Speech Assistant System

László Czap
Department of Automation and Infocommunication
University of Miskolc
czap@uni-miskolc.hu

Abstract

A speech assistant system has been developed at the University of Miskolc in cooperation with the University of Debrecen granted by the European Union. The project aims to help training of deaf and hearing impaired people to speak. The idea of the project has come from a three-dimensional head model for articulation presentation, called "talking head" developed by the University of Miskolc, and an audio-visual transcoder for sound visualization developed by the University of Debrecen.

Index Terms: talking head, visualization, speech training

1. Introduction

Integration of the results achieved so far and the knowledge accumulated in research works, as well as solving other basic and applied research tasks have opened the door to develop an application that supports speech understanding of deaf people and teaching them to talk more effectively than known methods.

Theoretic basis of the research is that studies have shown that integration of acoustic and visual modalities in brain is optimal for maximum intelligibility. In case of deaf people, the stronger the acoustic signal distortion, the more they rely on the visual signal. We expect that this can be met if the visual signal is not the image of the speaker, but a visual transcript of the sound. Some part of technology is language independent, and can easily be adapted to foreign languages, thus the involvement of foreign partners can contribute to the implementation of better society that is one of the three objectives (Key Objectives) appeared in the Program Horizon 2020, supporting by an FP8 project.

2. Audiovisual Transcoding

The concept of audio-visual transcoding is to shape the voices into abstract visual signals – videograms - by coding acoustic features into graphics. We have defined two types of videograms: a matrix representation depicting momentary voice, and a column diagram visualizing whole lexical units (word, sentence). In matrix depiction voice is represented by predetermined number of squares assigned to discrete frequency components. The sizes of squares depend on the sound intensity.

According to the frequency component the squares are color-coded and position-coded: the bass sounds are imaged by larger wavelength colors (red), treble sounds are presented by short wavelength colors (blue, magenta). The most important audio range of 125 Hz - 8000 Hz is transcoded. It is divided into fixed frequency bands, which are plotted on the display divided into appropriate sectors.

This partition allows the visual representation of momentary voice by squares. Therefore the voice frequency component is not only color-coded but also position-coded. It provides visual differences between the tropes, thus distinction between the words of a similar sound shape becomes possible.

By a microphone and sound card the sound is recorded at 16 kHz sample frequency, which is digitalized by the computer, and then using Fourier transform they are shaped and filtered into frequency-specific signals. Audio-visual transformation of the voice is real-time and is therefore suitable for mapping continuous speech.

Above this visualization that depicts the momentary transcription of a 40 msec frame, training of hearing impaired people to speak needs the visualization of a whole lexical unit being practiced. For this purpose we have transformed matrices to columns, enabling the representation a number of frames. During speech training lessons, comparing the current utterance of the trainee to the pre-recorded reference one is being demanded. A neural network based voice activity detector is trained with 4.5 hours of voices of more than 300 speakers. Acoustic samples for voice activity detection have been recorded by using different microphones, sound cards and PC-s, in a variety of noisy office, laboratory and home environment. After recording the current utterance the beginning and end of speech is detected. Its transcoded column representation is displayed in a window above the reference one's to be able to compare them. For column representation of voice we have introduced a third code above the position and color. A rectangle varies not only its size but its intensity highlighting more important segments.

3. Refining the Performance of Talking Head

Dynamic modelling of articulation is a basic research. Static properties of phonemes of speech can be found in various audio albums. However, data found about dynamic changes of the parameters can only be used for a given sample word. Regarding the non-visible (or partially visible) movements of speech organs, only few data are available. The lip-shapes are provided by a database that we created for audio-visual speech recognition, thus visemes requires only minimal adjustments. For tongue movements, it is very difficult to obtain tracking data for the realistic head model, it was not sufficient motion based animation found in voice albums. For presenting the voice formation, correction of tongue movements is needed in the head model with transparent face by the involvement of surdo-teachers and researchers.

The proposed client-server and Internet based concept allows to practice anytime and anywhere. Computers (laptop, desktop, tablet) and smartphones having different display size may require different enlargement of the display patterns.

In case of smartphones, close proximity and movement of the mouth should be plotted on the display, but all the details of the articulation of the whole head must have seen on a computer display. Tests are needed to determine the preferred view.

Making the articulation of scalable depth of the sign language interpreter stronger than natural lip movements of the average intensity is one of the features of the developed talking head. For determining the most appropriate pace of samples tests must be performed, the usual practice in daily life, or slightly slower speech rate, tests are necessary. Samples must be produced with different speech rates an indicator of the progress is the speed of speech production.

