DIFFERENCES

The two preceding implementations entail individual differences, within the ITD component as well as within the minimum-phase filters. On our 17 heads, these differences can lead to ITD mismatch as large as 100 microseconds and to an average distance between magnitude spectra that reaches 7dB above 8kHz ([5]). Perceptual tests have shown that these differences influence localization accuracy, and must therefore be taken into account. If a generic database has to be used by all listeners, this implies that these HRTF have to be chosen appropriately so as to realize an optimal trade-off between listeners. A first level of individual adaptation consists in proposing a discrete set of HRTF databases among which each listener has to pick up the most appropriate. In the following, we address another strategy, named “continuous adaptation”, which aims at synthesizing the listener’s HRTF knowing the individual setting of a few parameters. This setting can be made by the listener himself, typically by adjusting a few sliders, or can even be automatic if the parameters are linked to morphological features that a camera or other sensors may capture.

CONTINUOUS ADAPTATION OF ITD

ITD can be approximated using a spherical head model as a function of azimuth \( \phi \), elevation \( \theta \) by ([4]):

\[
\text{ITD} \approx \frac{r}{c} \times \left[ \cos(\theta) \sin(\phi) + \sin(\cos(\theta) \sin(\phi)) \right] \quad (2)
\]

If HRTF are known, the parameter \( r \), also called “radius”, can be estimated by minimizing the mean squared error over all positions. If HRTF have not been measured, eq. (2) can still provide a satisfying estimate of ITD if the radius is approximated by a weighted sum of the head’s dimensions, which gives with our data:

\[
r \approx 0.66 \text{width} + 0.11 \text{depth} + 3.33 \text{ cm} \quad (3)
\]
SPECTRAL WARping OF MINIMUM-PHASE HRTF

In [1], Middlebrooks proposes to transform one head into another by using spectral warping (see Figure 1). The warping coefficient (or “scaling factor”) is common to all positions, and is derived by minimizing a global distance between heads. This distance is defined as a mean of the squared difference between the log-magnitude spectra of two heads, from 3 to 13kHz.

![Figure 1](image)

**FIGURE 1.** Magnitude spectra of two heads: principle of individual adaptation by spectral warping.

In order to optimize the morphing procedure, we have introduced an interval-wise warping, splitting the full frequency band into a low-frequency band (frequencies below 1kHz), a mid-frequency band (from 1 to 5kHz), and a high-frequency band (above 5kHz). The three scaling factors obtained prove to be poorly correlated to each others, which shows that applying an unique scaling factor is sub-optimal. Warping of the low-frequency part does not lead to a fair reduction of spectral distance and should therefore be discarded. This may be explained by the flat characteristics of HRTF spectra on this interval. The largest reduction is observed for the high-frequency interval, for which the reduction reaches 30% for more than one third of the heads. Besides, the high-frequency scaling factor is highly correlated to concha height (see e.g. [1] for references on measurement of morphological features). The mid-frequency warping reduces significantly distance between heads, but no consistent link to morphology could be found.

Although not fully automatic, an estimation of warping parameters based on a limited number of HRTF would already be a practical solution. Matching a generic HRTF database to the listener’s HRTF would thereby require only a few measurements. For the high-frequency interval, again, a strong relationship can be found between the global warping coefficient and the estimate given from the “principal directions”. By construction, this estimate is also optimal for a multi-channel implementation involving the HRTF measured at these positions.

CONCLUSIONS AND PERSPECTIVES

Frequency warping is an efficient approach to transform one head into another, and consequently to adapt a generic set of HRTF to each listener. Our study has shown that it may be applied separately to mid- and high- frequencies, and that such transform is inappropriate at low frequencies. An additional parameter, the “radius”, has to be taken into account to control an analytical approximation of ITD. No robust relationship has been found between those three adaptation parameters.

To reduce the remaining inter-head distance, other transforms should be considered. Besides, although the perceptual improvement for localization has been investigated by Middlebrooks in the case of global warping ([2]), further studies should be undertaken in order to compare it to the improvement already provided by head-tracking and/or by the rendering of early reflections.

ACKNOWLEDGMENTS

The authors wish to thank V. Ralph Algazi and Richard O. Duda for making their HRTF data and morphological measurements available for this study.

REFERENCES

Spectral optimization of listening conditions in car cabins

G. Vandernoot\textsuperscript{a,b}, O. Warusfel\textsuperscript{b}

\textsuperscript{a} Renault Technocentre – 1, avenue du golf – 78288 Guyancourt – France
\textsuperscript{b} Institut de Recherche et de Coordination Acoustique et Musique (IRCAM Centre Georges Pompidou)
1, place Igor Stravinsky – 75004 Paris – France

Design of stereo reproduction system for cars is very specific because of cabin acoustical response and non-symmetrical loudspeakers positions. This behavior has motivated various researches aiming at improving listening conditions with the help of signal processing. Based on time-frequency analysis, a new spectral equalization is presented, considering independently the direct sound and the diffuse field spectra. This method allows an efficient correction based on spatial characteristics of the sound field. A perceptive justification was undertaken thanks to listening tests analyzed with multidimensional tools.

**INTRODUCTION**

Design of stereo systems in cars is very specific. The car cabin is a very small and damped space. The sound emitted by loudspeakers is strongly modified by the small volume of the compartment. Hence, an electronic equalization is highly recommended.

Time-frequency analysis of car cabin impulse response shows that the time support is less than 100 ms, and that the reverberation time decreases in low frequency down to a constant value of 50 ms above 800 Hz. At the risk of abusing this specific term, we will employ the word “reverberation” despite the shortness of the phenomenon. One has to carefully consider which strategy to adopt to conduct equalization. For classical room acoustics, it is generally considered that contributions of direct sound and of reverberation play independent roles, linked to the temporal distribution of reflections (early/late) and their spatial distribution (localized/diffuse). Due to temporal masking and integration phenomena, the question arises if this distinction is still valid in the case of a very short time support such as car cabin response. It might seem useless do distinguish the frequency response linked to direct sound from that of cabin effect. In order to answer this question, a double-blind listening test was carried out with synthetic impulse responses where the spectral contents of direct sound and cabin effect could be precisely and independently controlled.

**EXPERIMENT**

From a mean response that simulates the main characteristics of sound restitution in a car cabin (temporal and modal density, localization), we varied the time-frequency-energy distribution, i.e. the relative frequency contents of the reverberated field and the direct sound. The different considered configurations are shown on the figure 1. The curves represent the spectral content of the direct sound and cabin effect contributions. The frequency response was controlled in three frequency bands thanks to double-shelving filters. Crossover frequency between LF, MF and HF were 250 Hz and 8000 Hz.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Direct sound</td>
<td>Reverberation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Direct sound</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverberation</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>+6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>-3 +3 +3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>-6 -6</td>
</tr>
</tbody>
</table>

**FIGURE 1.** Energy levels in three frequency bands (low, mid, high)

It can be noticed that the energy ratio between the direct sound and the reverberated field is constant within each category A or B (0 dB at mid frequencies). On the other hand, the total energy, i.e. the sum of the direct and reverberated contributions, is constant for each configuration 1, 2 or 3. Subjective dissimilarity judgements between these 6 configurations (i.e. 15 pairs) were collected.

**MULTIDIMENSIONAL ANALYSIS**

Analysis was conducted with INDSCAL method that allows representing in a common space the different stimuli and individual weighting of the different
dimensions. The number of dimensions is traditionally determined, a posteriori, thanks to two parameters measuring the concordance of the model with the initial data. In our case the limited number of stimuli does not allow to consider more than 2 dimensions. The corresponding map explains 65% of the variance.

**FIGURE 2.** Group stimulus space for the first test (diffuse reverberation) and the second test (localized reverberation), given by INDSCAL analysis with two dimensions.

From raw data analysis together with INDSCAL internal parameter (additive constant), we could derive an estimation of the just noticeable difference. It is schematized, on figure 2, by a circle centered on stimulus 1A. The interpretation of the test can also be conducted from correlation between the perceptual dimension provided by INDSCAL and the objective characterization of the stimuli. Vectors superimposed in the stimuli space represent these correlations. In the following analysis, we single out the spectral modifications affecting either the whole response or the ratio between the direct sound and the reverberation.

**Global modifications of the spectrum** The pairs parallel to the tot vector represent the same perceptual effect, and relate to the global spectral content modification. Compared to the just noticeable threshold, only the pairs \{1A;3A\} and \{1B;3B\} can be easily discriminated.

**Swapping between direct sound and reverberation** Pairs parallel to the dir/rev vector show distances longer than the model’s threshold, and are all discriminated. The tot and dir/rev correlation vectors are orthogonal to each other. It can be concluded that the direct sound and “reverberated” sound played independent role in the dissimilarity judgments.

**Complex modifications** For the pairs \{1A;2B\} and \{2A;3B\} the modification of the energy coming from direct sound is even more important, although it is partly compensated by that of the reverberated field, while for the pairs \{2A;1B\} and \{3A;2B\} it is the reverse. The distances are greater when the spectral modification affects mainly direct sound. For the pairs \{1B;3A\} and \{1A;3B\}, the energetic modification is uniquely targeted on the direct sound or the reverberation. These pairs were judged to be very different.

Although the perception of the frequency content linked to direct sound is predominant, the analysis shows two independent perceptive factors linked to objective indices dealing with direct sound and cabin effect. So, despite the very short length of time of the cabin effect, our perception can distinguish the time-frequency distribution of energy. Indeed, by imposing an identical angle of incidence for direct sound and cabin effect, the repetition of this test shows that the subjects are no longer able to distinguish the frequency modifications carried out on one or the other of those two entities. So we can conclude that this distinction relies on the spatial distribution, but it doesn’t rely on the temporal effect.

**VEHICLE EQUALIZATION**

According to the results of the sound test, it is desirable to equalize separately direct sound and cabin effect. The spectral densities of the energy of the direct sound and that of the cabin effect are derived from the time-frequency analysis of the impulse responses. The corrections envisioned here are minimum phase equalizations, i.e. they only concern the corresponding transfer function magnitude. The inverse filters are implemented under IIR form.

**FIGURE 3.** Inversion by 16th order IIR

Considering the direct sound to come exclusively from front, its equalization only concerns the front loudspeakers. Taking into account for this first correction, the rear loudspeakers will then provide an added degree of freedom in order to perform the overall cabin reverberation equalization.

