6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

S P E C O
A MULTIMEDIA MULTILINGUAL TEACHING AND TRAINING SYSTEM FOR SPEECH HANDICAPPED CHILDREN

K. Vicsi– P. Roach† – A. Öster‡ – Z. Kacic^ – P. Barczikay # – I. Sinka†*
*Technical University of Budapest, Hungary, †University of Reading, United Kingdom,
‡Kungl. Tekniska Högskolan, Sweden ^University of Maribor, Slovenia, #RCS – Robot Control Software, Hungary

ABSTRACT

In the frame of the INCO-Copernicus program of European Commission we have started to develop an audio-visual pronunciation teaching and training method and software system for hearing and speech-handicapped persons to help them to control their speech production. A teaching method is drawn up for progression from the individual sound preparation to practice of the sounds in sentences.

The main aim is to develop an audio-visual articulation training and teaching system for all participant languages, these being English, Swedish, Slovenian and Hungarian.

The basic part is a general language-independent measuring system and database editor. This database editor makes it possible to construct modules for all participant languages and for different speech disabilities. Two modules are under development for its construction in all languages, one of them being for teaching and training vowels for hearing-impaired children, while the other one is for correction of misarticulated fricative sounds.

10.21437/Eurospeech.1999-209

1. INTRODUCTION

During the process of learning speech, children with normal hearing follow a product-oriented approach. They discover how to control their speech organs through reference to acoustic speech signals. In this way they develop the ability to generate all the acoustic effects occurring in speech. Speech impaired persons have some problems in this process. In traditional speech therapy a process-oriented approach is generally used [1] the speech therapist gives instructions on how to use the speech organs while forming sounds. During normal speech development, children never receive instructions on how to move or where to place their speech organs.

Instead of the process-oriented approach, or to supplement it, our project would like to offer a product-oriented one. In speech communication it is not the process of the articulation that is important, but the quality of the produced sound by which the information is transmitted to the other person. In our project developed for hearing-impaired children the produced sound is measured and visualised. The user here discovers how to control his or her speech organs through reference to the visualised pictures of the acoustic speech signals. Thus the system helps the patients to discover how to move their speech organs by simultaneously comparing the visual patterns (speech pictures) of the normal acoustic speech signal with the disordered one. In this way they may develop the ability to generate most or all of the acoustic effects occurring in speech.

For the project, there is a need for the cooperation of scientists who represent different fields, such as digital speech processing, speech acoustics, expertise in linguistics on different levels, in speech therapy, and knowledge of the newest technological facilities. With the cooperation of different experts we are developing the system for all participant languages.

2. GENERAL METHOD OF SPEECH THERAPY

The aim of our project is the correction of the disordered aspects of speech by visual presentation of the speech signal, using the healthy visual channel of the patient for additional processing. However, during practice we also use the patients' limited auditory channel, by giving auditory information synchronised with the visual information. (Fig. 1.)

Our system differs from those articulation teaching systems that have been developed on the base of current speech recognition systems (ASR). Usually these can do no more than distinguish between good and poor pronunciations of a known word spoken by a child. The normal goal of ASR is to classify all utterances correctly, even if they are not pronounced accurately. But in pronunciation teaching, whether the speech of a child is good or poor determines whether most people in their educational, public and cultural life can understand them easily or with difficulty [2]. Thus the two systems, ASR and pronunciation teaching systems have different aims. Nevertheless, many results and algorithms of ASR are used in our speech processing system.

A microphone picks up the speech, then the acoustic speech processing follows. The main problem is to decide which parameters are important for the speech processing and how to present these to the patient; we need to know what sort of parameters change simultaneously during articulation, and what sort of parameters patients can use to make the decision themselves as to whether their pronunciation is correct or not, and how far it is from the correct one.

2.1 The acoustic speech processing

The sound pressure-time function is not very informative and changes easily as a result of many factors. However, the decomposition of the complex sounds of speech into their component frequencies is an important first step of the analysis. Separation into frequency channels is maintained throughout the auditory nervous system, so it would obviously be advantageous to do this separation in a way that is similar to the human auditory system.

But we still do not have a complete understanding of the way in which even the simplest of speech sounds is processed by the auditory system and we know virtually nothing of the neural activity evoked by running speech or even sentences or phrases However there is accumulated knowledge derived from the neuro-physiological examination of the low level peripheral hearing system, which is in harmony with psycho-acoustic experiments. Such processing is for example the critical-band spectrum analysis and real-time masking.

Our acoustic speech processor is a simple auditory model, which imitates only the low level processing of the human hearing system, which we understand reasonably well [3, 4].
The model analyses the speech signal by approximately similar time, frequency and intensity resolution available to the human peripheral auditory system during speech perception. The data are valid for average speaking rates and average speech intensity level (65 dB). [5, 6]
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 were used. Visualising the output of these filters as a function of time gives the so-called cochleogram.
For pitch measurement the simple AMDF method will be used instead of a technique which would be suited to the auditory model.

