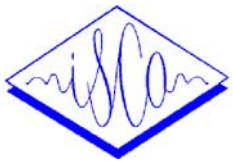


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

COMPUTER-AIDED SPOKEN-LANGUAGE TRAINING WITH ENHANCED VISUAL AND AUDITORY FEEDBACK

Jan Nouza

SpeechLab, Dept. of Electronics and Signal Processing
Technical University of Liberec, Halkova 5, 461 17 Liberec, Czech Republic
E-mail: jan.nouza@vslib.cz, WWW: http://itakura.kes.vslib.cz/kes/kes_lab.html

10.21437/Eurospeech.1999-49

ABSTRACT

A tool developed for computer-aided training of spoken language is presented in the paper. The tool's environment utilizes both visual and auditory feedback information to help a user in learning pronunciation and intonation in L1 or L2. The learning is supported by displaying the user's speech and its relevant parameters (volume, F0 and spectrum) in parallel with multiple reference templates. The templates may belong to the same utterance or make a minimum pair that can be used for contrastive training. The time plots are accompanied by textual labels (phonemes, syllables or words) that are automatically aligned to the user's utterance and by plots that identify the regions with major deviations with respect to the reference templates. The tool has been tested in two tasks: a) speech training of a deaf person and b) learning pronunciation and intonation in a foreign language.

1. INTRODUCTION

A modern multimedia PC has become a widely available platform for enhancing the quality of speech and language learning. It has been used with success in such tasks like training basic speech abilities, improving pronunciation, practicing intonation, gaining speaking fluency and evaluating the already acquired skills - as it was demonstrated at specialised workshops [1-2].

In most of these tasks the computer serves as a machine that records speech signals, plays them back and eventually visualises them in order to provide a user by feedback information. The visualised data - in form of plots, diagrams or spectrograms - are used in comparative or contrastive methods developed for training hearing-impaired persons [3] and for learning pronunciation in a foreign language [4]. Other systems may focus also on learning prosody and correct intonation, using either visualised F0 contours [5] or acoustic synthesised signals [6]. The most sophisticated systems can provide the user by animated artificial agents, like Baldi [7], whose articulators (the tongue, teeth, hard and soft palates) can be made visible and synchronised with natural or synthetic utterances. A lot of effort has been put also to automatic assessment of speech quality [8].

Practical experience shows that the best results in spoken language training are achieved when the trainee is guided by complex, both a priori and feedback, information. The a priori information must include explanation how to do; i.e. how to articulate, how to produce specific sounds, which prosody features should be improved and how, etc. The feedback information must clearly demonstrate how well the target was met. In practice it means that the trainee should have a chance to compare his/her utterance to a correct template (both visually and acoustically if is possible), to see and hear where the major mistakes in his/her speech occurred and to learn how good (at least relatively) was his/her current attempt.

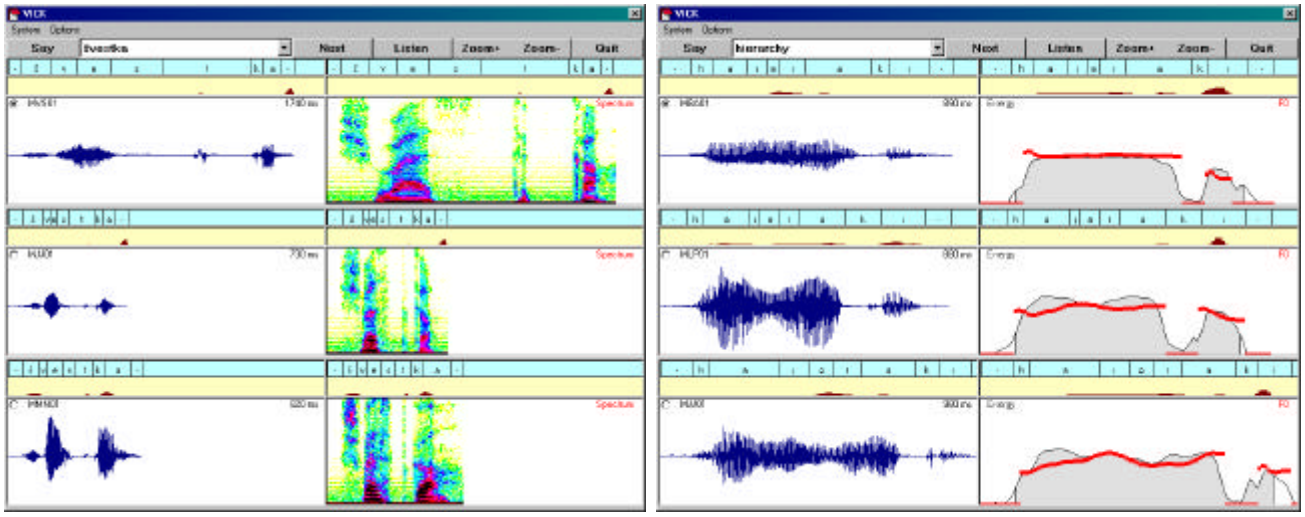
Our own experience comes from the several-year period of development [9] and evaluation [10] of speech training tools that where originally designed for the hearing-impaired users. The recent version of the training tool, named VICK, has been further enhanced to include support also for learning foreign languages. Now, the product provides both visual and auditory feedback and offers some kind of a priori information (e.g. phonetic transcriptions of utterances to be trained) as well as detailed analysis of the spoken attempt (plots of several parameters, phonetic markers and labels, identification of places with major deviations and an automatic evaluation module).