This method was applied on a vehicle, and we managed to achieve an instant comparison between the original and the corrected audio systems. Although informal, listening tests showed satisfying quality.
Early Reflection Thresholds for Virtual Acoustic Sound Field Simulation

Durand R. Begault\textsuperscript{a}, Bryan U. McClain\textsuperscript{b}

\textsuperscript{a}Human Factors Research & Technology Division, \textsuperscript{b}San José State University Foundation, \textsuperscript{a, b}Spatial Auditory Display Laboratory, MS 262-2, NASA Ames Research Center, Moffett Field, California 94035-1000 USA; dbegault@mail.arc.nasa.gov

Data on auditory thresholds for acoustic reflections were obtained for seven subjects as a function of both spatial position and time delay in a simulated 5.1 surround sound listening environment. Absolute thresholds (perception of any type of change) were measured at the 70.7\% level using a one up-two down staircase algorithm, for both anechoic and reverberant conditions. Reflection threshold data are useful in the context of building acoustics, since path length attenuation and absorption can make potential reflections inaudible. Audibility of reflections is desirable for 3-D sound headphone simulations that require sound source externalization. The information is also useful for determining engineering parameters for the real-time simulation of virtual acoustic environments, such as helmet-mounted displays that include head tracking.

INTRODUCTION

Data on auditory thresholds for early reflections have been reported by various workers using real sound sources \cite{1, 2}. The following study uses an ‘auralization’ technique for simulating direct and reflected sources corresponding to a 5.1 listening room configuration. The correspondence between real and virtual sound source thresholds allows an estimate of an auralization program’s capacity to predict perceptual responses to more complex room models for both psychoacoustic investigations and sound quality evaluation. Establishment of thresholds for early reflections is pertinent to determining necessary absorptive treatment for building acoustic treatment. Another goal previously described in \cite{3} is for management of computational resources for real-time auralization.

METHODOLOGY, SUBJECTS

Absolute thresholds were determined for time-delayed speech signals, relative to an undelayed version of the same signal corresponding to an acoustic “direct path”. The delayed signals, corresponding to acoustic “reflections” within an enclosure, were manipulated in terms of both time delay and location between experimental blocks; the level of the reflection was manipulated as the dependent variable. Stimuli were formed from one of 36 randomly chosen anechoic speech segments 1.3 s in duration \cite{4}. Azimuth-elevation angles (referenced to 0° at a point directly in front of the listener) were simulated via real-time head-read transfer function (HRTF)-filtering.

Seven subjects evaluated for normal hearing were run. Stimuli were presented in double-wall sound-booth having a background noise level of 15 dB (A-weighted). Stimuli were presented at a level of 65 dB (A-weighted) via stereo headphones (Sennheiser HD 430). The HRTFs were the non-individualized “slv” set used in several of our other studies and originally measured at University of Wisconsin-Madison.

A room modeling software package (Odeon 4.0) was used to obtain image model reflection timings and azimuths for a surround sound loudspeaker array within a room conforming to listening test standards (ITU). The room dimensions were 8 x 6 x 3 m, with the listener centered between the loudspeaker array and the left and right walls, 4.5 m from the back wall. Loudspeakers were modeled at 0° and 120° azimuth, corresponding to “center” and “surround” channels. For each direct path, 1\textsuperscript{st} and 2\textsuperscript{nd} order reflections were selected (ref. Table 1). To establish reflection delay time as an independent variable, the derived azimuth and elevation for a given reflection was subsequently investigated at 3, 15, and 30 ms. Data in the “Az. Dif.” column correspond to the inside angle subtended on the horizontal plane between the direct and reflected sound azimuths.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3, 15, 30</td>
<td>0, 0</td>
<td>0 - 50</td>
<td>0</td>
<td>Floor</td>
</tr>
<tr>
<td>3, 15, 30</td>
<td>0, 0</td>
<td>72</td>
<td>72</td>
<td>Right wall</td>
</tr>
<tr>
<td>3, 15, 30</td>
<td>0, 0</td>
<td>151</td>
<td>151</td>
<td>Back wall</td>
</tr>
<tr>
<td>3, 15, 30</td>
<td>120, 0</td>
<td>120 - 50</td>
<td>0</td>
<td>Floor</td>
</tr>
<tr>
<td>3, 15, 30</td>
<td>120, 0</td>
<td>72</td>
<td>48</td>
<td>Right wall</td>
</tr>
<tr>
<td>3, 15, 30</td>
<td>120, 0</td>
<td>- 76</td>
<td>164</td>
<td>Left wall</td>
</tr>
</tbody>
</table>
Using a two-alternative forced-choice paradigm, thresholds were obtained at the 70.7% level within a tolerance of 1 dB with a “one up-two down” staircase algorithm that adjusted the level of the reflection. The reference (R) consisted of only the direct path, while the probe (P) consisted of the direct path plus the amplitude-scaled reflection. Two sequential stimuli were presented either as P-R, R-R, P-R, or R-P. Subjects indicated their response via a custom push-button interface as to whether or not the sequential stimuli were “same” or “different”. Thresholds were defined as the mean value of the five final staircase reversals at the minimum level.

Subjects were run under each of the time-location configurations indicated in Table 1 using both “anechoic” and “reverberant” stimuli conditions, for a total of 36 blocks. Block ordering was randomized across subjects. Anechoic stimuli included simulation of only the direct sound and a single reflection. Reverberant stimuli were generated via convolution of the direct sound with a synthetic reverberation decay, formed from exponentially-decaying white noise decorrelated between the left-right channels and at a level −20 dB below the direct sound. This corresponds to a non-acoustically damped version of the modeled room. The mid-band reverberation time in the 500 Hz - 1 kHz octave bands corresponded to 0.63 s.

RESULTS AND DISCUSSION

Figure 1 indicates the mean values of the results across seven subjects for anechoic and reverberant stimuli. For both anechoic and reverberant stimuli, thresholds decrease monotonically with increasing time delay between the direct sound and the reflection. When the direct sound is at 120°, an increase in azimuth angle difference corresponds to decreased thresholds for both anechoic and reverberant stimuli. However, the azimuth angle difference does not correspond to decreased thresholds when the direct sound is at 0°. For example, the threshold for an anechoic reflection with 0° azimuth difference at 3 ms (corresponding to a floor reflection in the modeled room) is −14.6 dB, compared to −9.8 dB for an anechoic reflection with an azimuth difference of 72°.

Compared to anechoic stimuli, thresholds are increased for reverberant stimuli by nearly 10 dB on average. Thresholds decrease less with increasing time delay compared to anechoic stimuli, and the range between experimental conditions is greater.

For both anechoic and reverberant stimuli, the lowest thresholds are for the direct sound at 120°, with the 164° azimuth difference. Most likely, subjects attended to a binaural cue (image broadening) for that class of stimuli, which may be easier to detect compared to ascertaining the timbre cue present when the direct and reflected sound were azimuthally co-located at 0°, or separated by only 48°.

ACKNOWLEDGEMENTS

Support provided by the NASA AOS program and NASA-SJSU co-operative agreement NCC 2-1095.

REFERENCES

3-channel dummy head design for binaural recording and synthesis.

Sin-lyul Lee, Hee-Suk Pang, Koeng-Mo Sung

School of Electrical Engineering and Computer Science, Seoul National University, Seoul, Korea

A binaural 3-D sound system is usually composed of binaural recording and binaural synthesis. The conventional 2-channel binaural system has several problems, which are difficulty in front source localization, front-back reversals, and sound quality degradation. In this paper, a 3-channel dummy head for binaural recording and synthesis is proposed to solve these problems. The proposed system is added a cardioid microphone to the brow of the conventional 2-channel dummy head to amplify the front sound intensity. As a result, spectral differences in the cone-of-confusion are more pronounced within overall audible frequencies. Since the accent microphone also uses the precedence effect, front-back reversals and front localization problems can be eliminated. Listening tests show that the proposed system is superior to the conventional one, especially in front-back discrimination and sound quality. The HRTFs measured by the proposed method can be used for the virtual reality, entertainment and auralization program.

1. INTRODUCTION.

The conventional 2-channel binaural system has several problems; difficult front source localization, front-back reversals, sound quality degradation. 3 channel dummy head adds a cardioid microphone to the brow of the conventional 2-channel dummy head. The aim for additional accent microphone is as follows.

① Sound localization in 2-D.

All incident wave in two dimension can be obtained by combining the outputs of an omni-directional responses with two bi-directional responses at 90° to each other[1].

\[ L(\theta) = A + B \cos(\theta) + C \sin(\theta) \]
\[ R(\theta) = A + B \cos(\theta) - C \sin(\theta) \]

A + B \cos(\theta) is a polar pattern of single forward-facing cardioid microphone and \pm C \sin(\theta) is a polar pattern of a single bi-directional microphone. In binaural recording, 3-channel dummy head can localize full 2-D sound source more exactly within median plane.

② 12 o’clock HRTF.

Danny Lowe[2] uses a transfer function corresponding to a location directly in front of a listener, that is, at 12 o’clock. He uses precedent effect and achieves front source localization. 3-channel dummy head can also use “precedent effect” with accent microphone, and can localize front sound source more easily.

③ Equalization.

Equalization is usually classified “Free-field equalization” and “Diffuse-field equalization”. Recently, modified equalization technique, “Weighted Diffuse-field equalization”[3] is introduced for stereo compatibility and sound quality. This technique uses conventional diffuse-field equalization and weight front hemisphere HRTFs in order to simulate real sound recording environment. 3-channel dummy head can simulate “Weighted Diffuse-field equalization” more easily with accent microphone.

④ Analysis in a Cone-of-confusion region.

Individualized HRTF mainly differs standard HRTF in cone-of-confusion region. Overall spectral differences in a cone-of-confusion region can improve sound source localization. 3-channel dummy head can solve this problem by using accent microphone.

2. EXPERIMENT.

Binaural signal was recorded by using 3-channel dummy head. The accent microphone has a variable
polar pattern.

Table 1. Results of localization experiments

<table>
<thead>
<tr>
<th>Accent Microphone</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 channel</td>
<td></td>
</tr>
<tr>
<td>Cardioid</td>
<td>Good performance of overall sound localization.</td>
</tr>
<tr>
<td>Super cardioid</td>
<td>Difficulty of side source localization.</td>
</tr>
<tr>
<td>Bi-directional</td>
<td>Front back ambiguity problem.</td>
</tr>
<tr>
<td>Omni-directional</td>
<td>Almost same result as 2-channel case.</td>
</tr>
</tbody>
</table>

Modified HRTF is obtained with 3-channel dummy head. The advantage of the modified HRTF is as follows.

① It enhances the front source localization, because spectral differences in a cone of confusion region is amplified overall audible frequencies.

② Ming Zhang[4] exaggerates notch frequency and causes sound quality degradation. 3-channel HRTF exaggerates broad peak frequency and can reproduce a natural sound.

3. CONCLUSION.

3-channel dummy head amplifies the front sound intensity, and spectral differences in the cone-of-confusion are more pronounced within overall audible frequencies. The precedence effect of an accent microphone can eliminate the front-back reversals and front localization problems. The result of listening tests shows that the 3-channel HRTF is superior to the conventional one, especially in front-back discrimination and sound quality. The HRTF measured by the proposed method can be used for the virtual reality, entertainment and auralization program.

