TRAINING SPEECH THROUGH VISUAL FEEDBACK PATTERNS

Jan Nouza

SpeechLab, Department of Electronics and Signal Processing,
Technical University of Liberec,
Halkova 6, 461 17 Liberec, Czech Republic
jan.nouza@vslib.cz, http://www.fm.vslib.cz/~kes/kes_lab.html

ABSTRACT

The paper describes a new version of a visual feedback aid for speech training. The aid is a PC based speech processing system that visualizes incoming signal and its most relevant parameters (such as volume, pitch, timing, spectrum) and compares them to utterances recorded by reference speakers. The goal is to help a trained person in identifying the most severe deviations in his or her pronunciation. The learning through visual comparison is supported by displaying multiple reference utterances, including phonetic labels both to the reference speakers' and trainee's speech, indicating the areas with larger deviations in any of the displayed features and offering a simple tutoring assessment of the trainee's attempts. Primarily, the system was aimed at hearing-impaired users, but its features make it well applicable also for foreign language pronunciation learning and practicing. The latter possibility was verified in an experiment in which a group of subjects tried to learn pronunciation of a couple of words in an exotic for them foreign language.

1. INTRODUCTION

Recent developments in speech technology offer new challenges and opportunities on a fast growing field of computer aided language learning (CALL) and computer aided speech training (CAST) tools. While the former are aimed at mastering the spoken form of foreign languages, the latter are to help people with speaking and, in particular, hearing, disabilities. The most recent overview of the current research in the CALL and CAST domains as well as a description of available products and diverse applications can be found in [1].

Our interest in the CAST field dates from 1996 when we were asked to develop aids that could help deaf and hard-of-hearing people in training and improving their speech abilities. Since no such products were available for Czech we tried to apply our previous experience from the speech recognition research in the design of simple aids and games for the hearing-handicapped [2]. The main training aid was based on speech signal visualization and visual comparison [3]. In 1997-98 the aid was tested in a school for deaf children, which provided us with a lot of practical experience (both positive and negative, [4]) as well as by new ideas for further work. At the same time, a demand to extend our research also on CALL systems came from industry and that is why the new version of the speech training system has been designed for a more general usage.

2. THE VICK SYSTEM

The product of our recent research has been named VICK (as VIsual feedbaCK) system. Its latest version has number 2 to distinguish it from the previous system described in [4].

2.1. Concept

The VICK's concept has many features common to recent CALL and CAST systems [5-6]. Its design builds on two different areas: signal processing and visualization. The former accomplishes such tasks like speech data acquisition, computation of relevant signal parameters, measuring and grading the speech quality. The latter makes the results of the data processing visible and understandable for the user.

The main features of the VICK design can be summarized in the following paragraphs:

1. The system serves for training pronunciation of single words and short phrases. It provides support both for hearing and hearing-impaired users.
2. The learning and training strategy is based on the comparison between visual patterns formed from the user's speech and reference utterances.
3. The visual patterns have a form of plots, diagrams and labels that allow for both complex and simplified look at the captured signal.
4. The patterns can be extracted and displayed both for pre-recorded and on-line signals. Their processing is performed with minimum time delay (usually without waiting more than 1 s).
5. The graphic design allows even a non-expert user to orient him/herself in the plots. This is accomplished mainly by placing phonetic labels along the time plots.
6. The speech parts exhibiting larger deviations from the references are marked and displayed in special signal difference subplots.
7. The system includes a simple tutoring scheme that provides the user by an automatic assessment of his/her speech attempts.

