A Multilingual, Multimodal, Speech Training System SPECO

K. Vicsi - P. Roach† - A. Öster‡ - Z. Kacic^ - F. Csatári* - A. Sfakianaki† - R. Veronik^*

*Budapest University of Technology and Economics, Hungary, †University of Reading, United Kingdom, ‡Kungl. Tekniska Högskolan, Sweden, ^University of Maribor, Slovenia
vicsi@alpha.ttt.bme.hu

Abstract

The SPECO Project was funded by the EU through the INCO-COPERNICUS program (Contract no. 977126) in 1999. In the frame of the project a system has been developed which is an audio-visual pronunciation teaching and training system for 5-10 year old children. Correction of disordered aspects of speech is done by real time visual presentation of the speech parameters, in a way that is understandable and interesting for young children, while remaining correct from the acoustic-phonetic point of view. The development of the speech by our teaching method is made mainly on the basis of visual information using the intact visual channel of the hearing impaired child. However during practice we use their limited auditory channel too, by giving auditory information synchronised with the visual one. This multi modal training and teaching system has been developed for languages of all the SPECO partners; these are English, Swedish, Slovenian and Hungarian [1, 5, 6].

1. Introduction

The general concept of the SPECO system and the comparison of it with the other computer-based pronunciation training tools were described for the last EUROSPEECH99 [1]. In the present paper we present developments made during the last two years: the general structure of the system, the speech pictures and the databases which had to be constructed for the four different languages.

The system is composed of two parts: one is a general language independent frame program, the **LANGUAGE INDEPENDENT DATABASE EDITOR AND MEASURING SYSTEM** and the other is the language dependent database files. Two databases have been collected and built into the system for all participating languages, one being for teaching and training vowels for hearing-impaired children, the other one for correction of misarticulated fricative and affricate sounds. In this way, the **VOWEL AND SIBILANT SUPPORT** for all participating languages has been developed. This system contains many exercises and provides a teaching facility for developing fluent and intelligible speech using a large reference speech vocabulary.

2. Language independent database editor and measuring system

The Hungarian partners have prepared the database editor and measuring system. This editor measures the different acoustic-phonetic parameters of the speech signal, and helps the users to build a language-specific vocabulary, and to construct the appropriate Vowel and Fricative Support.

ACOUSTIC MEASUREMENT MODULE

This module is a language independent system, which measures and visualizes the characteristic parameters of speech, such as auditory spectrum, spectrogram, spectrogram differences, sound energy, pitch, voiced - unvoiced detection, intonation, rhythm, etc., as described in an earlier paper [1].

RECORDING MODULE

The recording module of the editor gives the possibility of creating the reference speech vocabulary of the Vowel and the Sibilant Support. This vocabulary contains the reference examples: sound sequences, words, word-pairs and sentences. A special vocabulary has been constructed for each language. Carefully selected children using normal sound intensity produced the reference examples by pronouncing all of the phonemes in the expression in a correct way in silent surroundings (the signal-to-noise ratio min. 35 dB).

SEGMENTER MODULE

The place of the phonemes is marked on the auditory spectrograms (see Fig. 4.). Thus segmentation and labelling of the reference examples was necessary at the phoneme level. In the dialog box of the segmented module of the EDITOR, there are two input fields to filled in. One is for a free form string that will be visible on the spectrogram. The other will never be displayed. It is used to identify the different sounds for the automatic feedback.

An automatic segmentation was developed by the Hungarian partners and built into the segmenter module of the EDITOR to help the work of the partners [4].

GRAPHICAL EDITOR MODULE

The visual presentation of the acoustic parameters (called speech pictures) gives the possibility to the user to process the speech further on the basis of this visual presentation. A detailed examination was prepared to decide what scale of loudness, of pitch contour, of spectral distribution, etc, gives the most informative visual presentation (speech pictures) about these parameters, and which kind of background pictures helps child best to understand the speech pictures. Amusing background pictures were involved to emphasise the important parts of the spectrograms. These pictures are displayed on the screen together with the spectral points. The editor gives the possibility of positioning these background pictures relative to the spectrogram points of the reference example. Moreover the editor helps the developer to handle the calling cards (pictures to enable the selection from the menu) and the symbolic pictures of phonemes (each phoneme has a symbolic picture in the program).

3. Reference speech vocabulary

The speech pictures of the reference speech examples are clear and easy to read. The aim of the patient during therapy is the production of sound pictures (visual pattern) similar to these

references examples in the vocabulary. In the vocabulary all trained phonemes must be present in isolated form, in carefully selected sound sequences, in words, in sentences and minimal pairs, in different sound positions and systematic sound sequences. These samples of the vocabularies are used during the training as reference speech.
The reference auditory spectrograms together with the corresponding background pictures form another type of sound picture, as presented in the case of 's' in Fig. 1.

