SPECO: Computer-Based Phonetic Training for Children

Anna Sfakianaki, Peter Roach, University of Reading
Klara Vicsi, Ferenc Csatari, Technical University of Budapest
Anne-Marie Öster, KTH, Stockholm
Zdravko Kacic, University of Maribor
Peter Barczikay, Robot Control Software, Budapest

1 Introduction

The idea of computer systems which exploit developments in speech technology for pedagogical or remedial purposes is not new, and a number of such systems have been developed. The SPECO Project is attempting to produce a new system which uses advanced speech technology and is adaptable to different languages. Although its primary purpose is for clinical remediation of children's speech problems, we believe that it has additional potential in pronunciation teaching.

2 Components of the system

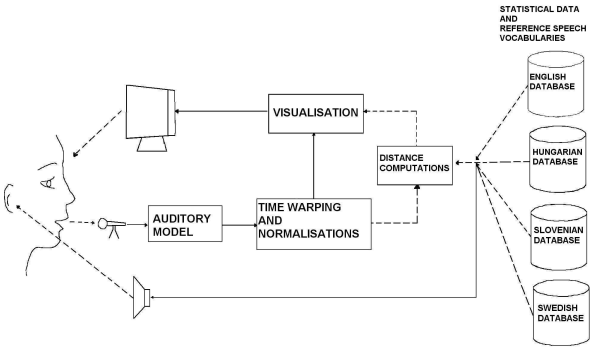

Figure 1 The components of the SPECO system

The system consists of two basic parts. The first part consists of a language-independent editor and measuring system which is used to construct the modules for all SPECO languages. This language-independent editor can be adapted to any European language. The second part consists of language-dependent speech databases. As can be seen in Figure 1, a microphone picks up the speech signal and the acoustic speech processing follows.

The acoustic speech processor is a simple auditory model, which imitates the low level processing of the human hearing system which we understand reasonably well (Zwicker, 1982; Zwicker & Terhardt, 1980). So the model analyses the speech signal with time, frequency and intensity resolution which is approximately similar to that available to the human peripheral auditory speech perception. The data is valid for average speaking rates and average speech intensity level (65 dB) (Vicsi, 1981; Vicsi *et al.*, 1990). The separation of the complex sounds into their component frequencies is done in critical filter bands, from 80 Hz to 8 kHz. In this range 20 critical band filters are used.

Different time-warping algorithms are used so that the corresponding sounds in the reference utterances and in the client's utterances are shown one immediately below the other, in spite of the fact that the client may use a different speaking rate. In this way the clients can easily compare their speech with the reference speech. Normalisation algorithms are also employed to reduce inter-speaker variability and make the system usable on a wide range of different speakers. This is still work in progress and must be done with great care, because the differences between normal and disordered speech must not be removed.

The participating languages are English, Hungarian, Slovenian and Swedish, and there are thus four reference speech databases, which the system uses in order to make a decision about the microphone input. Finally, the output of the filters and the computations is visualised. The visual presentation of the acoustic parameters is achieved through speech pictures, and clients are trained to recognise the significant differences between the reference speech picture and their own.

3 The Child Speech Database

SPECO is essentially a speech recognition system, and like other modern speech recognition systems is dependent on training with carefully prepared training data. Much of the work of the project has therefore consisted of building up a substantial database of children's speech in the four languages of the project. Each language has two databases: the reference-speaker database and the multi-speaker database. The four language versions are divided into two packages: the fricative and

affricate support and the vowel support. In the case of the English version, the fricative and affricate support includes the sibilants s, z, S, Z and the affricates tS, and dZ. The vowel support includes the five long vowels i:, Î:, A:, ł: and u:, and the six short vowels I, e, Q, Ã, Ł and U.