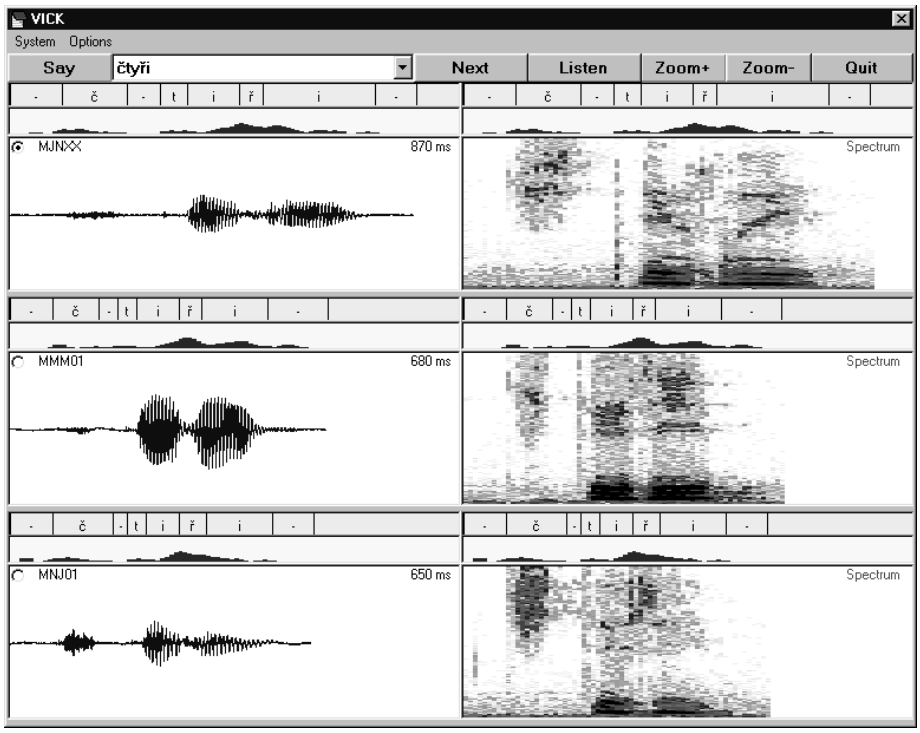

Figure 1. *A snapshot showing the VICK's screen during a training session. (Czech word 'four' is being practiced.)*

A trainee's utterance (in the topmost panel) is displayed together with two selected reference utterances (the lower panels). Signal waveforms are shown on the left side, color spectrograms on the right one. Each of the three panels has subpanels with phonemic label information and subpanels indicating distances between DTW aligned speech frames. This is to help the user in identifying the speech parts with the more severe deviations in pronunciation.

2.2. Design

To make the system available for wider use, for example, at schools, the VICK's design builds on the common personal computer platform, i.e. a PC with a 100 MHz processor at minimum, a monitor with at least 800x600 resolution, a 16-bit sound card, a microphone and loudspeakers or headphones. The computation-intensive routines have been compiled into fast 32-bit DLLs, hence the Windows95/NT operating system is a necessity. On the other side, the user interface has been written in Visual Basic, which allows for its easy modifications.

A typical VICK's screen is shown in Fig.1. We can see that its form 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.) In a standard situation, all the three parallel signals belong to the same currently trained word or phrase. Usually, the trainee's speech is hosted by 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 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.

Each panel 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 or F0 contours (see Fig.1 and 2). Above the main *signal subpanel* there is a pair of complementary subpanels. The higher one, the *label subpanel*, serves for positioning phonetic labels, the lower, the *difference subpanel*, indicates the parts of speech with major differences between the trainees' attempt and the references.

In the current VICK's version, the majority of classification and time alignment 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.

The user (if not hearing-impaired) 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. Similarly, a detailed visual analysis is available through zoom-in and zoom-out options.

2.3. System Features

When developing the software we have borrowed standard modules and parameters from our own speech processing and recognition system. The VICK runs with 16 bit/8 kHz sampling, 20 ms long frames (10 ms overlapped), a set of 20 features including 8 cepstral and 8 delta-cepstral coefficents, together with log energy and its first two derivatives and F0. No optimization of the parameters and the feature set with respect to the given task has been made yet. However, we are well