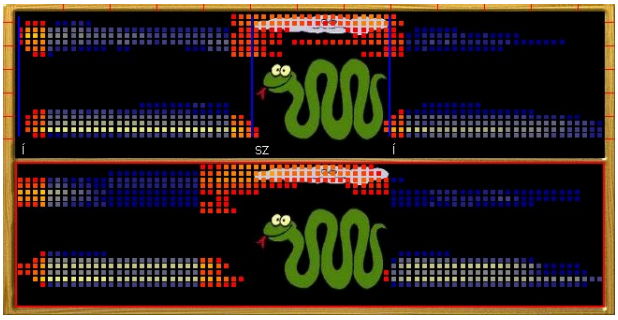

Figure 1. An example of the spectrogram-type sound picture of the fricative sound [s] in [isi] sound sequence

The background pictures emphasize the important parts of spectrogram. The task for the child in the case of sibilant sounds is to cover the clouds with the dots representing spectral energy, but not to cover the other parts of the background pictures.
Similar rules were constructed for vowels.

4. Multimodality

During the training the patients see the speech pictures of the reference speech and listen to the sound of it at the same time. Thus they use the human visual and auditory feedback during their speech learning in addition to the proprioceptive and tactile sensation.
One part of the training is the comparison of the actual pronunciation of the items with the stored references. Children first make the comparison visually. An automatic feedback has been developed to help comparison, based on distance calculation between the spectral components of the typical auditory spectrums or spectrograms and the actual speech.
One important feature of this system has been to find good distance measures that can help these comparisons between the spectral components of the reference and the actual speech. This distance measure must reflect the judgement of the listeners, i.e. the judgement of most of the people in the child's educational public and cultural life. [7]
The role of the automatic feedback is the most important in the teaching of severely hearing impaired children. This feedback provides a good opportunity for practice with no teacher but the system alone.

5. Collection of the corpus of children's speech, for statistical examination

All partners collected a corpus of children's speech, which contains examples of the pronunciation of fricatives, affricates and vowels that have been evaluated as "good", "acceptable" or "wrong". We refer to this database as 'the child speech database'. Good examples from the database were selected by listening tests [3]. These good examples were used to calculate the typical spectral patterns. The text material of this speech database contains all of the practiced fricatives, affricates and vowels in isolated form, in sound sequences, in words (70-80) and in sentences (from 5 to 10).
The Hungarian corpus contains recordings from 72 children between 5-10 years of age. The database description and the distance score experiments were presented at EUROSPEECH99 [2]. The English partner compiled a corpus with 36 children between 5-10 years of age in primary schools. The Swedish partner collected a corpus of 31 children between 7-9 years of age. In Slovenia a corpus of 50 speakers has been recorded.

6. Statistical models of the practiced phonemes

6.1. Typical auditory spectra of phonemes

The average spectra of the good examples from the 'child speech database' are named as typical auditory spectrum patterns. These patterns of all vowels and sibilants were calculated from the good examples of Hungarian, Swedish, English and Slovenian 'child speech databases', in isolated pronunciation and in words. A typical auditory spectrum pattern is calculated as shown below:

$$\overline{A}_n[phoneme] = \frac{1}{1-4}\sum_{i=3}^{I-2} A_i[phoneme]$$

$$\overline{A}[phoneme] = \frac{1}{N}\sum_{n=1}^{N} \overline{A}_n[phoneme]$$

where

$$A_i = \{a_1, a_2 .. a_f\},$$

f = 20 number of the critical band filters,

I = total number of frames in the phoneme,

N = number of speakers in children speech database

The calculated typical average spectrum patterns of sibilants are presented on the Fig. 2. The Hungarian partner made these calculations, and incorporated the spectrum patterns into the Support facility. All of these spectrum patterns are presented for the children, using one type of speech picture, shown as it is presented in case of the vowel 'u' in the Fig. 3. During the training the actual spectrum lines must fall within the two spectrum lines on the speech picture.