4. REFERENCES.


2. USPTO patent no. 5,371,799


Extraction Process of Spectral Cues from Input Signals to Two Ears in Median Plane Localization

K. Iida\textsuperscript{a}, M. Itoh\textsuperscript{b}, E. Rin\textsuperscript{b}, and M. Morimoto\textsuperscript{b}

\textsuperscript{a}Multimedia Solution Labs, Matsushita Comm. Ind. Co., Ltd, 600 Saedo, Tsuzuki, 224-8539 Yokohama, Japan
\textsuperscript{b}Environmental Acoustics Lab, Fac. of Eng., Kobe University, Nada, 657-8501 Kobe, Japan

Most previous studies on median plane localization have treated spectral cues as a monaural phenomenon. Morimoto demonstrated that both ears contribute to the perception of the vertical angle even in the median plane. However, it is not yet clear whether spectral cues are extracted from the input signal to each ear independently, or they are extracted after some integration process of input signals to two ears. In this study, localization tests were carried out using the stimuli, which provided HRTF of different vertical angle in the median plane to the right and left ear, respectively. The results show that the subjects localized a sound image to either direction of HRTF, or sound images to both directions of HRTF. This suggests that a listener extracts spectral cues directly from the input signal to each ear, independently.

\section*{INTRODUCTION}

Most previous studies on median plane localization have treated spectral cues as a monaural phenomenon. Morimoto \cite{1} demonstrated that both ears contribute to the perception of the vertical angle even in the median plane. Moreover, the contribution of the near ear is more than that of the far ear, when a sound source shifts laterally from the median plane \cite{2,3}. Though some models on the extraction process of the spectral cues have been proposed \cite{2,3}, it is not yet clear whether spectral cues are extracted from the input signal to each ear independently, or they are extracted after some integration process of input signals to two ears. In this study, two hypotheses were made and the validity of them was examined.

\section*{EXTRACTION PROCESS OF SPECTRAL CUES}

According to the previous studies, the following two hypotheses on the extraction process of the spectral cues for the perception of the vertical angle could be built:

1) Hypothesis 1: Integration of the Spectra
   The spectra of the input signals to two ears are integrated into one spectrum. The spectral cues are extracted from the integrated spectrum.

2) Hypothesis 2: Integration of the Cues
   The spectral cues are extracted from the spectrum of the input signal to each ear independently. Listeners perceive the vertical angle of a sound image by integrating those spectral cues.

\section*{METHOD OF LOCALIZATION TESTS}

The HRTF of the subjects in the upper median plane were measured at seven elevations of every 30° from the front to the rear. The source signal was the band-limited white noise (280Hz - 11.2kHz). Stimuli were prepared by convolving the noise with the measured subject's own median plane HRTF. The HRTF from the sound source at different vertical angles were provided to the left and right ears, respectively. Namely, directional information provided to the left and right ears are different. Stimuli were presented to the subjects by the sound field simulation system through near-ear loudspeakers. This system compensates the transfer function from the near-ear loudspeakers to the entrance of the ear canal of the subject by DSP, and the crosstalks between left and right ears are negligible small. Each stimulus was presented 10 times in random order at 60±0.4 dBA at the entrance of the ear canal of the subjects. The duration of the stimuli was 1s and the interval between two stimuli was 9s. The task of the subjects was to mark down the perceived elevation of the sound image on the recording sheet. Subjects were four males with normal hearing sensitivity.

\section*{RESULTS AND DISCUSSIONS}

The subjects reported that they perceived all sound images outside the head. Figure 1 shows an example (Subject: IT) of the responses to the stimuli in case that the elevations provided to two ears are the same. This shows that he localized sound images accurately with the simulation system.

Figure 2 shows an example (Subject: IT) of the responses to the stimuli, in case that the HRTF for
different elevations were provided to the left and right ears. Figure 2(a) shows the 10 responses in case of HRTF for 0° and 180°. The subject sometimes perceived two sound images simultaneously, and sometimes one sound image. The perceived elevations agree with the provided ones. This means that the listener perceives the elevation of a sound image from the spectrum of the input signal to each ear, independently. Figure 2(b) shows the responses in case of HRTF of 30° and 60°. The responses agree with either elevation of two HRTF. Figure 2(c) shows the responses in case of HRTF of 120° and 30°. The subject perceived a sound image around 30° in most trials, and two sound images once. These were common behavior to other subjects. Therefore, it seems reasonable to consider that the hypothesis 2 is credible.

The mean localization error, $e$ was obtained by Eq. (1).

$$e = |R - S|$$

where $R$ is the perceived angle, and $S$ is the simulated angle. Table 1 shows the errors when the elevations provided to two ears are the same (Case 1), and when they are different (Case 2), are shown in Table 1. Since two elevations are provided in the latter case, the smaller error is regarded as the error. Table 1 indicates that the localization error in Case 2 is almost the same as that in Case 1.

**CONCLUSION**

These results of the localization tests infer that the spectral cues for the vertical angle perception were extracted from the spectrum of the input signal to each ear, independently. It seems reasonable to consider that the hypothesis 2 is credible.

**REFERENCES**


**Table 1.** Mean localization error $e$

<table>
<thead>
<tr>
<th></th>
<th>Case 1 (the same HRTF)</th>
<th>Case 2 (the different HRTF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e$ (deg.)</td>
<td>11.6</td>
<td>9.3</td>
</tr>
</tbody>
</table>
Binaural Interaction of the Balanced Modulation with a Pure Tone

E. Ozimek, and J. Konieczny

Institute of Acoustics, A. Mickiewicz University, Umultowska 85, 61-614, Poznań, Poland

The study concerns the evaluation of the modulation depth evoked by the binaural interaction of the balanced modulation (BM – two sidebands only) presented to one ear and a pure tone of frequency set between two sidebands presented to the other. Such binaural stimuli (target) have produced a sound image close to the AM signal. Two successive stimuli were presented at a random order on each trial of the run: the target and the AM signal being a reference one. The modulation depth of the reference was matched to the target, according to the one-up, one-down adaptive procedure. Estimation of the binaurally perceived modulation depth was made for different carrier and modulation frequencies. The experimental data showed that the perceived modulation depth of the target was significantly lower than modulation depth which one could expect as a result of a simple summation of the BM and the pure tone. This holds true for all modulation and carrier frequencies applied.

INTRODUCTION

The results of [1] concerning the binaural perception of the modulation depth of AM signals with the interaural difference in modulation depth have shown that the interaction of those signals is not simple. A question arises of how this interaction works in the case of the binaural perception of the balanced modulation (BM) and a pure tone with a determined frequency. More specifically, what is the value of the binaurally perceived amplitude variations when the BM signal is presented to the one ear and the pure tone to the other ear. How does it depend on the modulation rate or carrier frequency? The answers to these questions are the basic issue of this study.

BALANCED MODULATION AND BEATS

Balanced modulation (BM) is similar to amplitude modulation (AM) except that there is no distinction between the modulating signal and the carrier. If a signal \( x_m(t) \) modulates a signal \( x(t) \) in a balanced modulator, the output signal \( x(t) \) has a form \( x(t) = x_m(t) x_c(t) \). Assuming that \( x_m(t) \) and \( x_c(t) \) are sine and cosine waves with frequencies \( \omega_m \) and \( \omega_c \) and relative phase equals zero one can write

\[
x(t) = \sin(\omega_m t) \cos(\omega_c t)
\]

The spectrum of \( x(t) \) consists of only two sinusoidal components with the sum and difference frequencies.

\[
x(t) = \frac{1}{2} [ \sin(\omega_c + \omega_m) t + \sin(\omega_c - \omega_m) t ]
\]

Unlike in the case of AM, in the BM (having no a DC component), there are no original frequencies \( \omega_c \) and \( \omega_m \). Equation (2) indicates that there is no real difference between balanced modulation and the beating of two sinusoidal tones with equal amplitudes.

Assuming that these amplitudes are equal 1, the envelope of the signal for beating tones is given by

\[
A_m(t) = 2 \cos(\Delta \omega t/2)
\]

where \( \Delta \omega \) is the frequency difference between beating tones. By contrast, the envelope fluctuation of the AM signal is given by

\[
A_{AM}(t) = 1 + m \cos(\omega_m t)
\]

Equations (3) and (4) show that the rate of the amplitude changes due to beats is the same as that of the changes due to AM if the frequency difference between beating tones \( \Delta \omega \) is the same as the modulation frequency \( \omega_m \) of AM. For equal rate of the amplitude changes (variations) and modulation index \( m = 1 \), one might expect that both kinds of amplitude changes would lead to the same amount of the perceptual changes in amplitude. This is not the case, because the envelope of the AM tone is sinusoidal, whereas the envelope of the beating tones has a sharper structure. Therefore, one might assume that the binaural perception of the amplitude variations in the BM signal may be less effective than the changes in the AM signal. The aim of the experiment 1 was to verify this assumption.

In experiment 2, the binaurally perceived amplitude variations were determined when the BM signal was presented to the one ear and the pure tone with determined frequency to the other ear. When the frequency of the pure tone is chosen so that it lies between frequencies of the two BM components and both stimuli are added, the AM signal is obtained. The modulation depth of AM signal produced this way depends on the amplitude ratio of the pure tone and the BM components. In this light, it was interesting to examine how the binaural system processes the BM signal presented to the one ear and the pure tone to the other ear. In other words does the binaural interaction of those stimuli leads to formation of AM signal? If so, what is the value of the modulation depth of such AM signal?
signal for determined amplitude ratio of the tone and the BM components?

**EXPERIMENT AND RESULTS**

Experiment 1 concerned the binaural perception of the amplitude variations (beats) when two BM signals, with determined amplitude ratio of their components, were presented to the left and to the right ear. The aim of the experiment was to estimate the perceived variation in amplitude of BM signal as a function of the modulation frequency. Two successive stimuli were presented on each trial: the BM stimulus (target) with determined amplitude ratios of the components and the matching stimulus (AM). Measurements of estimates of the amplitude variations in BM were made for carrier frequency of 1000 Hz, modulation frequencies of 4, 32, 64, and 128 Hz, and the amplitude ratios of the BM components equal: 1:1; 1:0.5; 1:0.25. The modulation depth of the matching stimulus was adjusted during the run according to a one-up, one-down adaptive procedure. Figure 1 shows the matched modulation depth of AM signal as a function of the modulation frequency, for three amplitude ratios of BM components.

