ATR CALL: A speech perception/production training system utilizing speech technology

Reiko Akahane-Yamada, Hiroaki Kato, Takahiro Adachi, Hideyuki Watanabe, Ryo Komaki, Rieko Kubo, Tomoko Takada, Yuko Ikuma, Hiroaki Tagawa, Keiichi Tajima, and Hideki Kawahara

1 ATR Human Information Science Laboratories
2 Wakayama University

Abstract

A computer-based instruction system for language learners, dubbed ATR CALL, is described. This system is a collection of speech perception, production and/or comprehension training tools, which have been mainly developed for second language learners. Each component focuses on the acoustic-phonetic, prosodic, lexical, or semantic decoding level of spoken language. Speech analysis tools for language learners are also provided. State-of-the-art technologies are used for each tool. The current system consists of a Japanese-teaching system as well as an English-teaching system.

1. Introduction

There are thousands of mutually unintelligible languages in the world, of which typically only one is acquired as the native language. While one’s native language is acquired quite effortlessly, once the native language is established, it is often difficult to master a second language. In our laboratory, we have been undertaking basic research on second-language acquisition by human learners, through extensive training experiments, under various environments (in the laboratory, at school, over the Internet, etc.). In order to acquire data from many participants efficiently, we have developed the “ATR CALL” system, and have been utilizing it in our research. The system consists of three parts, (1) Multimedia instructional tools, (2) Speech analysis tools, and (3) Training tools. In this paper, we describe each part.

2. Multimedia Instruction Tools

In order to teach the phonetic system of the target language, a multimedia phonetic symbol chart has been developed, by utilizing movies (human face and three-dimensional animation), and speech sounds. Furthermore, for interactive use of the chart, speech analysis and recognition programs are also installed in the chart, so that learners are able to try to produce the target word, and analyze and/or evaluate their own pronunciations.

3. Speech Analysis Tools

In speech production training, visualization of acoustic properties is one possible type of automatic feedback regarding the goodness of learners’ pronunciation. However, in order to understand acoustic properties, knowledge of acoustic phonetics is necessary. Thus, tools that interactively analyze learners’ speech and show spectrographic representation with or without formant tracking result overlaid, and prosodic properties have been developed. In addition, some of the tools represent formant frequencies in a real-time manner, and plot them on an F1-F2 plane or an F2-F3 plane.

4. Training Tools

4.1. Perception Training

Perception Training consists of two parts, (1) identification training of phonetic contrast, and (2) syllable counting training. Identification training uses mainly a two-alternative forced-choice (2AFC) task in which two words comprising a minimal pair contrasting in English phonemes (e.g., /r/-/l/, /b/-/v/, etc.) or Japanese moraic structure (e.g., /seto/-/se:to/, /furin/-/fu:rin/, etc.). Many minimal pairs are presented in each training block. On each trial, a member of a minimal pair is played over the headphones or the speaker. At the same time, orthographic representations of both members of the minimal pair are presented on a screen. Learners have to choose one of the words. Immediate feedback about the response is provided. Training stimuli vary in talkers, phonetic environment of the target phoneme, speech style (word in isolation or word in carrier sentence), and context (word in semantically meaningful sentence or word in semantically neutral sentence). Note that more than two choices are presented in certain conditions.
4.2. Production Training

Feedback in a production training has been a challenging issue. In our system, three types of automatic feedback for learners have been pursued: (1) visualized acoustic properties, (2) pronunciation scores produced by an HMM-based speech recognition system, and (3) model speech with the phonetic attributes of a model talker and with the voice quality of the learner, so-called “mosaic speech synthesis”.

In (1) visualized acoustic properties, a target word is played back and presented on a CRT monitor in an orthographic form. Learners reproduce the word, and immediately after the production, the spectrographic representation of the speech by a model talker and the learner are displayed side-by-side together with formant-tracking results. Learners try to produce the same word until they were satisfied.

In (2) pronunciation scoring, each utterance was transformed into a sequence of feature vectors, each consisting of 21 MFCC-based values, including MFCC deltas and delta energy, calculated for a 20ms frame of speech at a 10ms frame rate. The HMM for each phoneme category was assigned 3 states, each state consisting of a mixture with 8 Gaussian components. Posterior probabilities were directly calculated at the segment level from the segment-level log-likelihoods. The given values are normalized into scores from 0 (worst fit) to 100 (best fit).

In (3) mosaic speech synthesis, auditory morphing using STRAIGHT [1,2] was utilized. This system generates synthetic speech with the correct pronunciation but with the learner’s voice quality, based on a model utterance and the learner’s utterance (Figure 1.2). Importantly, the quality of the generated synthetic speech is quite high and it sounds as if it was a production by a real person.

4.3. Vocabulary Training

Vocabulary training is based on word translation, in which a two-alternative forced-choice task was used. An orthographic representation of a word in one language is presented on the computer display. Simultaneously, orthographic representations of word choices in another language are presented. The learner’s task is to decide which is the better translation and to respond by clicking that word. Immediate feedback about the response is provided.

5. Conclusion

The current ATR CALL system is a collection of various instruction and training tools for second language learners. Since each tool is developed for use in basic research that examines the process of speech perception and production learning by humans, the training effects of the system have been evaluated already, which is an advantage of this system. Further development of each tool and the instructional design is necessary to develop the current ATR CALL system into an integrated second language learning courseware.

6. Acknowledgements

Authors are grateful for Mr. Koji Kikuchi, Mr. Tomohiro Kishimoto, Ms. Naoko Ohtuki, and Mr. Tatsuya Ueda for their technical support. This research was supported in part by the Telecommunications Advancement Organizations of Japan.

7. References