2.2 The visual presentation

The further processing of the output of the model is done for the purpose of better visualisation. The visual presentation of the acoustical parameters (are called speech pictures) gives the possibility to the user to process the speech further on the basis of this visual presentation. (Comparative physiologists and neuroethologists have presented data which support the interpretation that specialised sensing and motor functions have emerged from common neural and/or physical mechanisms [7])
A detailed examination has been 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. How it is possible to draw the children's attention to the areas of maximum energy in the spectrogram. How is it possible to encourage the child to use correct loudness and intonation levels. How the child can recognise if he/she pronouncing sounds with a inappropriate rhythm etc. Generally we use different amusing background drawings to help the child to find the important part of these speech pictures.

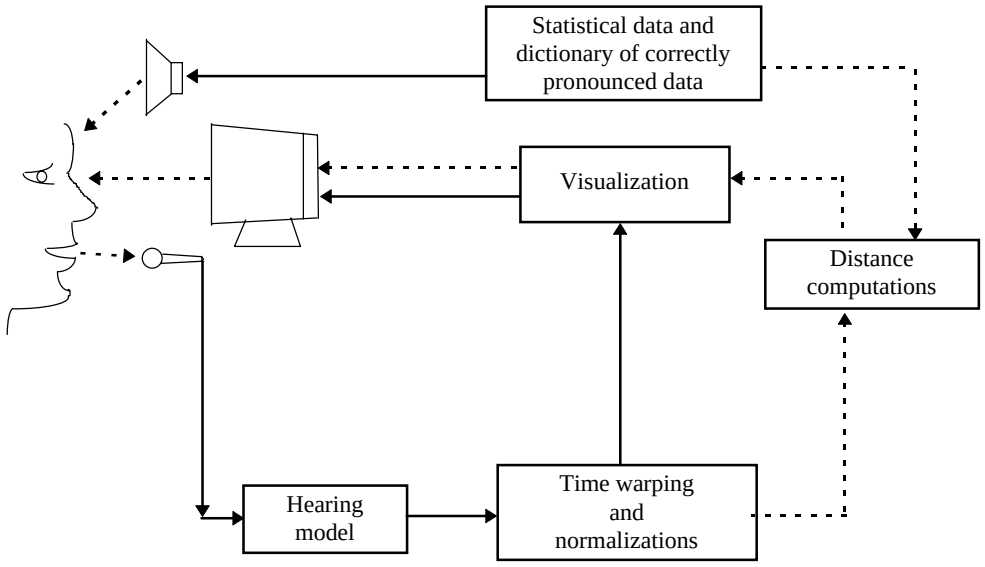

Figure 1. System structure of the teaching and training method

An automatic begin/end processing is involved to measure the beginning and the end of the phrases. With the help of different time-warping algorithms [8,9], the corresponding sounds in the reference speech and in the imitation are shown one immediately below the other, in spite of the fact that the duration of the reference speech and imitation may be different. In this way the sounds are easily comparable in the reference speech and in the imitation.

2.3 Vocabularies

While the acoustic speech processing and the visual presentation are language independent separate vocabularies must be constructed for all participant languages. In the construction of the vocabularies the language specific characteristics and specialised training methods of the traditional national therapy has been taken into consideration. In the vocabulary all trained phonemes must be presented 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 speech examples are produced by carefully selected persons who speak clearly. Their speech pictures must be clear and easy to read. The aim of the patients during therapy is the production of sound pictures similar to these references in the vocabulary.

3. DIFFERENT STEPS IN THE THERAPY

Two modules are under development in all languages, these being English, Swedish, Slovenian and Hungarian. One of them is for teaching and training vowels for hearing-impaired children, the so called VOWEL SUPPORT, and the other one is for correction of misarticulate fricative sounds and affricates, the FRICATIVE SUPPORT.
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, and training in words and automation. We have constructed specific tasks in a specific order involving the teaching experiences of the teachers of a given language. For example in case of the sound development of a linguapalatal fricative consonants /Σ/ we start to practice this sound together with a back vowel /u/, because the pronunciation of this sound connection is easier, than for example together with a front vowel like /i/. Thus one of the most important tasks for all partners has been to construct a well-defined text.
In the sound preparation we have the possibility to train the adjustment of different speech parameters. Four different types of practice are planned: loudness, spectrum, pitch and pitch-loudness.
In the sound development the child can choose a phoneme and start to develop it.

- The Fricative module involves all fricatives and affricates of the language.
- The Vowel module involves all vowels and diphthongs of the language.