The fricative and affricate support material was recorded with our reference speaker, CM, when he was eight years old, and the vowel support material was recorded about a year later. All recordings were carried out in the sound-deadened recording room in the speech lab at the University of Reading, using the special editor incorporated in the SPECO system. Each utterance was recorded three times and the best one was saved and chosen to appear as the reference example in the exercises. The reference database was segmented using a special application within the SPECO editor. The reference examples were segmented so as to feed the system with information about the normal range of each phoneme and to demonstrate the arbitrary limits of each phoneme in the exercise window to assist the speech therapist and the client in training.

The English multi-speaker database contains a portion of the reference material. 36 children aged between 7 and 11 were recorded. Each recording session took approximately 8-12 minutes, depending mostly on how fast the child could read the utterances from the cards. The speakers were selected from three different schools in the Reading and London area. The recordings were made at the schools and in rooms isolated from the rest of the classrooms and as quiet as possible. It may be worth noting that some children had problems articulating certain fricative and affricate sounds, most commonly [Z] and [dZ], especially in isolation. There were also some articulation problems concerning the sounds [¨] and [T]. The multi-speaker database was also segmented but this time using software (WASP) not incorporated in the editor itself.

Both databases have been used to establish norms which guide the teaching or remediation process. The segmented material was used in the construction of fricative and vowel spectra, referred to as "spreadlines" in the project, and determined the allowed spectral deviation. The spreadlines are constructed for each language separately (Figure 2) and constitute the actual background of the exercise (see also Figure 6).

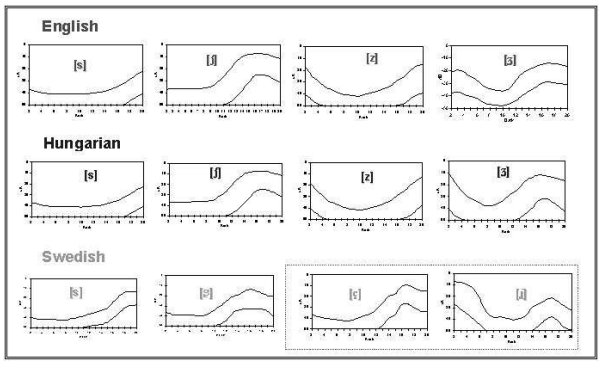

Figure 2 Typical fricative spectra of three SPECO languages produced in isolated pronunciation, with the allowed spectral deviation

4 Types of display

The concept of the SPECO system is to visualise speech at a low level of speech processing and to let clients use their high level information processing ability to work on this. Teaching children how to obtain information from speech pictures is preferable to giving articulation instructions. A detailed examination has been carried out to decide what scale of loudness, pitch contour, spectral distribution, etc., gives the most informative visual presentation (speech pictures) of these parameters. How can we draw children's attention to the areas of maximum energy in the spectrogram? How can we encourage them to use correct loudness and intonation levels? How can children recognise if their rhythm is appropriate? Generally we use different amusing background drawings to help children find the important parts of the speech pictures. First of all, each phoneme is assigned its own symbolic picture so that the child very quickly finds out which are the significant parts of the screen (Figure 3).

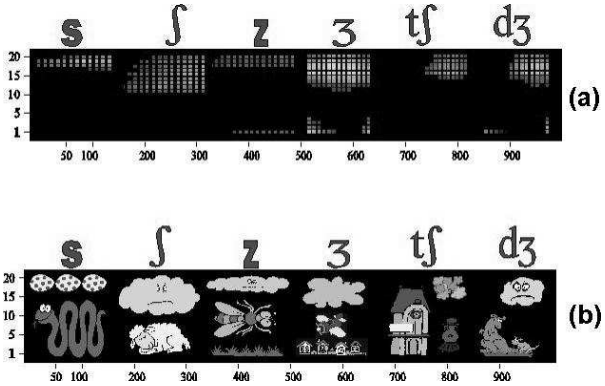

Figure 3 (a) Typical cochleograms and (b) pictures for sounds.

