METHODS AND APPLICATION OF PHONETIC LABEL ALIGNMENT IN SPEECH PROCESSING TASKS

Jan NOUZA, Martin MYSLIVEC
SpeechLab at Dept. of Electronics and Signal Processing
Technical University of Liberec
Hálkova 6, 461 17 Liberec
Czech Republic

Abstract

The paper deals with the problem of automatic phonetic segmentation of speech signals, namely for speech analysis and recognition purposes. Several methods and approaches are described and evaluated from the point of view of their accuracy. A complete instruction for creating an annotated database for training a Czech speech recognition system is provided together with the authors' own experience. The results of the work have found practical applications, for example, in developing a tool for semi-automatic speech segmentation, building a large-vocabulary phoneme-based speech recognition system and designing an aid for learning and practicing pronunciation of words or phrases in the native or a foreign language.

Keywords

speech processing, phonetic segmentation, dynamic time warping, hidden Markov models, recognition

1. Introduction

Modern speech analysis and synthesis techniques require large databases of recorded speech signals. The recordings must be annotated, i.e. accompanied by precise phonetic transcription of their content. Often we also need to know exact positions of individual words and phonemes in the signal. This leads to a task that is referred to *as phonetic segmentation and label alignment*. It is an important task whose results find application in speech synthesis (to make an inventory of basic speech sounds), in recognition (for training models of speech units, usually phonemes), in speech learning and rehabilitation (for identifying wrongly spoken parts of an utterance).

Though the segmentation problem has been investigated for many years and several comprehensive studies have been published (e.g. [1]), for many researchers it still remains a challenging task. Recently, we could meet new methods based on artificial neural networks [2], analysis through synthesis [3], edge detectors [4]) or multi-level segmentation [5]. The problem of speech detection and segmentation in extremely noisy environment is dealt, for example, in [6].

In this paper we describe our approach and the methods we applied for creating a large annotated database designed for the purpose of the recognition of spoken Czech. Because no such database did not exist before, we had to start from scratch. To minimise the tedious work associated with recording and annotating speech data we have developed methods and tools that allowed us to perform signal segmentation and phonetic label alignment in a semi-automatic and even automatic way. We believe that learning our approach and experience may be useful for those who must accomplish a similar task, either in Czech or other language.

The paper is structured as follows: In the next section we state briefly the technical parameters of the database that was the main subject of the segmentation work. Section 3 gives a detailed overview of the methods we used. Section 4 deals practical implementation issues and section 5 is focused on comparing the accuracy of the methods. Practical applications based on the research are mentioned in section 6.

2. A training database for speech recognition of Czech

In 1998 we started to collect a large corpus of read Czech sentences. The structure of the corpus was designed so that the data could be used for training a phoneme based continuous speech recognition system for Czech language. We adopted a phoneme inventory defined in [7] (with a minor modification that consisted in including also the rarely occurring phoneme „schwa") and constructed a set of 82 phonetically rich sentences. The set, which is an extension of that proposed in [8], covers all 42 phonemes in various context. The less frequent phonemes have at least 16 occurrences while the vowels, for example, appear in several hundred instances. Up to now the sentences have been recorded by 60 people, who were selected with respect to the gender and age. Several speakers passed two recording sessions, so the complete corpus contains almost 14 hours of speech, now.

The signal was recorded by a common head-mounted microphone set directly to a computer using 8 kHz sampling rate and 16 bit resolution and stored in the WAV format. Each sound file was further processed to get parametric representation that is standardly used in our speech recognition systems: i.e. a 20-dimensional feature vector composed of 8 LP cepstral and 8 delta cepstral coefficients, log energy together with its 1st and 2nd derivatives and a value of the spectral variation function (SVF - see section 4). The feature vectors were computed every 10 ms for 20 ms long frames. The energy and SVF values are used only for phoneme boundary identification, not for distance or probability computation.

3. Methods of speech segmentation

The phonetic segmentation problem is defined as follows: Having a speech signal and its phonetic transcription (containing also symbols for pauses and noise events), the goal is to align the transcription with the signal, i.e. to find the starting and ending times (frames) for all the symbols in the transcription. There are several possible ways to reach the goal and here are the most commonly used ones:

1. *Manual segmentation.* A human expert listens to the signal while simultaneously watching its waveform and places manually the phoneme labels. In general, this is the most accurate approach, although it is known that even skilled experts (phoneticians) may differ slightly when positioning phonemes in the same signal [1]. Obviously, this approach is very tedious and expensive.
2. *Segmentation based on DTW.* This method can be employed if we have a reference template for the utterance to be processed and this template is already segmented and labelled. Using the well-known Dynamic Time Warping (DTW) algorithm (see, for example [9] or [10]) we can try to match and align both the signals and their transcription symbols. The method assumes that we have somehow produced the templates (usually using the manual way).