The development starts with isolated pronunciation. On the screen the change of the energy measured in each frequency band is visible. The form of the distribution lines of the correctly pronounced phonemes is different in each case and characterises the phonemes themselves. Training in syllable sized units is under development. The vocabulary contains sound sequences constructed so that the phonemes being practised occur in different positions and contexts.
For FRICATIVE SUPPORT fricatives and affricates are presented in CV, VCV, VC and VC-VC-VC position and connected with all vowels and diphthongs. The order of presentation of sound sequences could be important, so we grade those from the easier pronunciations to the more difficult ones.
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.

The **contrast pairs** are presented to the child to show the differences between the visual pictures of two phonemes in similar words.
Our aim in the therapy is to reach that speech level at which the patient speaks correctly without having to concentrate on the articulation. In this system the children can practise the sounds **in sentences** too. Specially designed sentences, first simple, then complex ones have been collected in the vocabulary of all participant languages.

4. PRELIMINARY CLINICAL EXPERIENCES

We have made some preliminary experiments with the draft version of the Hungarian Fricative Support module in The School of Hearing - impaired Children, in Budapest.
Generally it is not a simple task to assess the effectiveness of a multimedia teaching system. What facts determine the goodness of a system? First we have tried to work out a concept, of the kind of examination which could be useful for the evaluation of such a new multimedia tool and for comparison with more traditional methods. In our view these examinations would be the followings:

a) How quickly can the correct pronunciation of each sound be achieved?
b) How much time does it take to reach fluency in the pronunciation of the correctly formed sounds?
c) How good is the quality of the formed sounds?
d) Does the system make it possible to maintain the good quality of the sounds produced?

Such a detailed examination is planed for the future. Now we have some preliminary results.

a) The time requirement for forming the fricatives and affricates has been examined with the Fricative Support. We put together 4 different groups, with different degree of hearing-impairment, 5 children being collected in each group. Summarised results can be seen in Table 1.

5-10 year-old children with small hearing impairment the high frequency regions	in each case at the first occurrence within 1/4 hour
severely hearing-impaired 5-10 year-old children	in each case at the first occurrence within 1/2 hour
deaf 5-10 year-old children	in each case at the first occurrence within 1/2 hour
speech handicapped 5-10 year-old children with normal hearing	in each case at the first occurrence but the results here are widely dispersed

Table 1. Average development-time of fricatives and affricates with the Fricative Support

b) We have investigated the time requirement for the automation in 1 totally deaf child, 4 children with small hearing impairment and 4 severely hearing-impaired, aged 5-10, as well as in 2 children aged 6 of normal hearing. By using the Fricative Support we found that a consistently shorter time was required for the fixing and automation of a speech sound, than was the case with corresponding children of similar mental ability and impairment level, who had been instructed by the traditional method. However it is difficult to give values in figures because the result depended on many other factors (for instance one highly important factor was how much additional help the child received at home).

In speech handicapped children with normal hearing, the result was not completely clear. They went through the training with pleasure and used the computer happily, but we did not experience the same remarkable acceleration as was evident with the hearing-impaired children.

c) We have not yet run a subjective auditory examination for the evaluation of the quality of the formed sounds.

This investigation will be continued with the new versions of the Fricative and Vowel Support.

This is the first report of the SPECO Group; the duration of the project is 3 years, from 31st September 1998 to 30th August 2001.

ACKNOWLEDGEMENT

The research has been supported by European Community in the frame of Copernicus Program and by the Hungarian Scientific Research Foundation.

REFERENCES

[1] Povel, D.J. 1991. The Visual Speech Apparatus: Theoretical and practical aspects. *Speech Communication*, Vol. 10, 59-80.

[2] J.L. Wallace at al. 1998. Applications of Speech Recognition in the Primary School Classroom ESCA – Still 98 Workshop Proceedings, Marholmen, 21-24.

[3] Zwicker, E. 1982, Psychoakustik (Springer Verlag, Berlin)

[4] Zwicker, E. and Terhardt, E. 1980. Analytical expressions for band rate and critical bandwidth as a function of frequency, J. Soc. Am. Vol. 68, 1523.

[5] Vicsi, K. 1981. The Most Relevant Acoustical Microsegment and Its Duration Necessary for the Recognition of Unvoiced Stops ACOUSTICA Vol. 48, 53-58.

[6] Vicsi, K. Matilla, M. and Berényi, P. 1990. Continuous Speech Segmentation Using Different Methods, Acustica, Vol. 71, 152-156. Video Voice, Micro Video, 210 Collingwood, Suite 100. PO Box 7357 Ann Arbor, MI 48107.

[7] R. Cambell at al. 1998. Hearing by Eye II. Psychology Press 1998.

[8] Sakoe, H. and Chiba, s. 1978. Dynamic programming algorithm optimatization for spoken word recognition, IEEE Trans. Acoust. Speech and Signal Process, Vol. ASSP-26, 43-49.

[9] Rabiner, L.R. and Levinson, S.E. 1981. Isolated and connected word recognition - theory and selected application. IEEE Trans on Comm. Vol. Con. 29., No. 5, 621-659.