Figure 3 shows (a) typical cochleograms of the English fricatives and affricates trained by the system and (b) the picture corresponding to a particular sound so that the client can make the necessary association when looking at the speech picture. For example, the correct production of an [s] (which is symbolised with a snake) would cover most of the eggs with dots.

Some examples of types of speech pictures are shown below: energy changing with time (Figure 4); pitch; voiced - unvoiced detection; intonation; spectrum; spectrogram (cochleagram); spectrogram differences.

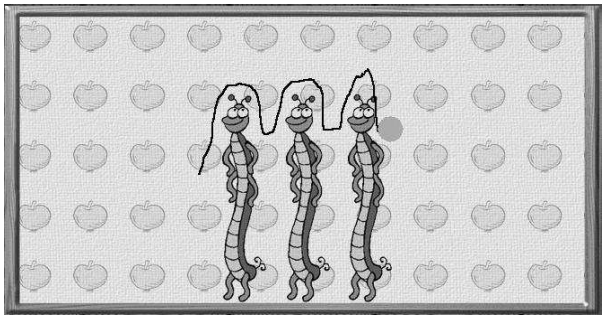

Figure 4 In saying pi pi pi, the child must make the yellow ball jump over the heads of the worms with the appropriate rhythm.

The system is based on up-to-date technology, but we follow the steps of traditional speech therapy in both modules. These are sound preparation, sound development, followed by training in words and automation (meaning the achievement of a reliable production not requiring further instruction).

At the stage of *sound preparation* children are trained to pay the necessary attention to the screen. They start to familiarise themselves with the way curves form on the screen according to sound energy and the position of the speech organs. There is the possibility to train the adjustment of different speech parameters: loudness, rhythm, spectrum, pitch, voicing, intonation.

In *sound development* we start with the forming of individual phonemes. This stage includes working with articulation pictures, isolated pronunciation practice and syllable training. The articulation pictures (Figure 5) show the child which is the right position of all the organs (mouth, tongue, teeth etc.) that play a role in sound production.

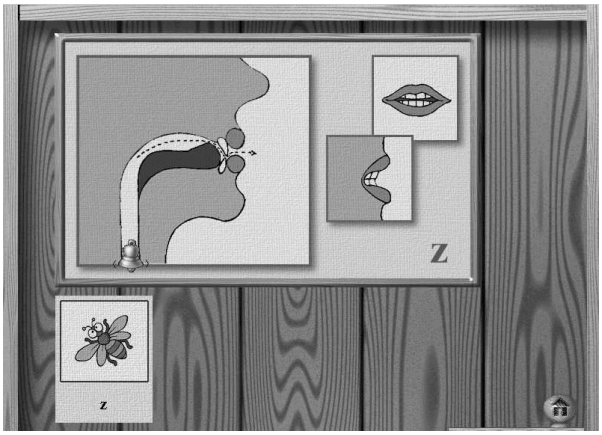

Figure 5 Articulation picture for fricative [z]; the little bell ringing indicates that there must be voicing when producing this sound.

The change of the energy measured in each frequency band is visible on the screen. The form of the distribution lines of the correctly pronounced phoneme is different in each case and characterises the phoneme itself (Figure 6). This exercise has three levels of difficulty: the limits of "spread" can be selected according to the level of the learner.

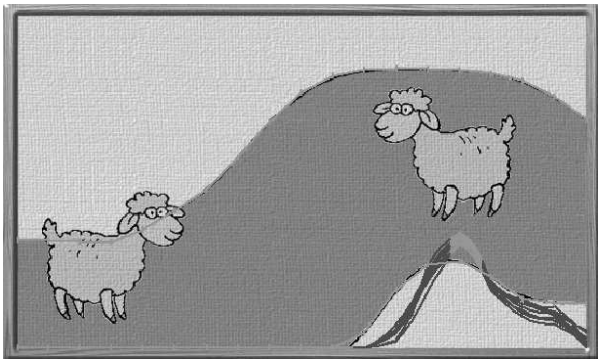

Figure 6 The spectrum of the English fricative [ʃ] presented as a speech picture by the program. The objective is to produce and sustain a line within the limits of the "green field".