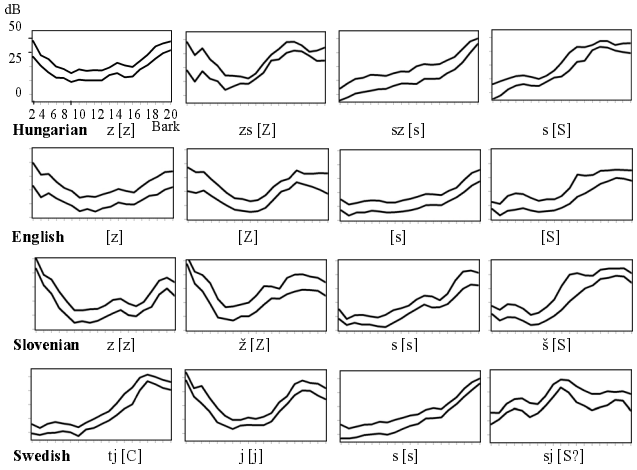

Figure 2. Typical auditory spectra of fricatives of four languages

6.2. The typical auditory spectrogram and the positioning of the background pictures

One part of the training is the comparison of the actual pronunciation of the items with the stored references. The auditory spectrogram of the stored speech of the reference speaker (the reference auditory spectrogram) is presented in the upper part of the screen and the spectrogram of the actual pronunciation on the lower part of the screen. Background pictures have been constructed and presented together with the reference auditory spectrogram to emphasize the important parts of it.

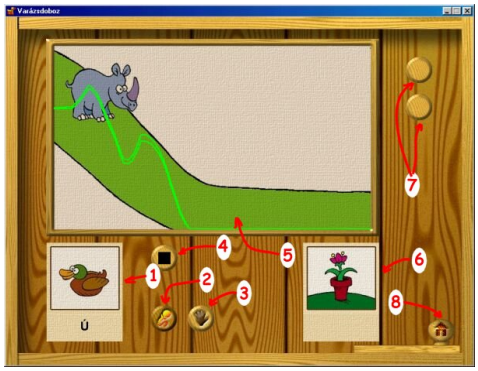

Figure 3. An example of the spectrum typed speech picture of vowel sound 'u' (5). Calling card (1). Cuff-links (2, 3, 4, 7, 8). Card to show the result of the automatic feedback with the flower. It is blooms in the case of good pronunciation (6).

The typical average spectrograms (upper part of Fig. 4.) help the construction and the positioning of the background pictures of each sound (lower part of Fig. 4.). The typical average auditory spectrograms have been calculated by averaging the energy values of spectrograms of the phoneme examples in words of the good and acceptable examples from the child speech database [2]. This typical average auditory spectrogram is not identical to the reference auditory spectrogram but the spectrogram of a good reference speaker is close to it.

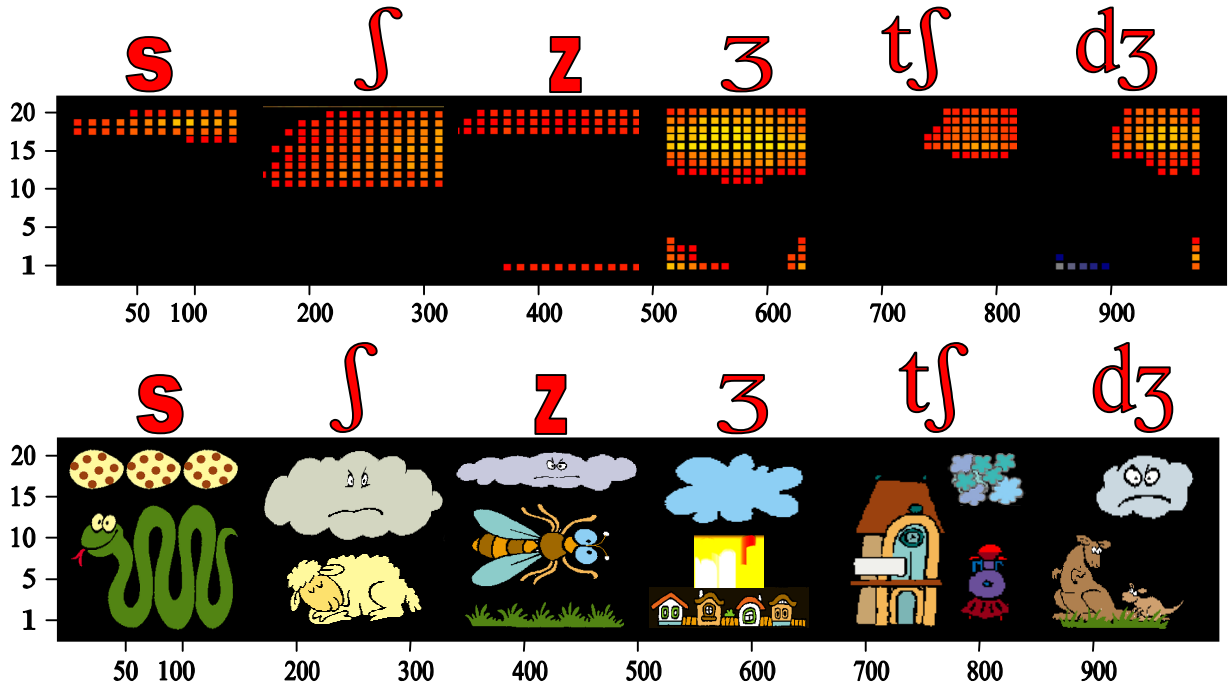

Figure 4. The typical average auditory spectrograms (upper) and the position of the background pictures (lower) for English sibilants. Text of the speech and picture database