The data depicted in Fig. 1 imply that for the low beating rate equalled 4 Hz (for which amplitude variations are perceived as changes in loudness) and for BM with equal amplitude of components (that corresponds to 100% of beats), the beats perceived correspond to changes in amplitude of AM signal with the modulation index $m \approx 40\%$. For the beating rates of 32 and 64 Hz (for which beats are perceived as a sensation of roughness), the modulation index $m \approx 70\%$, that is what Terhardt [2] found for monaural perception. For a beating rate higher than the roughness frequency (at which the subject perceives a sensation of timbre), the matched modulation index monotonically decreases with increasing modulation frequency reaching $m \approx 45\%$ for $f_m = 128$ Hz. The results obtained in Exp. 1 generally indicate that the binaural evaluation of amplitude variations for beating tones (BM signal) is less effective than the evaluation of amplitude changes in the AM signal.

Figure 2 shows the data obtained in Exp. 2, for which the BM signal is presented to one ear and the pure tone is presented to the other ear. As can be seen, the values of the matched modulation depth fall within the range of only 5 – 10\%. They are much lower than those which might be expected from a sum of the BM and pure tone. Moreover, the matched modulation depths do not depend on the modulation and carrier frequency. The obtained data cannot be interpreted on the basis of central masking of the BM components by a pure tone since this value of masking is relatively small.

**FIGURE 1.** Dependence of the binaurally matched modulation depth of AM signal to the amplitude variations (beats) in BM as a function of modulation frequency. Data averaged across three subjects. Each curve corresponds to a different amplitude ratio of beating tones in BM.

The experiment 2 generally shows that the binaural sound image consisting of the BM signal and pure tone applied does not form AM signal with the modulation depth corresponding to a linear sum of those stimuli.

The results obtained in Exp. 1 generally indicate that the binaural evaluation of amplitude variations for beating tones (BM signal) is less effective than the evaluation of amplitude changes in the AM signal.

**FIGURE 2.** Dependence of the modulation depth of AM signal binaurally matched to the amplitude variations formed by the BM and pure tone signals, as a function of modulation frequency. Data averaged across three subjects.

**ACKNOWLEDGEMENTS**

This study was supported by the State Committee for Scientific Research (KBN), Grant No 8 T11E 017 17.

**REFERENCES**


Compatibility of models with data on dichotic pitch

F. A. Bilsen and J. Raatgever

Applied Physics Department, Delft University of Technology, Lorentzweg 1, 2628 CJ Delft, The Netherlands

Dichotic pitch phenomena are generally considered natural byproducts of the mechanisms of binaural hearing. Two main aspects of dichotic pitch are the specific value of the pitch given the interaural phase relationship and the lateralized position of its pitch image. Models like cross correlation (CC), (modified) Equalization-Cancellation (EC and mEC) and the Central Spectrum theory (CAP-CS) have to predict these aspects correctly to be qualified as generally applicable.

CENTRAL SPECTRUM MODEL

Dichotic pitches are perceived when the same white noise is presented to both ears but with a particular interaural phase relationship. The following dichotic pitches have been reported: the Huggins Pitch (HP) for a 2π-phase transition in a limited frequency range [See e.g. 1, 2; for variations 6], the (a)symmetric Fourcin Pitch (aFP and sFP) for two uncorrelated noises with interaural delays \( T^1 \) and \( T^2 \) [2, 3], the Dichotic Repetition Pitch (DRP) for only one single interaural delay \( T \) [1, 2, 3, 4], the Multiple Phase Shift Pitch (MPSP) for a series of 2π-phase transitions equally spaced in frequency [1], the Binaural Edge Pitch (BEP) for a π-phase transition in a limited frequency range and the Binaural Coherence Edge Pitch (BICEP) [2, 5]. Acronyms and interaural phase configurations are summarized in Fig. 1 (columns 1 and 2).

HP, BEP and BICEP have a pure tone character, while DRP, FP and MPSP behave like periodicity pitch. In addition, a dichotic pitch has a more or less well-defined binaural image separated, in general, from the (diffuse) image of the generating dichotic noise itself. As both pitch value and pitch image position (lateralization) have been shown to be correctly predicted by the Central Spectrum (CS) theory [1, 2, 3, 4, 6], existing data are “summarized” by CS equations in columns 3 and 4 of Fig. 1.

In accordance with CS theory, pitches and their lateralizations can already be prognosed from the interaural phase patterns (column 2) by inspecting the dash-dotted lines. Being straight and going through the origin (0 phase, 0 frequency), these lines symbolise an internal delay \( \tau \) (similar to an interaural delay \( T \)). For example, for HP and MPSP the intersection with the phase pattern indicates the value and position of peaks in the central activity pattern (CAP) at \( \tau = 0 \). For aFP the dash-dotted line runs parallel to the dashed line \( T^2 \) and shifted by \( \pi \), thus indicating a straight valley of zero power from noise 2 in the CAP at \( \tau = T^2 \), which “highlights” the central spectrum part due to noise 1 at this internal delay. Different highlighting is obtained in the case of sFP by the additive interference of \( T^1 \) and \( T^2 \) at \( \tau \). Such highlighting is absent with the DRP stimulus, which therefore offers an infinite range of central spectra each with its own pitch and lateralization [4].

CROSS CORRELATION

Now it is examined to what extent also other current theories comply with these data. A summary is given in columns 6 to 8 of Fig. 1. Correct prediction is indicated with + and incorrect or non-prediction with – for pitch value and lateralization respectively (+,–).

The inadequacy of the concept of cross correlation (CC) is manifest already from the simple fact that identical pitch values are predicted for the aFP and sFP cases, which is in conflict with the data [3]. Further, it is unclear how ambiguity of pitch should be predicted from one cross-correlation peak pair, the more so as the peaks have equal polarity as in the case of aFP. Also, it is not clear how a negative peak at \( T^2 \) should predict a pitch image position corresponding to an internal delay \( T^2 \).

Alternatively, one might consider the possible virtues of a “Summary Cross Correlogram (SCCG)”, to be defined as the result of the “addition” of peripherally-filtered cross correlation functions, very much in analogy with the Summary Auto Correlogram (SACG) as promoted in studies on monaural periodicity pitch [7]. It has been shown that the SACG resembles the wide-band auto correlation function in its main features (e.g. position of first peak). Likewise, the SCCG resembling the wide-band cross correlation function should be expected to be unable to explain dichotic pitch behaviour for reasons similar to those mentioned above [3].

EQUALIZATION-CANCELLATION

Durlach’s original Equalization-Cancellation (EC) model is basically able to predict HP, MPSP, BEP and BICEP values [2, 5]. Also aFP is correctly predicted in addition mode. However, the model has to switch to subtraction mode for sFP [2, 3]. Further, sFP data are not predicted by the EC model, simply because equali-

---

1 Recent observations by the first author show that the sFP stimulus can be used successfully to draw one’s attention to one spectrum from the continuum of central spectra of a DRP stimulus.
zation by interaural delay always recovers the difference between the two delays, not the averaged value. The interaural delay needed in the cancellation process could possibly be extracted as an indicator for pitch-image position. But given this possibility, we still are faced with the problem that multiple images are not predicted. Moreover, the correct prediction of both pitch value and lateralization always calls for addition instead of subtraction in the cancellation process. Therefore, in column 7 of Fig.1, we choose to consider the EC model in its addition mode only (Note that this implies a deviation from the general preference for subtraction in the modelling of BMLDs). Further, it is assumed that the EC mechanism (in the absence of a signal) strives for maximum reduction of the noise.

Culling and colleagues [2] proposed a modified Equalization-Cancellation (mEC) model performing an equalization by adjustment of internal delay (and/or level) in each frequency channel (auditory filter) independently. An obvious reason for its failure to predict sFP along with aFP is its unique way of operation, i.e. to generate only one optimal “recovered spectrum”. For DRP, the mEC model does not predict any recovered spectrum at all. Further, lateralization is not dealt with by the mEC model, because the possibility to extract a single equalization delay as an indicator for laterality, is essentially absent.

REFERENCES

A very efficient software tool for real-time multichannel convolution is presented. The software is intended to the reproduction of Ambiophonic surround sound. The program runs on a Windows based PC using compatible sound cards. The binaural input signal is acquired by the sound card and then processed by the software producing outputs for the stereo-dipole and for the surround loudspeakers. Using state of the art PCs is possible to achieve set-ups with more than 10 loudspeakers and simulate room impulse responses larger than three seconds.

INTRODUCTION

The multichannel surround sound systems, well established in the industry of the cinema, still lack for providing a true periphonic space sensation. Also new difficulties are added in the production and mixing of multichannel sound and also new special techniques of mixing are needed. On the other hand, the systems based on the HRTF, binaural systems and transaural systems using cross-talk cancellation filters, have also experienced a considerable evolution. These systems have the advantage that only two channels are required for their transmission.

Recently a novel system of sound reproduction called Ambiophonics has been proposed [1]. This system combines the well-known techniques of cross-talk cancellation using closely-spaced loudspeakers to reproduce the front signal [2], and a group of loudspeakers surrounding the listener in order to reconstruct hall ambience signals. These signals are derived from the left and right channels coming from the CD, convolved with a set of real hall impulse responses. Only two channels are stored or transmitted and the sensation of 3D space is quite good.

However, powerful computational resources are needed in order to generate the signals that feed the different loudspeakers in a typical Ambiophonics set-up. A cross-talk cancellation system is required to obtain the stereo-dipole signal. The cross-talk filters are usually not too long. On the other hand a set of quite long impulse responses that simulate the original room impulse response (RIR) are convolved with the source signals. The software presented in this work can carry out in real-time the necessary multichannel convolutions in order to implementing an Ambiophonics set-up using a PC.

HOST SIGNAL PROCESSING

The new techniques in the sound field demand more and more processing power to implement the digital signal processing algorithms. The manufacturing companies of Digital Signal Processors (DSP) are aware of this fact and bring periodically faster and more powerful DSP to the market. Different companies coexist that manufacture different DSP models, all them incompatible to each other. Also the development tools for these DSP are not in many cases easy to manage. All these difficulties, together with the difficulty of designing the complementary hardware to produce a complete system of audio processing using DSP, made us to think about other ways of carrying out digital audio processing. The personal computer has experienced an impressive evolution in the last years, as much in the process speed as in the multimedia devices that can be connected to it. As a result of the aforementioned statement, the personal computer can be used for real-time digital signal processing applications.

In order to implement multichannel convolutions, software and mathematical optimizations have been employed. We use fast convolution methods (overlap-save) based on Fast Fourier Transforms (FFT). Therefore optimization of the FFT algorithm is an essential software need. There are mainly two public domain software libraries that supply FFT routines for software developers: the FFTW and the Intel Signal Processing Library (SPL). The first one provides an open code of C language and can be compiled to run on different platforms and operating systems. The second one is distributed already compiled in two formats: static libraries (.LIB) and dynamic libraries (.DLL). This software library works only for Intel (and compatible) processors over Win32 environments. The
Intel SPL has been used in this project because it yields lower execution time than FFTW over Intel processors and also because we had some experience in its use, gained from previous developments [3].