2. THE SPEECH TRAINING TOOL

The tool's previous versions were named VICK (as VIsual feedbaCK). The name has remained unchanged also in the latest version, even though the tool provides both visual and acoustic feedback, now.

2.1 The features

The VICK has been developed as a non-expensive computer-aided system that can help people in training and practicing pronunciation of single words as well as intonation of whole sentences.



a) A hearing-impaired person training a Czech word - using spectrographic representation of speech signals

b) A student training pronunciation of an English word - using energy and F0 representation of speech signals

Fig.1 - *Snapshots demonstrating the course of training speech through visual and acoustic feedback*

The main features of the VICK tool are summarized in the following paragraphs:

1. The tool supports both comparative and contrastive methods of speech training. This means that the trainee may compare his/her current utterance either to several reference recordings of the same utterance or to a pair made of a correct and an incorrect template.
2. The tool displays (and optionally replays) the currently spoken utterance together with the templates. Displayed and replayed can be also selected parts of the speech, e.g. those where major deviations occurred.
3. The displayed patterns have form of plots, diagrams and labels that help the user to identify individual elements (phonemes, syllables or words) in the currently spoken and prerecorded utterances.
4. The trainee is guided in his/her attempts by optional hints, like e.g. the phonetic transcription and stress marks. This is helpful namely in foreign language pronunciation training.
5. The tool also includes a simple tutoring scheme that provides the user by an assessment of his/her attempts. The assessment is based on measuring the similarity between the spoken utterance and the templates.

2.2 The design and layout

The tool has been designed for the most common computer platform - a PC (100 MHz CPU at minimum) and the MS Win95/98 operating system. The speech signal is acquired and reproduced (usually) through a head-mounted set of a microphone and headphones attached to a 16-bit sound card. The signal processing routines have been written in C to make them fast enough for real-time operation while the user interface has been designed in Visual Basic in order to allow for easy modifications.

The screen layout of the VICK is shown in Fig.1. Its main window is vertically split into three identical *panels*. Each panel is used for displaying a single speech signal. (As default, the displayed signal consists of the automatically detected utterance accompanied by 10 frames before and after the speech endpoints.)

When the comparative training method is applied all the three parallel signals belong to the same currently trained word or sentence. Usually, the trainee's speech is displayed in the topmost panel, while the lower two panels show the references. The system was designed to be able to operate with more than one reference per a trained item. Typically two or even more speakers' recordings are used as templates at the same time. However, displayed are only those two candidates that were classified as the closest to the trainee's speech. This is to make the assessment less dependent on individual characteristics of the reference speakers.

In case of the contrastive training method the two lower panels are used to display the correct and the incorrect pronunciation (e.g. minimal word pairs differing in the trained phoneme). The correct template is displayed in the middle panel. The classifier, however, tells the user which of the two templates was found closer to the spoken utterance.