The developed system is not only a measuring tool, but provides a method for teaching the child using this new multimodal program. All partners have been involved in the construction of the text and pictures of the database, i.e. the development of exercises for the Vowel and Sibilant Support from 'sound preparation' through to 'fluency'. The exercises are appropriate to the phonological structure of their languages, and traditional teaching practices have been taken into account in their construction.
The structure of the exercises is the following:

+SOUND PREPARATION
* loudness
* spectrum
* pitch
* rhythm

+VOWELS
-- sound development:
* articulation
* isolated pronunciation
* sound sequences about 30 items /phoneme
-- training in words: about 20 items /phoneme
* one-syllable
* polysyllabic
--fluency: * sentences 10/phoneme
* word pairs 10/phoneme

+SIBILANTS
-- sound development:
* articulation
* isolated pronunciation
* sound sequences about 30 items /phoneme
-- training in words about 30 items /phoneme
practised sound is at the beginning
practised sound is in the middle
practised sound is at the end of the word
-- fluency: * sentences 10/phoneme
* word pairs 10/phoneme

+ INTONATION

The developers were in direct communication with speech therapists throughout the development, and the advice of the therapists was taken into consideration in the construction of exercises and especially in the construction of the texts for the different languages.

7. Evaluation

The SPECO group has constructed an educational multimedia system for four languages, which is suitable for use with individuals, pairs and small groups. It follows most of the criteria of EvaluTech [8] (Software Evaluation Organisation of SREB in US.) During the development all partners kept in close contact with speech therapists. They have been provided with the Support material recently, and their opinion has been sought from time to time. Questioners were filled out by therapists from different educational fields to reveal any problems of the usage, and show the need for any corrections. This has improved the system.

When the system was being developed a detailed evaluation examination was organized in the Hard of Hearing School of Budapest. Groups of children with different problems were taught with the help of the new system and a control group trained by the traditional way was compared. The intelligibility of the speech of these children after a half-year of training was compared. In those groups which used the Support material the intelligibility was 3.2 times better on average after the training than before. In the control group this rate was only 1.9. A detailed presentation of this evaluation will be given during the conference.

8. Acknowledgement

This research has been supported by the European Union in the framework of the Inco-Copernicus Program (Contract no. 977126) and by the Hungarian Scientific Research Foundation.

9. References

[1] Vicsi, K., Roach, P., Öster, A., Kacic, Z., Barczikay, P. and Sinka, I.: A Multimedia Multilingual Teaching and Training System for Speech Handicapped Children (Proceedings of EUROSPEECH'99, pp. 859-862)
[2] Csatári, F., Bakcsi, Zs. and Vicsi, K.: A Hungarian Child Database for Speech Processing Applications (Proceedings of EUROSPEECH'99, pp. 2231-2234)
[3] Vicsi K., Csatári F., Bakcsi, Zs. and Tantos A.: Distance Score Evaluation of the Visualised Speech Spectra at Audio-visual Articulation Training (Proceedings of EUROSPEECH'99, pp. 1911-1914)
[4] Vicsi, K., Vig A.: LIAS: Language Independent automatic Segmentation Technique Using Sampa Labelling of Phonemes. First International Conference on Language Resources Education, Granada Spain, 1998. 1.317
[5] Kacic, Z., Vicsi, K., Roach, P., Sinka, I., Öster, A., Ogner, M. and Barczikay, P.: Development Perspectives of Multimedia Multilingual Teaching and Training System for Speech handicapped Children
[6] Öster, A., Vicsi, K., Roach, P., Kacic, Z. and Barczikay, P.: A Multimedia Multilingual Teaching and Training System for Speech and Hearing Impaired Children - SPECO (FONETIK'99, pp. 149-152)
[7] J.L. Wallace at al. 1998. Applications of Speech Recognition in the Primary School Classroom ESCA – Still 98 Workshop Proceedings, Marholmen, 21-24.
[8] EvaluTech (Software Evaluation Organisation of SREB in US). http://www.evalutech.sreb.org/