**Intel Signal Processing Library**

The Intel Signal Processing Library provides a set of signal processing routines optimized for general-purpose Intel (and compatible) processors rather than specialized DSP processors. The library includes functions for finite impulse response (FIR) and infinite impulse response (IIR) filters, fast Fourier transforms (FFTs), wavelet transforms, tone generation, and many vector operations.

Computation times for the FFTs are very good (compared with commercial DSP). For instance a 1024 points single precision real FFT takes only 28 \(\mu\)s in an AMD K7 at 1 GHz and a \(2^{14}\) points FFT takes 0.71 ms. Using state of the art PCs it is possible to achieve set-ups with more than 10 loudspeakers and simulate room impulse responses larger than three seconds.

**SOFTWARE DESCRIPTION**

The software tool has been developed using the Borland Delphi compiler. Delphi uses Object Pascal and has a versatile IDE and powerful debugging tools. This compiler produces very optimized code for the Intel processors family. The software presented along this paper has been named Ambiovolver. Figure 1 shows Ambiovolver main window. A friendly GUI (graphic user interface) provides to the user an easy handling and fast learning. From the main window the user can reach different configuration windows, Load the bank of filters and Play or Stop the convolution.

In order to provide the impulse responses (IR) to the software, a simple and flexible method is used. Files containing the IR must be stereo .WAV files. The files have to be named as follow: ‘anythingN.wav’ where N represents the number of the output channels. The first two files (N=1 and N=2) must correspond to the stereo-dipole pair. For each stereo WAV file, the left channel contains the impulse response that has to be convolved with the left program signal, and idem for the right channel. Results of two convolutions are added and sent to the output after applying a gain factor selectable by the user. A wide variety of WAV file formats are accepted by the software. Particularly 20, 24, and 32 integer and floating-point formats are supported.

**ACKNOWLEDGMENTS**

This work has been partially supported by Generalitat Valenciana grant no. GV00-094-14. The authors want to thank Angelo Farina for their valuable suggestions about Ambiophonics that helped us to improve the software.

**REFERENCES**


Localization Accuracy of Speech in the Presence of Non-Directional and Directional Noise Maskers

K. Abouchacra\textsuperscript{a,b} and T. Letowski\textsuperscript{b}

\textsuperscript{a}American University of Beirut Medical Center, Department of Otolaryngology – Head & Neck Surgery, Beirut – Lebanon; Human Research and Engineering Directorate, \textsuperscript{b}U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland, 21005-5425, USA

Sound-field audiometry (SFA) is a commonplace in most audiological centers. However, the focus of SFA is currently limited to hearing thresholds and speech recognition data. This investigation was aimed at establishing localization norms for signals presented in noise. A speech signal (S) was presented in non-directional speech spectrum noise (NDN), or in the presence of a single-source directional noise (DN) of a constant level [65 dB (A)]. Localization accuracy was measured for 40 normal-hearing subjects in 40 sound-field conditions (35 S+DN and 5 S+NDN conditions). The S and DN sources were placed at 0-, +45-, +90-, +135-, or 180-deg azimuth. Results indicated that localization accuracy was similar for DN and NDN conditions at all sensation levels. In the DN conditions, localization accuracy varied, in a complex manner, as a function of the absolute location of S and the amount of separation between the S and DN. Data are discussed in terms of accuracy, front-back errors, and directional shifts to and from the DN source.

INTRODUCTION

Limited data in the literature exists on how the presence of a secondary sound source affects the perceived location of a speech signal [1,2]. The specific purpose of the study was to determine how the presence of directional or non-directional noise affects the listener’s ability to determine the location of an incoming speech signal in the horizontal plane.

METHOD

Subjects. Forty listeners between the ages of 18 and 29 participated in the study. Listeners had pure-tone hearing thresholds better than or equal to 15 dB HL at audiometric frequencies from 125 - 8000 Hz [3]. The difference in hearing thresholds in both ears did not exceed 5 dB at any test frequencies.

Stimuli. The interfering stimulus (N) was speech-spectrum noise presented at 65 dB (A) at the listener’s head location. Noise was presented through either (1) six matched loudspeakers distributed in space to create non-directional background noise (NDN) or (2) through a single, boom-mounted loudspeaker to create directional noise (DN). The target signal (S) was a single, spondaic word “northwest” presented through a second boom-mounted loudspeaker. Both boom-mounted loudspeakers were located at 1m from the listener (ear level). All loudspeakers were Bose 108515K. Both S and DN sound sources could be placed at 0°, ±45°, ±90°, ±135°, or 180° azimuth.

Procedure. Each listener participated in two, 2-hour sessions. Testing was conducted in an anechoic chamber (2.7x2.7x1.9m). Localization performance was evaluated at four sensation levels: 0, 6, 12 and 18 dB SL\textsuperscript{1}. To determine the decibel level needed to achieve the appropriate SLs for a given listener in a given environment, we measure the listener’s detection thresholds (DTs) for S in each NDN and DN condition.

\[ \begin{array}{c|c}
0 \text{ dB SL} & DT \\
6 \text{ dB SL} & DT + 6 \text{ dB} \\
12 \text{ dB SL} & DT + 12 \text{ dB} \\
18 \text{ dB SL} & DT + 18 \text{ dB} \\
\end{array} \]

A touch-sensitive response pad (KoalaPad\textsuperscript{TM}) and stylus was given to the listener to make localization responses. The pad displayed a drawing of the subject’s head and a pointer ring for making localization responses. The test consisted of 320 trials (4 sensation levels x 40 combinations of S and DN or NDN sources x 2 sessions). After each presentation, the subject touched the stylus to a position on the pointer ring corresponding to the perceived S location. (Subjects could respond in 1° increments anywhere along the 360° azimuth circle). No feedback was given to subjects about their responses. Subjects participated in a 30-min training session prior to testing.

RESULTS AND DISCUSSION

Localization Accuracy. The results of the study indicate that the listener’s ability to determine signal direction in DN or NDN decreased as SL decreased. The rate of degradation varied as a function of the S and N locations. To measure localization accuracy, we established a cut-off point of ±15° azimuth of the true S location to determine correct responses. In the 18 and 12 dB SL conditions, localization accuracy (within ±15°) was highest for S=0° (90-100%; Fig 1) and for

\begin{figure}[h]}

\end{figure}
Sensation Level (dB SL) and poorest for S=±135° (25-55%) and S=±180° (45-65%). In the 6 dB SL condition, performance was highest for S=±90° (42-68%), slightly lower for S=0°, S=±45°, and S=±180° (25-55%), and lowest for S=±135° (10-32%). At 0 dB SL, highest accuracy (20-40%) occurred at S=0° (Fig. 1), S=±45°, and S=±90°; lowest performance occurred at S=±45° (10-20%).

Front-Back Errors (FBEs). An analysis of FBEs revealed that back-front errors were more common than front-back errors for most S locations and listeners. Both types of errors had tendencies to increase as SL was lowered, with the least FBEs occurring when S was at 0° or ±45° azimuth (5-15%), and the most FBEs when S was at ±135° or 180° azimuth (15-40%).

Directional Shifts. When signals were presented at ±45° or ±135° azimuth, listeners made always more lateral (toward 90°) than medial (toward 0°) errors for a noise source located at 0, 180, or ±90° azimuth. These data are presented in Fig. 2 as a percentage of medial errors to all errors made in +45°/−45° range around the S source position. This result indicates that listeners had a tendency to hear the sound source more lateralized than it actually was. We refer to this phenomenon as a lateralization effect.

When the S=±90° azimuth, listeners had a tendency to respond more medially (toward 0°) regardless of the noise type (Fig. 3).

Fig 1. Localization accuracy for S=0° (within ±15° of S).

Fig 2. Localization shifts for S=45° and S=135°. Bold dashed line indicates no shift. (a) p=A/(A+B) x 100%, (b) p=D/(C+D) x 100%, (c) p=E/(E+F) x 100%.

When the S=±90° azimuth, listeners had a tendency to respond more medially (toward 0°) regardless of the noise type (Fig. 3).

CONCLUSIONS

Overall, the results of this study indicate that the listener’s ability to determine the direction of a S in DN or NDN decreased as the SL was lowered. The rate of degradation varied as a function of the relative S and N locations. The results support also the notion that about +9 dB signal-to-noise ratio is needed for accurate localization of a speech source [4,5]. Finally, directional shifts in listener judgments of target source location were noted. These data can be used to establish localization norms for signals presented in noise in a sound field.

ACKNOWLEDGEMENT

The US Army Research Laboratory supported this research.

REFERENCES


*Sensation level is defined as the decibel level about the established noise in a sound field.

1 Sensation level is defined as the decibel level about the established noise in a sound field.

Fig 3. Localization shifts for the signal sources located at 90° and -90°. (a) p=B/(B+C) x 100%, (b) p=G/(G+F) x 100%.
Using Binaural Hearing for Localization in Multimodal Virtual Environments

Matti Gröhn, Tapio Lokki, Lauri Savioja

Telecommunications software and multimedia lab., Helsinki Univ. of Tech., P.O.Box 5400, FIN-02015 HUT, Finland

In this article we present the results of localization experiment accomplished in a real virtual room. The spatial audio reproduction methods used in experiment were HRTFs (for headphones), direct loudspeaker reproduction, and vector based amplitude panning. Also the effect of visual stimulus was explored.

BACKGROUND

Virtual environment is an interactive immersive multisensory 3-D synthetic environment; it uses position-tracking and real-time update of visual, auditory, and other displays (e.g., tactile). Currently spatial audio has been mostly used for sense of presence.

There are four common tasks in immersive visualization (localization, orientation, navigation, and data representation). In this research localization is defined as user ability to define direction and distance of the target. The other tasks and more detailed description of this research project as a whole is in article [1].

In this research we are comparing head-tracked binaural headphone reproduction (using HRTFs), direct loudspeaker reproduction, and vector based amplitude panning (VBAP) using loudspeakers [2]. The task of the user is to point to direction of the perceived sound source, and click the button of a interaction device. According to Djelani et al. [3] pointing is an appropriate method for localization experiments. In our test the user can freely move inside the virtual environment (as they typically do while they are using some application).

LOCALIZATION TEST

According to Blauert [4] the absolute lower limit for localization accuracy in front is one degree. In our experiment we have measured the localization accuracy inside a real virtual room, where there are many factors, that decrease the localization accuracy. For example, we didn’t use individualized HRTF’s, because they typically are not available. Instead, HRTF’s were measured a from dummy head. On the other hand, the screens and room reverberation will deteriorate loudspeaker reproduction accuracy.

Test environment

Localization tests has been made in our virtual room [5] at Helsinki University of Technology. For spatial audio reproduction we use 15 Genelec 1029A loudspeakers. Alternatively headphones (Sennheiser 590A) can be used. Our audio reproduction system is described in article [6] in more detail.