Each of the three panels is further divided into two halves that are used to display different types of parameters of the same signal, e.g. the time waveform, the spectrogram, the energy and F0 contours. Above the main *signal subpanel* there is a pair of complementary subpanels. The

higher one, the *label subpanel*, serves for positioning (phonemic, syllabic or word) labels, the lower, the *difference subpanel*, indicates the parts of speech with major differences between the trainees' attempt and the references.

The VICK's design heavily relies on classification, similarity measuring and time alignment. These tasks are accomplished through the dynamic time warping (DTW) technique. So, for example, the label positioning is done automatically using a DTW-driven alignment and previously stored reference label information. Similarly, the difference panels display the distance between DTW-aligned pairs of the utterance and reference frames. The distance is evaluated either for the whole set of features describing the speech signal or for a specific feature subset such as log energy and/or F0, depending on the type of the plot displayed in the adjacent main subpanel. Usually, the difference panels are set up to indicate only the major deviations that exceed prescribed thresholds.

In general, the VICK's design is language independent. It is because the alignment and assessment procedures are based on reference signals. The only language dependent part is the phonetic labeling module that must be set up for the given phonetic alphabet. For Czech we use the single-letter phonetic transcription alphabet introduced in [11]. For English the phonetic symbols used in the Longman series of dictionaries [12] have been adopted.

Most of the feedback information is provided in the visual form. Though, the user can make an additional benefit from replaying each of the utterances displayed on the screen. Moreover, he or she can listen to and compare speech segments belonging to individual phonemes, simply by clicking on the phoneme label.

Optionally, the VICK may also provide the user by an automatic tutor. It has a form of a small window with comments on the duration, volume, intonation and classification score of the currently spoken utterance.

2.3 Technical implementation

The tool operates with a front-end partly borrowed from our own speech recognition system. It automatically detects a speech signal and converts it into a sequence of 20-feature vector. (The features include the first 8 cepstral coefficients, their deltas, log energy together with its first and second derivatives and F0 value computed by the AMDF algorithm.) Before the data enters the DTW classification procedure and similarity measurement, it is modified by removing the DC offset, normalizing the log energy and passing through the CMS (cepstral mean subtraction) procedure. This is to eliminate variations caused by the recording facilities.

Positioning the labels is done automatically by backtracking the DTW path and aligning the known reference label markers to the frames of the trainee's speech. The label alignment is performed with the reference that has the minimum distance to the utterance. The difference subplots are evaluated in the same manner; using the (smoothed) values of local distances computed between the aligned pairs of frames. (For more details, see [9]).

3. WORKING WITH THE TOOL

The VICK design includes all the facilities that are necessary for the preparation of the training, for the training itself and for its evaluation.

3.1. Preparing a training session

The preparation of the VICK environment for a new training session consists either in selecting an already existing list of utterances or in creating a new one.

In the latter case the teacher types in the list of training items together with their phonetic transcription. Then the VICK asks the teacher (or another person) to record the reference templates. (For the contrastive training method a pair of utterances per each item must be recorded.) After an utterance is said it is displayed together with estimated positions of the phonetic labels. Any correction in their placement can be done by moving the markers by mouse. For each reference speaker, the template files store the raw signal, its feature vectors, the spectrogram and information about the labels and their positions.

3.2. Training and practicing

The tool allows the user or the supervisor to set up various options and parameters, such as the choice of the actual training list, the selection of the reference speakers, parameters for the endpoint detector (like, for example, the maximum length of intra-utterance pauses), options and thresholds for the plots and diagrams, etc.

Training list items can be practiced in an arbitrary order. In the standard setting, the speech from microphone is displayed in the topmost panel. However, it is possible to direct the input to any of the other two panels. This may be useful, for example, if the user wants to compare recent attempts with a previous one, or if the supervisor wants to add his or her utterance, e.g. to provide a contrastive sample.

3.3. Reporting and evaluation

For each training session, the VICK makes a report that can be later used for evaluation. The report contains all the system settings, scores and tutor comments for each utterance.

4. EVALUATION EXPERIMENTS

Recently we have conducted two different types of experiments: a) with a hearing-impaired woman trying to improve her pronunciation of some Czech words, b) with a group of students practicing pronunciation and intonation of English words and sentences.