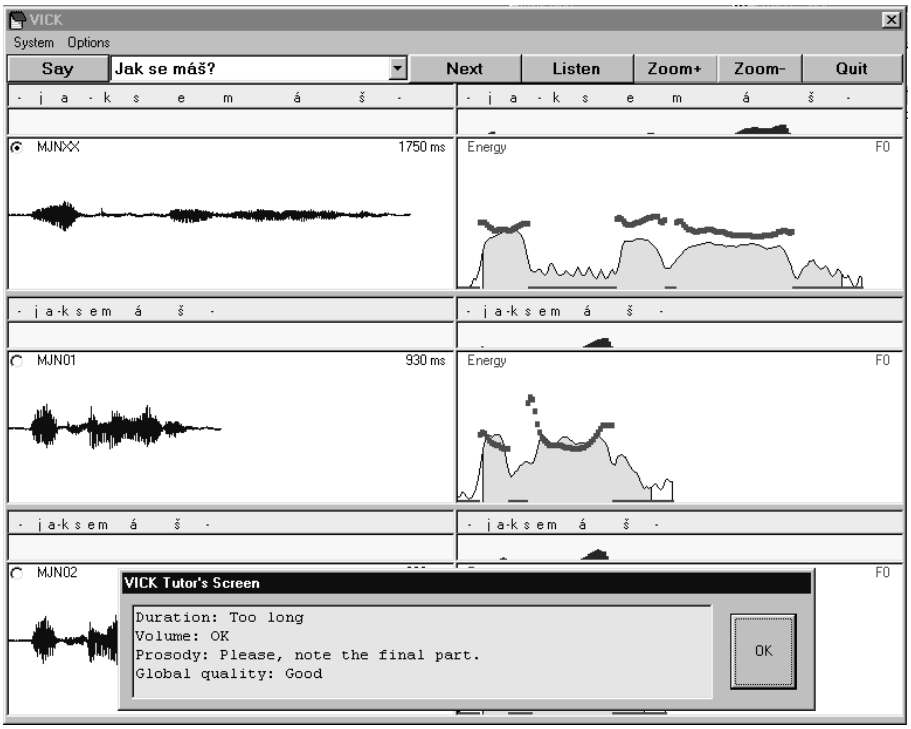

Figure 2. *Another example of the VICK's screen.*

This time, short Czech phrases are being trained. Instead of spectrograms, speech intensity (signal energy) and prosody (F0) features are displayed. Again the subpanels with distance diagrams can be utilized for identifying the utterance parts that differ most from the reference ones - either with respect to the complete signal description or just as to the F0 contour. Also a simple tutoring screen is available that tries to comment the trainee's attempt from several viewpoints such as duration, volume, prosody and global quality. The latter is based on the recogniser score.

aware of the fact that speech recognition and speech quality assessment may be rather different tasks.

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 [7].

2.4. Methods and Algorithms

Besides the standard speech processing modules we had to develop and implement several additional routines, e.g. for F0 estimation, label alignment, signal difference computation.

The F0 estimation is based on the classic AMDF method applied to 500Hz-lowpassed speech signal. For each frame the method provides an ordered list of F0 candidates. The resulting F0 sequence is optimized by selecting those candidates that make the smoothest contour. Voiced/unvoiced decision is based on energy, zero crossing rate and analysis of the AMDF shape.

Label positioning is done automatically by backtracking the path of a DTW utterance-to-reference match and aligning the reference label markers to the frames of the trainee's speech. The label alignment procedure is performed with the reference that has the minimum distance to the utterance.

Difference subplots indicate local deviations between the utterance and the references. They display a distance between frames aligned during the DTW match. If an utterance frame i represented by vector $u(i)$ has been aligned to frames j_1, .. j_N of reference k represented by feature vector $r_k(i)$, we define the frame difference function $D(i)$ as:

$$D(i) = \frac{1}{N} \sum_{j=j_1}^{j_N} d(u(i), r_k(j)) \qquad (1)$$

where $d(u(i), r(j))$ is a local frame distance computed for the actual feature set. When K multiple references are available, we can modify eq. (1) in the following way:

$$D(i) = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{N_k} \sum_{j=j_{1k}}^{j_{Nk}} d(u(i), r_k(j)) \right) \qquad (2)$$

The utterance subpanel displays the function $D(i)$ computed according to either eq. (1) or (2), while reference subpanels plots the function $D(i)$ computed by eq. (1) with reversed alignment. Peaks in the plots indicate the parts where the trainee's speech exhibits largest deviations.

Automatic evaluation of the global speech quality is not a simple problem. Anyway, the VICK tries to provide at least an estimate that is based on the classification results. The system has two in-built classifiers. Primarily, the DTW classifier is used because it is already employed in all the alignment procedures and does not require any models to be trained. The second classifier is a whole-word CDHMM recogniser. It may be alternatively utilized, provided there is some minimum number of reference recordings that can be used for HMM training. An assessment is based on means and variances computed from the scores achieved with the reference speakers.