Method

To find out the localization accuracy, a listening test was conducted. The auditory stimulus was pink noise. The task of the subjects was to point to the direction of the perceived sound source using a wand. We had eight male test subjects. Each subject accomplished four different tasks, and each task had 17 different sound source locations. Locations have been played to subjects in randomized order to avoid learning effect. Each subject had 34 locations with HRTF reproduction, 15 locations where sound was reprocud using one loudspeaker (LS), and 19 locations with VBAP reproduction.

In first two tasks there were no visual stimulus available on the screens. The order of headphone, and loudspeaker reproduction was randomized, in no visual stimulus, and with visual stimulus tasks. In the last two tasks the visual stimulus was a large 3D model of the estrogen receptor displayed on screens. (The locations of sound sources were not related with the molecule.)

We measured the pointing accuracy using Ascension magnetic tracking device. We measured the azimuth, and elevation angle separately. The duration of finding each location was also stored.

RESULTS

The azimuth localization accuracy was in average 9.7 degrees. With current reproduction methods, the perceived elevation accuracy was not so good, average error
was 27.0 degrees. Especially with HRTF’s the perceived sound source location was in average much higher, than given location. The average and median values of azimuth and elevations errors are shown in Table 1.

Table 1. Average and median values of azimuth and elevation errors for each reproduction type.

<table>
<thead>
<tr>
<th></th>
<th>Average Azimuth</th>
<th>Average Elevation</th>
<th>Median Azimuth</th>
<th>Median Elevation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRTF</td>
<td>10.6</td>
<td>31.7</td>
<td>8.3</td>
<td>27.4</td>
</tr>
<tr>
<td>LS</td>
<td>8.7</td>
<td>22.5</td>
<td>7.5</td>
<td>20.1</td>
</tr>
<tr>
<td>VBAP</td>
<td>8.9</td>
<td>23.0</td>
<td>5.6</td>
<td>18.9</td>
</tr>
</tbody>
</table>

We use an analysis of variance (ANOVA) model for the analysis. There was no significant difference between no-visual stimulus, and visual stimulus tasks. The visual stimulus was not related with sound source locations, which explains the result.

HRTF reproduction was significantly worse both in azimuth \( (p < 0.01) \), and elevation \( (p < 0.01) \) accuracy compared with other reproduction methods. On the other hand there was no significant difference in accuracy between the direct loudspeaker reproduction, and VBAP. Boxplot of the azimuth and elevation localization accuracy is seen in figure 1.

There was significant difference in localization time between the HRTF reproduction, and the direct loudspeaker reproduction \( (p = 0.004) \). The localization last in average longer with HRTF reproduction.

**CONCLUSIONS AND FUTURE RESEARCH**

In a real virtual room with used reproduction methods the average azimuth error was 9.7 degrees, and elevation error 27.0 degrees. Non individualized HRTFs were significantly more inaccurate than loudspeaker reproduction methods. The VBAP was in this experiment as accurate reproduction method as direct loudspeaker reproduction. In this experiment the visual stimulus doesn’t have on effect on localization accuracy.

Azimuth accuracy achieved in this experiment is convenient for our purposes. On the other hand more research should be carried out to find out how we can increase the elevation accuracy to convenient level.

Future research with other auditory stimuli should be conducted. Another area for future experiments will be testing the effects of integrated auditory and visual stimuli.

**ACKNOWLEDGMENTS**

This research was partly funded by Helsinki Graduate School in Computer Science and Engineering. The authors like to thank Juhani Forsman, and Tommi Ilmonen for their effort, and time with our audio reproduction system, and virtual environment.

**REFERENCES**

The dolphin’s sonar and binaural hearing

G. L. Zaslavskiy

University authority for applied research, Ramot, Tel-Aviv, Israel

The dolphin was required to determine which of two transducers transmitted the first of two identical broadband clicks delayed relative to each other. A minimum angle between the transducers and minimum delay between the clicks just detectable by the dolphin were measured. A maximum angle between the transducers was measured as a function of the click amplitudes.

Binaural hearing similar that in humans is thought to provide the dolphin with ability to localize targets and a sound source. On the other hand, the dolphin possesses directional hearing [1], let alone directional click transmission, which are absent in humans. As far as humans are concerned, the spatial hearing does not involve any substantial hearing directionality. The directionality appears to negate essential ability of the binaural hearing to instantaneously localize a sound source.

We investigated how the directionality of the dolphin’s auditory system affected a particular binaural phenomenon is usually referred to as the “echo suppression”. The “echo suppression” phenomenon suggests that echoes in a reverberant room are not heard by human subjects [2]. The dolphin is easily trained to determine a transducer of a primary sound. However, it is not clear whether a delayed sound is suppressed in the auditory system or the dolphin can hear both sounds and simply determines, which of the sounds comes first. A minimum angle between the transducers as well as a minimum delay between the clicks at which the dolphin was able to locate the transducer of a primary click were measured. We also measured the maximum angle between the transducers as a function of the click intensities. For a directional receiver, near the absolute threshold for a click, the bigger angle between transducers, the larger click amplitudes should be if both clicks are to be heard together. Therefore, such a function may describe the dolphin’s receiving beam pattern.

METHOD

The subject was adult male bottlenose dolphin. The experiments were conducted in a 28 × 13 × 4 m concrete pool using a two-response forced choice paradigm. 1.2-cm spherical transducers were positioned at each side of a net partition at 1-m depth. The partition set a 5-m range, from the dolphin to the transducers. The dolphin was required to approach the transducer transmitting a primary click. The peak frequency of the clicks was at 110-120 kHz. Click repetition rate was 5 Hz. The method of constant stimuli was used, to measure the dolphin’s auditory thresholds. First, for a delay between a primary and a delayed click of 0.3 ms, the minimum angle between the transducers, at which the dolphin was able to locate the transducer of a primary click, was measured. Second, for a 1.5°-angle between the transducers, a minimum delay between the clicks was measured. The sensation level of the clicks was around 40 dB (above absolute threshold). Finally, a maximum angle between the transducers as a function of the click amplitudes was measured. The measurements were made for 8 different amplitudes of the clicks above a single click absolute threshold.

RESULTS

For a 0.3-ms delay between the primary and delayed clicks the minimum angle between the transducer was found to be around 0.4° (figure 1).

FIGURE 1. The dolphin’s correct response as a function of the angle between the transducers.

Clearly, at such a small angle, the dolphin could not resolve the transducers by azimuth with the receiving beam pattern as wide as 13.7° at 120 kHz, [1]. On the other hand, it is highly unlikely, that the dolphin was able to locate imaginary source created by the “suppression echo” phenomenon at the 0.4°-angle between the transducers simply because a minimum audible angle measured in the dolphin was around 1.5°. Regardless, a very small angle between the transducers, the dolphin scanned with its head at the angle as big as 8 – 15°. The minimum angle between the transducers was a slightly larger than a threshold angle between targets [3]. In order to determine which of two identical targets separated in the horizontal plane by a very small angle was
at a closer range, the dolphin directed one of steep slopes of a transmitting beam pattern at both targets, causing the target echoes to be different in amplitude. The dolphin likely used the same technique to locate the transducer of a primary click. Which of the clicks is louder should depends on which of the slope (right or left) of a receiving beam pattern is directed at the transducers. The dolphin likely to perceive a primary and a delayed click as a double-click coming from a single source. However, when it scanned with its head from left to right at the angle larger than the angle between transducers, the amplitude ratio between If this was the case, the dolphin discriminated between the double-clicks with reversed temporal order of a low and a loud click.

At the angle between transducers of 1.5°, the minimum delay between the primary and delayed clicks detectable by the dolphin was as small as 17-18 µs (figure 2).

FIGURE 2. The dolphin’s correct response as a function of the delay between the primary and delayed clicks.

It seems doubtful, that such a small delay could cause a delayed click to be suppressed in the dolphin’s auditory system. The dolphin rather heard the primary and delayed clicks as two time separate acoustic events and determined which of them came first. The threshold delay is essentially the same value as the interaural time difference threshold measured with the phones attached to the panbone of the lower jaw of the dolphin [4]. The two practically coincide with the theoretical time resolution constant of the dolphin click and slightly better than the estimate of the dolphin’s auditory time resolution of 20-30 µs [5].

The larger angle between the transducer, the higher amplitude of the clicks should be for the dolphin to locate the transducer of a primary click (figure 3). The dolphin responded to a change in the click amplitudes like a directional receiver with the beam pattern described by the plot in figure 3. If the dolphin directed the acoustic axis of the receiving beam pattern at one of the transducer, the plot describes a half of the receiving pattern with the ordinate being the acoustic axis. If this was the case then our results are very close to Au and Moore measurements at 120 kHz [1]. If the dolphin made a decision when the acoustic axis was directed somewhere between the transducers, then the receiving beam pattern was even narrower.

FIGURE 3. The threshold amplitude of the clicks, (relative to the absolute threshold for a single click), as a function of the angle between the transducers.

The extremely small threshold angle between the transducers of 0.4°, and the threshold delay between the primary and delayed clicks of 17-18 µs, as well as the directionality of the auditory system can hardly be attributed to the binaural hearing phenomena known for the human’s auditory system. The dolphin’s auditory system at high frequencies performed simply as a directional receiver rather than the omnidirectional one characterized by the “echo suppression phenomenon”. As far as the active sonar is concerned the dolphin always knows the target’s direction just because the transmission of the echolocation clicks is directional. The directional hearing makes the target localization more accurate.

REFERENCES

Three-dimensional Sound Image Localization by Interaural Differences and the Median Plane HRTF - Part I. A New Sound Localization Method Based on Lateral and Vertical Angle Perception -

K. Iida\textsuperscript{a}, M. Itoh\textsuperscript{b}, E. Rin\textsuperscript{b}, and M. Morimoto\textsuperscript{b}

\textsuperscript{a}Multimedia Solution Labs, Matsushita Comm. Ind. Co., Ltd, 600 Saedo, Tsuzuki, 224-8539 Yokohama, Japan
\textsuperscript{b}Environmental Acoustics Lab, Fac. of Eng., Kobe University, Nada, 657-8501 Kobe, Japan

Morimoto and Ando have shown that 3-D localization can be attained, if HRTF are accurately simulated. However, a numerous number of HRTF are required to represent the entire 3-D auditory space. Furthermore, decrease of the localization accuracy due to the individual difference of HRTF remains unsettled. One of solutions of these issues is to reduce the number of HRTF to be used. Morimoto and Aokata have reported that the lateral and vertical angles of a sound image are determined by the binaural disparity cues and spectral cues, respectively. Moreover they suggested that the spectral cues are same in any sagittal plane. In this study a new 3-D localization method by simulating interaural differences and the median plane was proposed. Localization tests were performed to examine the method. The results indicate that the perceived lateral and vertical angles agree with the simulated ones.