4.1. A speech training experiment

The subject in this experiment was a post-lingually deaf person (a frequent collaborator of our team) whose speech is almost intelligible though it suffers from missing acoustic feedback. We have made a list of 40 words, in which her pronunciation differed significantly from the correct one. The list was recorded by four reference speakers with similar voice characteristics.

The subject passed through several training sessions with a supervisor who explained the meaning of the VICK's visual patterns and provided her by professional assistance. After these introductory sessions the subject was able to read the plots, to identify the parts of her speech (phonemes and phoneme groups) with wrong pronunciation and to make attempts approaching the correct one.

4.2 A L2 training experiment

This experiment was focused on testing the possibility of training the correct pronunciation and intonation in a foreign language. We chose English and created a list 20 words that are often pronounced wrongly by Czech students (words like „aerodynamics", „hierarchy", „maths", etc) and another list of 15 prosodicly rich sentences. Both the lists were recorded by two native speakers in several repetitions.

The subjects in this experiment were four Czech students. Their goal was to learn the pronunciation of the words and the correct intonation of the sentences with the use of the VICK. They were explained how use the tool and than they had a week for self-learning. At the end of the experiment we could compare the results of their learning with their initial recordings that had been made before they started using the VICK. In all cases the subjects significantly improved their pronunciation of the given words and sentences. In the word training part they appreciated mainly the labels and difference plots that helped them in identifying and comparing (both visually and acoustically) the wrong segments in their own pronunciation. Observing the F0 contour in the sentence training part they could see (not only hear) the difference in the intonation. Moreover, the VICK gave them a chance to see the intonation curves aligned with individual words in the sentence. Eventually, due to this alignment they were able to mimic the prosody features of the native speakers.

5. CONCLUSIONS

The major disadvantage of similar products is that the users may have problems with reading and understanding the meaning of the visualized data. Our tool tries to overcome this problem by adding labels to the plots and by pointing to those parts of speech where major mistakes in pronunciation or intonation seemed to occur.

The preliminary experiments mentioned in the previous section show that the tool may be helpful both for supervised training of hearing-impaired people as well as in self-learning foreign languages.

ACKNOWLEDGMENTS

The work described in the paper have been supported by the Czech Grant Agency (GACR) under the project No.102/96//K087.

REFERENCES

[1] Proceedings of the ESCA Workshop on Speech Technology in Language Learning (STiLL), Marholmen, May 1998.

[2] Proceedings of the ESCA Workshop on Method and Tool Innovations for Speech Science Education (MATISSE), London, April 1999.

[3] Öster A.-M.: Clinical Applications of Computer-Based Speech Training for Children with Hearing Impairment. Proc. of ICSLP'96, Philadelphia, October 1996, pp.157-160.

[4] Öster A.-M.: Spoken L2 Teaching with Contrastive Visual and Auditory Feedback. Proc. of ICSLP'98, Sydney, Dec.1998, pp. 2663-2666

[5] Auberg S., Correa N., Rothenberg M., Shanahan M.: Vowel and Intonation Training in an English Pronunciation Tutor. Proc. of STiLL, Marholmen, May 1998, pp.69-72.

[6] Sundström A.: Automatic Prosody Modification as a Means for Foreign Language Pronunciation Training. Proc. of STiLL, Marholmen, May 1998, pp.49-52.

[7] Cole R. et. al.: New Tools for Interactive Speech and Language Training: Using animated Conversational Agents in the Classrooms of Profoundly Deaf Children. Proc. of MATISSE Workshop, London, April 1999, pp.45-52.

[8] Sevenster B., de Krom G., Bloothooft G.: Evaluation and Training of Second-Language Learners' Pronunciation Using Phoneme-Based HMMs. Proc. of STiLL, Marholmen, May 1998, pp.91-94.

[9] Nouza J., Madlikova J.: Evaluation Tests on Visual Feedback in Speech and Language Learning. Proc. of STiLL, Marholmen, May 1998, pp.151-154

[10] Nouza J.: Training Speech through Visual Feedback Patterns. Proc. of ICSLP'98, Sydney, Dec.1998, pp.3293-3296.

[11] Nouza J., Psutka J., Uhlir J.: Phonetic Alphabet for Speech Recognition of Czech. Radioengineering, vol.6, no.4, Dec.1997, pp.16-20.

[12] Longman Active Study Dictionary of English. London 1995.