Detailed observation of articulation and sound visualization in a speech segment helps ex-post freeze on the display. Articulation needs to be adapted to speech rate, since rapid speech features less approach the nominal value. Interpolation rules have to be able to manage the pace of speech as well. As a feedback, emotional charge associated to the head model can enhance the attractiveness. The server-side application serves the client applications via an Internet connection, registers the results, the evaluations and the selection of the next practice pattern. Continuous updates of feedback messages stored in server database may enhance the acceptance. The client-side application has to be developed for different operating systems (Windows, Linux, Android, Windows Phone). The client-side application may vary according to the target group's age and condition.

4. Suprasegmental features

Speech technology has improved a lot in recent years, and a sort of technically difficult problems, including use of the Internet or displaying right prosody, is now solved. The prosody analysis and visualization is essential for learning correct prosody. Adequate formation of the suprasegmental features is an important indicator of clarity and naturalness. The most striking feature of hearing impaired speech is the bad use of prosody. The prosodic parameters, intonation, rhythm proper emphasis, normalization, time warping are critical parts of displaying during the training. Visual perception of similarity and difference of the reference and the actual pronunciation should be easy for small children as well. This visual assessment is planned to be enhanced with an automatic response feature. The complexity of the task and the appropriate allocation of processes dictate the implementation of prosodic analysis of server-side and client-side display function, transmission of digital voice tag for the current quirk to the server and the result is performed online to the client.

5. Conclusions

This paper describes a speech assistant system that forms the base of a research project that aims at creating a Hungarian audio-visual speech aid, a cognitive development system, training hearing impaired people to learn speech articulation. Fine tuning of the probabilities for natural animation and avoiding mechanical, rule based repetition of gestures resulted from detailed study of speech production have been introduced as directions for human-like pronunciation. The speech training started in last September with a group of 30 children trained by this system and with a control group of 30 ones. First evaluation of the progress will be provided after one year of training by intelligibility tests. The system conception can be extended to other languages and correction of speech disorders.

6. Acknowledgements

This research was carried out as part of the TAMOP-4.2.2.C-11/1/KONV-2012-0002 project with support by the European Union, co-financed by the European Social Fund.

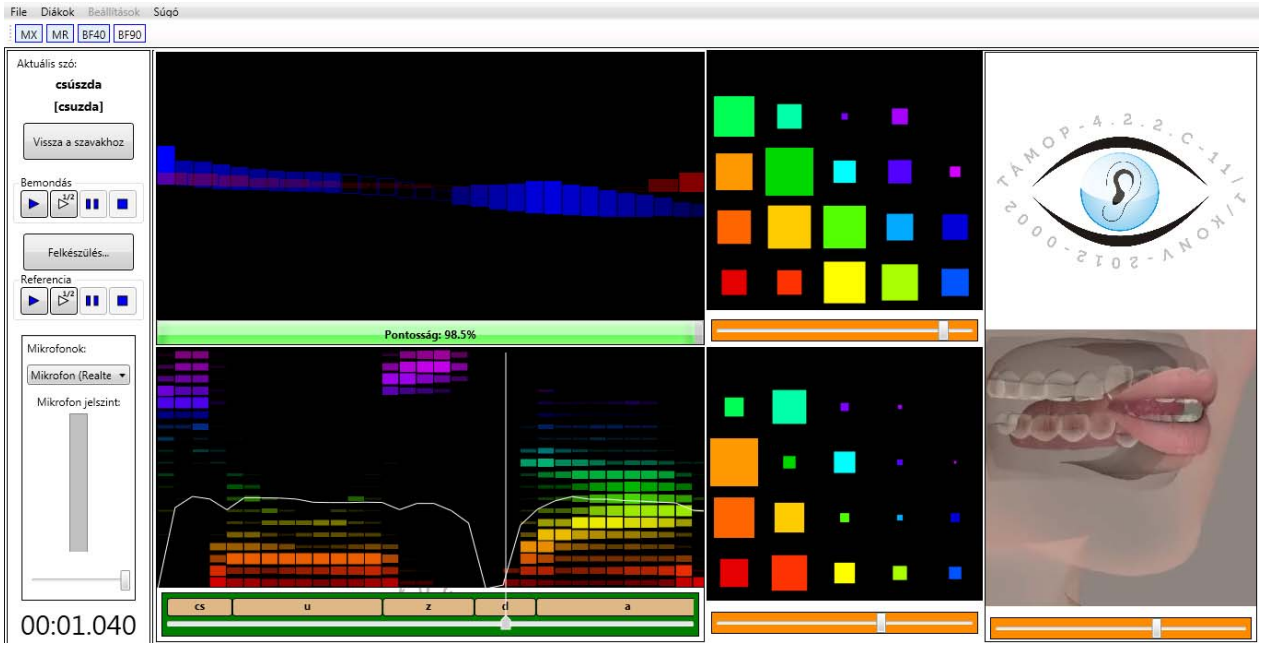

Figure 1: *Layout of the speech assistant system.*