INTRODUCTION

Morimoto and Ando [1] have shown that 3-D localization can be attained, if HRTF are accurately simulated. Almost of the recent studies on 3-D localization are based on this principle. However, a numerous number of HRTF are required to represent the entire 3-D auditory space. Furthermore, decrease of the localization accuracy due to the individual difference of HRTF remains unsettled. One of solutions of these issues is to reduce the number of HRTF to be used. Morimoto and Aokata [2] have demonstrated that the lateral angle $\alpha$ and vertical angle $\beta$ of a sound image are determined by the binaural disparity cues and spectral cues, respectively, by using the coordinate system shown in Fig. 1. Moreover, they suggested that the spectral cues are same in any sagittal plane. In this study, based on these findings, a new 3-D localization method by simulating interaural differences and the median plane HRTF is proposed, and its localization accuracy was examined.

METHOD OF LOCALIZATION TESTS

Localization tests were carried out to confirm the localization accuracy of the proposed method. The HRTF of the subjects in the upper median plane were measured at every 30° from the front to the rear. Interaural differences were also measured at the four lateral angles (0°, 30°, 60°, 90°) in the right side of the horizontal plane. The source signal is the band-limited white noise (280Hz - 11.2kHz). Stimuli were prepared by convolving the noise with the measured subject’s own median plane HRTF, and adding interaural differences to them. Twenty-eight kinds of stimuli (7 HRTF x 4 interaural differences), which simulated the sound sources in the upper hemisphere, were presented to the subjects by the sound field simulation system through near-ear loudspeakers. The duration of the stimuli was 1s. The task of the subjects was to mark down the perceived azimuth and elevation of the sound image on the recording sheet. They were transformed into $\alpha$ and $\beta$ after the experiment. Subjects were four males with normal hearing sensitivity.

RESULTS AND DISCUSSIONS

The subjects reported that they perceived all sound images outside the head. Figure 2 shows examples of...
the responses (subject: IT, $\alpha = 0^\circ, 30^\circ, 60^\circ$, $\beta = 0^\circ, 90^\circ, 180^\circ$). The circular arcs denote the lateral angle $\alpha$, and the straight lines from the center denote the vertical angle $\beta$. The simulated $\alpha$ and $\beta$ are shown in bold lines. The intersection of two bold lines indicates the simulated direction. This figure shows that the perceived lateral angles almost agree with the simulated ones. The localization accuracy, however, decreases to some extent for $\alpha = 60^\circ$. With the vertical angle, the localization accuracy is high at the front and rear of the subjects, but relatively low at the above. This behavior is similar to the responses for the real sound source [3]. The perceived vertical angles agree with the simulated ones even in the cases of the lateral sound sources. In addition, no front-back confusion is observed. These were common behavior to other subjects.

Furthermore, the mean localization error, $e$ was obtained by Eq. (1).

\[ e = |R - S|, \]  

where $R$ is the perceived angle, and $S$ is the simulated angle. Table 1 shows the errors of the lateral and the vertical angles for each lateral angle of the simulated sound source. The larger simulated lateral angle becomes, the larger lateral angle error becomes. These errors show the similar tendency to the jnd of horizontal plane localization [3]. The vertical angle error is almost the same as the localization error in the median plane[3]. However, the vertical angle error shows relatively large at $\alpha = 60^\circ$. The reason seems to be that a change of vertical angle along a circular arc in the lateral sagittal plane is sensitive when its radius is small.

<table>
<thead>
<tr>
<th>Simulated Lateral Angle $\alpha$ (deg.)</th>
<th>0</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lateral Angle Error(deg.)</td>
<td>1</td>
<td>7</td>
<td>16</td>
<td>23</td>
</tr>
<tr>
<td>Vertical Angle Error(deg.)</td>
<td>15</td>
<td>13</td>
<td>21</td>
<td>-</td>
</tr>
</tbody>
</table>

**CONCLUSION**

The localization tests demonstrate the realization of 3-D localization by simulating interaural differences and the median plane HRTF.

**REFERENCES**

Three-dimensional Sound Image Localization by Interaural Differences and the Median Plane HRTF
- Part II. Effects of ITD and ILD on Perception of Lateral Angle -

M. Itoh\textsuperscript{a}, K. Iida\textsuperscript{b}, E. Rin\textsuperscript{a} and M. Morimoto\textsuperscript{a}

\textsuperscript{a}Environmental Acoustics Lab, Fac. of Eng., Kobe University, Nada, 657-8501 Kobe, Japan
\textsuperscript{b}Multimedia Solution Labs, Matsushita Comm. Ind. Co., Ltd, 600 Saedo, Tsuchiura, 224-8539 Yokohama, Japan

In this paper, localization tests were performed to examine the effect of each interaural difference on the perception of the lateral angle. When the HRTF on the median plane and ITD were simulated, subjects perceived both of lateral and vertical angles accurately. On the other hand, when the HRTF on the median plane and ILD were simulated, the subjects' responses did not agree with the simulated direction. They perceived lateral angles near the median plane regardless of the simulated lateral angles. Moreover, the listeners sometimes perceived unnatural sound images. These results mean that ITD is dominant on the perception of the lateral angle. This agrees with the findings by Wightman et al.

INTRODUCTION

In Part I [1], it is shown that the sound image of any direction could be localized using the HRTF in the median plane and interaural differences. The interaural differences consist of interaural time difference (ITD) and interaural level difference (ILD). In this paper, localization tests were performed to examine the effect of each interaural difference on the perception of the lateral angle.

LOCALIZATION TESTS

The method is the same as Part I, with the exception that either ITD or ILD is simulated as interaural difference.

Figure 1 shows examples of the responses to the stimuli simulated by using the HRTF in the median plane and ITD. Similarly, Figure 2 shows the responses to the stimuli simulated by using ILD for ITD. Both results were obtained from subject IT. Table 1 and 2 indicate the mean localization error when ITD and ILD were simulated, respectively.

Figure 1 shows that perceived lateral angles almost agree with the simulated lateral angles when ITD was simulated. The mean localization errors are about the same as the test in which both interaural differences were simulated (see Table 1 in Part I). On the other hand, Figure 2 shows that perceived lateral angles do not agree with the simulated lateral angles, and all responses distribute near the median plane \((\alpha = 0^\circ)\) when ILD was simulated. Accordingly, the mean localization error becomes larger as ILD increases.

These results agree with Wightman’s [2] that ITD is dominant to other cues when the signal contains low frequency components. When ILD was simulated, subjects localized near the median plane because ITD was zero.

Concerning vertical angle, perceived angles almost agree with the simulated ones in both cases.

In each localization tests, 840 responses were obtained in total. Unnatural sound images, such as multiple images, appeared 6 and 41 times when ITD and ILD were simulated, respectively. The reason can be considered as follows: The simulation system used in the tests does not reproduce the frequency dependence of ILD that is small at lower frequencies and is large at higher frequencies. Then the subjects perceived unnatural interaural differences. Furthermore, unnatural sound images hardly appeared in the tests in Part I where both interaural differences were simulated. From these results, it can be inferred that ILD contributes for the sound image localization to some extent.

CONCLUSION

Localization tests in which ITD and ILD are simulated separately indicate that ITD is dominant on the perception of the lateral angle.

REFERENCES

### Table 1. Mean localization error when ITD is simulated

<table>
<thead>
<tr>
<th>Source angle α (deg)</th>
<th>Perceived angle α</th>
<th>Perceived angle β</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>30</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>60</td>
<td>22</td>
<td>16</td>
</tr>
<tr>
<td>90</td>
<td>29</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 2. Mean localization error when ILD is simulated

<table>
<thead>
<tr>
<th>Source angle α (deg)</th>
<th>Error</th>
<th>Perceived angle α</th>
<th>Perceived angle β</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>60</td>
<td></td>
<td>33</td>
<td>16</td>
</tr>
<tr>
<td>90</td>
<td></td>
<td>67</td>
<td>-</td>
</tr>
</tbody>
</table>

**FIGURE 1.** Examples of the responses to the stimuli simulated by using median plane HRTF and interaural time differences. Bold lines indicate simulated lateral angle α and vertical angle β.

**FIGURE 2.** Examples of the responses to the stimuli simulated by using median plane HRTF and interaural level differences. Bold lines indicate simulated lateral angle α and vertical angle β.
HRTF interpolation using a narrow-band envelope and instantaneous phase components

H. Nomura\textsuperscript{a}, S. Yamanaka\textsuperscript{b}, M.Matsumoto\textsuperscript{b} and M. Toyama\textsuperscript{b}

\textsuperscript{a}Department of System Engineering, University of East Asia, Shimonoseki, Yamaguchi, Japan
\textsuperscript{b}Department of Informatics, Kogakuin University, 2665-1 Hachioji, Tokyo, Japan

It is shown that directional HRTF records can be rendered by interpolation of the instantaneous envelopes and phase in an analytic signal representation. The rendering precision of the log envelopes was 20 dB on the average for both lateral and ipsilateral source-directional conditions in the horizontal plane. Measuring the HRTFs in all source directions, even in the horizontal plane, is not a realistic task in a practical application such as an immersive audio system. The authors therefore formulate the HRTF impulse records in the complex-time domain in order to HRTF representation suitable for rendering or morphing. After taking 1/4-octave sub-band records, we adopt the analytic signal form to decompose each 1/4-octave band signal into its envelope and phase components. We can approximate the linear instantaneous phase in the complex time domain to most significant sinusoidal component. This modeling precision was over 20 dB on the average (SDR).

**INTRODUCTION**

Networking audio technology, including immersive audio systems, has recently been developed [1-2]. The authors have also been developing an ISFN (which stands for interactive-sound-field network), as shown in Fig.1, under the support of the Telecommunications Advancement Organization of Japan (TAO).

Networking audio technology including 3D-sound-field control normally requires 3D-directional HRTF records for sound image localization. However, measuring many 3D-directional HRTF records is not a realistic task. To overcome this problem, we have developed a rendering method that uses parametric representation for 3D-directional HRTF records [3-4]. An HRTF record is a representation in the complex frequency domain. In the current study, the authors formulate the HRTF impulse response records in the complex-time domain based on the analytic signal form using the instantaneous envelope and phase components [5]. We extract the instantaneous phase which can be approximated to be almost linear for every 1/4-octave sub-band record. Thus, this type of parametric representation will make it possible to render the HRTF records.