For syllable training, the vocabulary contains sound sequences constructed so that the phonemes being practised occur in different positions and contexts. These syllables appear on the screen in the form of cochleagrams. For the English fricative and affricate support, fricatives and affricates are presented in CV, VCV, VC and VC-VC-VC position and connected with the five long vowels. Whereas the English vowel support contains all vowels in syllables along with front stops, like [p, t and b].

The order of presentation of sound sequences could be important, so we grade those from the easier pronunciations to the more difficult ones. In this exercise, the reference syllable is demonstrated on the upper half of the screen, while the syllable the client produces appears in the bottom half of the screen (Figure 7). The client attempts to match his picture with the reference one as closely as possible.

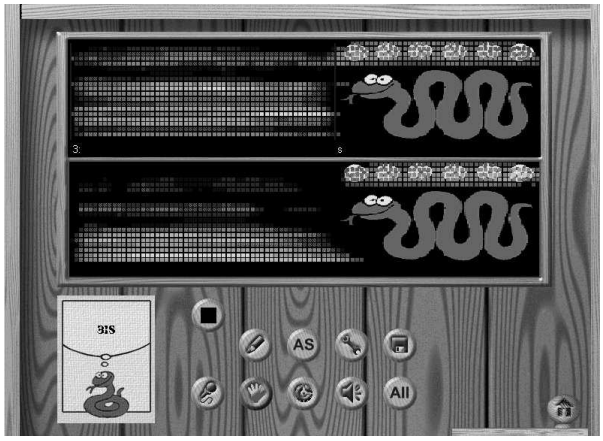

Figure 7 The reference syllable [ɜ:s] appears on the top half of the screen and the client's production below. The blue dots correspond to the vowel and the red dots to the fricative. The aim is to cover as much of the eggs as possible with red dots and leave the snake uncovered.

In the *training in words* the grouping of words is different in fricative support and vowel support. In fricative support all phonemes are presented in initial, medial and final position in words. In vowel support all phonemes occur in one-syllable words and in words of two or more syllables. Again the upper half of the screen shows the cochleogram of the reference word and the client has to produce the same pattern in the bottom half of the screen.

The *automation* (or "continuity" for the English version of the system) consists of two parts: contrast pairs and phrases. These exercises work on the basis of cochleograms as well. The contrast pairs are presented to the child to show the differences between the speech pictures of two phonemes in similar words. For example one of the word pairs chosen to train the phoneme /z/ word-initially is "zip-dip". The phrases contain the trained phoneme at least once and they are specially designed and graded from simple and short to complex and longer ones. Our aim in the therapy is to reach that speech level at which the client speaks correctly without having to concentrate on the articulation. Therefore, besides the practice with phrases, we have included a category called "free exercise". The therapist produces the reference example and the client attempts to produce the same example correctly. The example can be a syllable, a word, a pair or a sentence according to the client's level, wishes and needs.

5 Evaluation

In Budapest where the project is directed and there are extensive contacts with schools for the deaf and speech therapists, a detailed and large-scale evaluation system, both qualitative and quantitative has been designed and tested.
The other partners will also carry out evaluations in their own environment. The evaluation for the English version will be carried out in the Communication Disorders Centre (CoDisC) at the University of Reading.

6 References

Vicsi, K. (1981) The most relevant acoustical microsegment and its duration necessary fot the recognition of unvoiced stops. *Acoustica*, 48: 53-58.

Vicsi, K., Matilla, M. & Berényi, P. (1990) Continuous speech segementation using different methods. *Acoustica*. 71: 152-156.

Zwicker, E. (1982) *Psychoakustik*. Berlin: Springer Verlag.

Zwicker, E. & Terhardt, E. (1980) Analytical expressions for band rate and critical bandwidth as a function of frequency. *JASA*, 68:1523.