2.5. Tutoring

As it is shown in Fig.2, the VICK has a small tutoring screen that may optionally tell the user some comments about the quality of the utterance, namely on the duration, volume, prosody and the global quality. The comments contain either a simple assessment or try to point to the detected deviations.

3. PRACTICAL USAGE

Practical work with the VICK system typically consists of three phases.

3.1. Preparation

The VICK's environment must be prepared by a supervisor. The supervisor decides about the list of the words or phrases to be trained. Then, in the VICK's preparatory mode, he or she types in the list items together with their phonetic transcription.

After that the system is ready for recording the reference utterances. When an utterance is said it is displayed together with estimated positions of the phonetic labels. It is a good practice to check acoustically the placement of the labels and the markers. Any correction in the placement can be easily done by moving the markers by mouse. After recording the complete list, the recordings and the label files are automatically stored on disk. If the HMM recogniser is to be used, an external model training software is invoked.

3.2. Training and Practicing

Training and practicing is possible either with or without the supervisor. The supervisor is important, particularly, in case of hearing-impaired users where his or her role is primary in the teaching. A non-handicapped user, however, may suffice with the acoustic and visual information provided by the system.

The VICK 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, options and thresholds for the plots and diagrams, etc.

Words and phrases can be practiced in an arbitrary order. Only if a reference recording (e.g. for later evaluation) is to be made, its items must follow the listed order. Standardly, the on-line speech is displayed in the topmost panel. However, it is possible to direct the input to any of the other 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, distances, scores and tutor comments for each utterance.

4. EXPERIMENTS

The experiments conducted with the VICK's previous version (reported in [4]) demonstrated the prospects of the speech training supported by visual patterns. Two types of tests were carried out: a) with a group of hard-of-hearing children in a specialized school, b) with a group of students in an environment simulating a missing hearing channel. In both cases, the subjects were able to benefit from the provided visual feedback information. Measured through HMM recognition scores, the acceptability of their pronunciation has improved in average by some 20%. For details, see [4].

Introducing the new features into the VICK design, mainly the phonetic labels and difference plots, further helps in improving the learning process. Recently we have conducted a preliminary test with a group of Czech subjects trying to learn pronunciation of 20 words in an exotic for them language (Vietnamese). The test conditions were similar to those reported in [4]. Results showed that several practical sessions helped the subjects to achieve the pronunciation level represented by 87% correctly accepted words. (In the previous tests the level was about 70%, for native, Vietnamese, speakers it was above 98%.) However, larger field test are yet to come.

5. CONCLUSIONS

The new version of our speech and language training visual feedback system has several novel features that may help in practicing and improving pronunciation of single words or short utterances. It is namely the introduction of multiple reference environment, phonetic labeling, difference measurement, scoring and tutoring support.

REFERENCES

1. Proceedings of the ESCA Workshop on Speech Technology in Language Learning (STiLL), Marholmen, May 1998.
2. Nouza J., Hajek D.: Speech Training and Motivating Tools for Hearing-Impaired People. In Studientexte zur Sprachkommunikation, Heft 13, Berlin, Nov. 1996, pp.154-159.
3. Nouza J.: Visual Processing of Speech: Tools for Education, Aids for Handicaped. Proc. of Int. Conference on Speech Processing (ICSP'97), Seoul, Korea, August 1997,pp.667-682.
4. Nouza J., Madlikova J.: Evaluation Tests on Visual Feedback in Speech and Language Learning. Proc. of STiLL, Marholmen, May 1998, pp.151-154
5. Hiller S., Rooney E., Laver J., Jack M.: SPELL: An Automated System for Computer-Aided Pronunciation Teaching. Speech Communication, no. 13, 1993, 463-473.
6. Larsson H.: Lingus - A General Purpose Computer Aided Language Learning System which Could Serve as a Platform for the Implementation of Speech Analysis Tools. Proc. of STiLL, Marholmen, 1998, pp.131-134
7. Nouza J., Psutka J., Uhlir J.: Phonetic Alphabet for Speech Recognition of Czech. Radioengineering, vol.6, no.4, Dec.1997, pp.16-20.

ACKNOWLEDGMENTS: The work referred in this paper has been supported by grant of the Czech Grant Agency (GACR) no.102/96/0351 and by grant of the Czech Ministry of Education in program PREZENTACE.