**RESPONSE MODELING IN THE COMPLEX-TIME DOMAIN**

Suppose that $h(n)$ is a directional HRTF impulse response record. Its analytic signal $\hat{h}(n)$ can be defined by using the instantaneous envelope $|\hat{h}(n)|$ and phase $\theta(n)$ according to the Hilbert transform. If we introduce the minimum and all-pass phase decomposition, which was originally devised in the complex frequency domain for transfer function analysis, into the instantaneous phase in the complex-time domain, we can get the formula: \[\hat{h}(n) \equiv \hat{h}(n)| e^{j\theta(n)} \equiv \hat{h}(n)| e^{j\theta_\text{min}(n)+j\phi(n)}\] (1)

Here we define the modeled-HRTF impulse response $\tilde{h}(n)$ as \[\tilde{h}(n) \equiv \tilde{h}(n)| \left[A \cos(\theta_\text{min}(n)+\omega n) + B \sin(\theta_\text{min}(n)+\omega n)\right] \] (2)

where $\theta_\omega(n)$ is represented as a linear-phase with a modeling parameter set composed of a single angular frequency $\omega$ and constant initial phase $\phi$ for $A = \cos(\phi)$ and $B = \sin(\phi)$. These modeling parameters can be determined by solving Eq. (2) according to the least-square-error criterion [6].

**MODELING-ERRORS EVALUATION**

We modeled the HRTF impulse response records of MIT [7]. Modeling precision can be numerically evaluated by signal-to-deviation ratio (SDR) as \[SDR = 10 \log_{10} \frac{\sum_{n=0}^{N-1} \left| h(n) - \tilde{h}(n) \right|^2}{\sum_{n=0}^{N-1} \left| h(n) \right|^2} \] (dB),

where $N$ denotes record length. If SDR increases, the modeling error decreases. Figure 2 shows the modeling precision (defined as SDR) for the 45-degree-HRTF impulse response record of the left ear at every 1/4-octave band from 53 to 19,000 Hz. It is clear that SDR decreases (i.e., modeling error increases) as the frequency of interest increases.

Figure 3 shows the directional variations in the modeling parameters for the left-ear HRTF impulse response record at 500 Hz-centered of the 1/4-octave band ($\alpha$: frequency of...
the linear-phase; b: initial phase; c: A and B; and d: logarithmic envelope).

**RENDERING HRTFs BASED ON INTERPOLATION USING MODELED HRTF**

The parameter variances due to changes in the sound source directions are shown in Fig. 3, where some are smooth and some are a little too sensitive. In this research, we try to interpolate 15 and 30-degree HRTF impulse responses from 0 and 45 degrees, in the same way as the interpolation at 45-degree angular intervals, as shown in Fig.4. Figure 5 shows precisions of log envelope interpolation were 20 dB on average (SDR) for HRTF impulse response records at 500 Hz for both ears.

**SUMMARY**

A method using the analytic signal form for HRTF rendering of parametric representation of HRTFs in the complex time domain was developed. The modeling precision was 30 dB on average for frequency bands below 700 Hz, and rendering precision (SDR) due to interpolation of the log envelope was 20 dB on average. These precisions are being now subjectively evaluated by listening tests. We introduced the magnitude and phase concept into the analytic signal form in the complex-time domain as well as the concept in the transfer function analysis in the complex frequency domain. This complex-frequency time analysis will reveal new transfer function signature.

**REFERENCES**

7. MIT website: http://sound.media.mit.edu/KEMAR.html

**FIGURE 2.** Modeling errors in SDR for the 45-degree-HRTF impulse response record of the left ear at every 1/4-octave band from 53 to 19,000 Hz

**FIGURE 3.** Modeling parameters variances due to the changes in source directions, (a) frequency making the linear phase, (b) initial phase, (c) A and B, and (d) Logarithmic envelope

**FIGURE 4.** The directions of interpolated HRTF impulse response

**FIGURE 5.** Interpolation precision (in SDR) for HRTF impulse response of 500 Hz-centered 1/4 octave-band.
Acoustical design and requirements in an aquatic park

V. Tarabusi, L. Tronchin

Department of Energetical, Nuclear and Environmental Control Engineering,
Bologna University, 40136 Bologna, Italy

During architectural design, it is nowadays quite important to analyse from an acoustical point of view the behaviour of spaces devoted for music, like Opera Houses and Concert Halls. Nevertheless, it become quite important to investigate the acoustical aspect of other environments that are not particularly devoted to music or speech, especially in consideration with other conditions which could characterise the good working of the structure. In this work the main acoustical problem was to solve the requirements of a good acoustics in three different halls that are strictly in conjunction with rooms with different variety of fishes, included sharks. The problem has been carried out studying both acoustical characterisation of the special effects in the rooms and the special needs of the sharks. The solutions coming from numerical simulations have been therefore adapted with the requirements of a quite high transmission loss of the walls, especially at low frequencies.

INTRODUCTION

On the Adriatic Sea there are many different attractions that characterize that particular area. In summer season millions of tourists are crowded the beach in the northern Adriatic Sea, especially from Ravenna to Pesaro. Among all the towns on that area there were no aquatic parks, apart of the new one that’s going on close to Cattolica.

This new attraction will be characterized for difficult conditions. In some halls multimedia shows will be presented. Many different fishes, including sharks, will populate the swimming pool. The acoustical requirements had to be added to other supplies, concerning safety, lighting, indoor air quality. Different spaces were therefore studied.

SOUND QUALITY FOR MULTIMEDIA SHOWS

The acoustical supplies for multimedia halls extend from the necessity of good sound envelope for special effects, to low reverberation time, to a dynamic range of sound intensity, to a high insulation of the walls.

These requirements entail peculiar condition in the acoustical characteristics of the partitions, like floors, walls and doors. Besides, in the case of aquatic park vibrations have to be limited in order to avoid dangerous effect with sharks, which are particularly carefully with low-frequencies noise.

THE CASES

In 1990s the town of Cattolica decided to restore a dismissed area and to build on a new attraction, following other Italian and European experiences. The project intended to re-utilize the buildings and to transform them into a set of swimming pools ad halls with multimedia shows, ranging from the “big bang hall”, to the “thunderstorm hall” and the “square hall”.

![Wireframe model of big bang and square hall](image)

Each of them had peculiar acoustical requirements, both of sound quality and acoustical and vibrational insulation.
LINEGUIDES IN THE ACOUSTICAL DESIGN

In the acoustical design a proper electro-acoustical system was studied for many of the enclosed space. The position of loudspeakers was chosen with simulation of the characterization of sound surrounding special effects. Therefore the acoustical treatment of the rooms was limited to reduce the reverberation time to low values, (around 0.6÷0.8 s), to increase clarity and definition index to high values, whereas the spatial parameters were emphasized, as well as intelligibility. This allowed a better realization of special sound effects. In the square hall, where no multimedia shows were planned, no electro acoustical systems were positioned; therefore reverberation tail was slightly higher.

FIGURE 2 big bang hall: EDT; in Italic, adopted solution

FIGURE 3 storm hall: STI; in Italic, adopted solution

The insulation requirements were usually achieved by using light plasterboard partitions, with high transmission loss in a short thickness. In conjunction with a specific calculation of flanking transmission, specific solutions were given to the architectural designer.

ACKNOWLEDGMENTS

The authors wish to thank Giorgio Guidotti for his help during the drawing of the wire-frames of the halls at the computer desk.

REFERENCES

2. CIARM, Guidelines for acoustical measurements inside historical opera houses, Ferrara, 1999
A common problem of head-related transfer function (HRTF) simulation using numerical methods is the complexity of the models and the highly individual nature of the modeling problem. On the other hand, spherical head models sometimes used in spatial audio lack the needed idiosyncratic behavior of HRTFs. In this paper a solution to these problems is approached by approximating the HRTFs using simplified numerical models. The modeling scheme is based on a simplified geometrical model of a subject’s head and shoulders, which still preserves the important individual properties. The HRTFs are solved with the Boundary Element Method (BEM) in the frequency domain from 50 to 4000 Hz, applicable for speech bandwidth.

INTRODUCTION

This paper addresses the issue of creating low-complexity head-related transfer functions (HRTF) based on numerical modeling of a highly simplified head and torso system. An HRTF describes the free-field transfer function from a point in space to a point in a human or dummy head ear canal. In virtual acoustics and spatial hearing the HRTF is an essential concept, because it can be efficiently realized using digital filters [1-2]. One of the drawbacks of HRTF-based technologies, however, is the highly individual nature of the responses. In the literature, the generalized systems (using non-individualized HRTF's) for 3-D audio have usually been reported to perform worse than those individually calibrated. Both coloration and localization inaccuracies (for example front-back confusion) occur, including inside-the-head localization in headphone listening. The optimal way to overcome this problem would be to use each person's measured HRTF. Unfortunately the measurement procedure is rather complicated (and time consuming). Therefore it is desirable to try to find ways to customize the virtual 3-D audio reproduction system to fit to the user. Some studies have been done in this field, attempting to, for example, customization of HRTFs, or by creating simplified generic filter design models.

Detailed numerical modeling of HRTFs has been proposed by some researchers (see, e.g., [2-4]) as a solution for obtaining HRTFs. One relatively unexplored alternative, however, is to look at the problem via simplified numerical modeling of the head and torso system (HATS) that is the main contributor to the HRTF. In particular, if a perceptually well matching numerical model could be created using as little information (or, as few dimensions) as possible, it would open possibilities to customize the reproduction system in a fairly simple and convenient way. Some commercially available dummy heads are indeed based on simplified models of the head and torso. The aim of the current paper, however, is to discuss individualization of HRTFs based on the models, which is not achievable using constructed dummy heads.

NUMERICAL MODELING OF HEAD AND TORSO SYSTEM

The acoustical models including the HATS geometry tend to become computationally heavy. To avoid this problem we have created a simplified model of HATS. The transfer functions have been numerically calculated using the Boundary Element Method (BEM) of I-Deas Vibro Acoustics [5]. The research goal is to develop a modeling scheme that would make possible individualization HRTF synthesis using only handful parameters for model creation and which would be computationally effective. Also, if an effective modeling scheme can be treated with a few parameters, it enables systematical investigation of the correspondence between the geometric dimensions and the HRTF.

The selection of the parameter set was based on results from psychoacoustics and preliminary simulations. The parameters and how they generate the model are shown in Figure 1.
FIGURE 1. A simple HATS model is defined by five size parameters.

RESULTS

The results achieved through the current modeling scheme appear to possess promising properties. If the simulated and measured HRTF of the same individual are compared, it can be seen that many of the global features are common (see Figures 2 and 3). If the results for different individuals are compared, the corresponding models give results that differ from each other in a consistent way. Differences between models prepared of different target person HRTFs have been shown. A more detailed comparison should, however, be carried out using perceptual listening experiments.

FIGURE 2. Magnitude responses of a set of measured HRTFs (male subject, 0° elevation).

FIGURE 3. A simulated set of HRTF responses corresponding to the measurements shown in Fig. 2.

ACKNOWLEDGMENTS

While conducting the work presented in this paper crucial help and advice has been provided by the people at Laboratory of Acoustics and Audio Signal Processing of Helsinki University of Technology and the colleagues at the Speech and Audio Systems Laboratory of Nokia Research Center.

REFERENCES