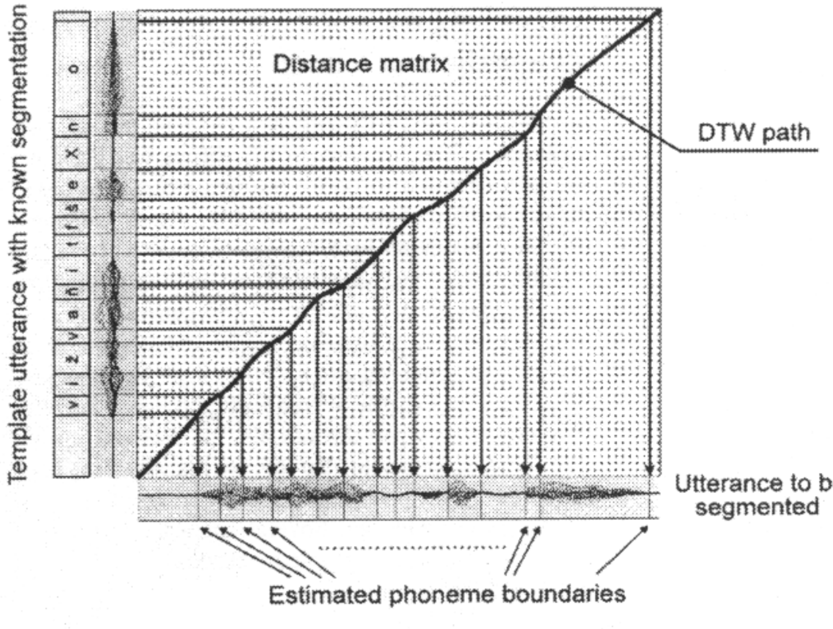

Template utterance with known segmentation
v
i
ž
v
a
ň
i
t
ř
š
e
X
n
o
Distance matrix
DTW path
Utterance to be segmented
Estimated phoneme boundaries

Fig.1. Illustration of the DTW based phonetic segmentation method

3. *Segmentation based on HMM.* Provided we already have statistic models (hidden Markov models - HMM [9,10]) of all acoustic (phonetic and noise) units we can use them to create a composite model based on the transcription and search for the best alignment between this model and the signal. However, to get good models of the phonemes, we need to train them on large amount of previously segmented and extracted samples of each phoneme.
4. *Embedded segmentation with HMM.* This most complex approach is based on an iterative procedure in which HMM training and signal segmentation is performed simultaneously. It assumes a large set of phonetically transcribed signals, a rather powerful machine and a lot of patience because many iteration steps are needed to achieve good models and an acceptable segmentation performance. This approach is the only one that does not require any manual work because it may run in fully automated, unsupervised manner. However, several published works show that the models and the quality of the segmentation are more or less inaccurate compared to the results achieved by a proper combination of the previously mentioned methods.

In the following subsections we shall take a closer look at the most convenient methods that are based on the DTW and HMM techniques.

3.1 DTW based alignment method

Let us have utterance template **Y**. Its signal is split regularly into frames that are represented by feature vectors $y = \{y_1, y_2, \dots y_N\}$. The template is phonetically transcribed as a sequence of K labels $a = \{a_1, a_2, \dots a_K\}$ that have been aligned with the frame vectors. Sequence $k = \{[y_{b1}, y_{e1}], [y_{b2}, y_{e2}], \dots\dots [y_{bK}, y_{eK}]\}$ denotes the pairs of left and right boundaries (frames) belonging to each label a_i. Due to obvious initial and continuity conditions, we get $y_{b1} = y_1$, $y_{bn} = y_{en-1} + 1$ and $y_{eK} = y_N$.

Now we have another utterance **X** with the same phonetic transcription a and feature vectors $x = \{x_1, x_2, \dots x_M\}$. We need a procedure that will find sequence $h = \{[x_{b1}, x_{e1}], [x_{b2}, x_{e2}], \dots\dots [x_{bK}, x_{eK}]\}$ that defines the phoneme boundaries in that utterance.

The problem can be solved by applying the Dynamic Time Warping algorithm. Searching for the minimum DTW distance between **X** and **Y** using the formula

$$D(\mathbf{X},\mathbf{Y}) = \min_{w} \sum_{m=1}^{M} d(x_m, y_{w(m)}) \tag{1}$$

we obtain the best alignment relation w between frames x and y and hence also the best alignment (in the sense of the DTW distance) between sequences k and h. The boundary points defining sequence h are determined as

$$x_{bi} = y_{w(bi)} \quad \text{and} \quad x_{ei} = y_{w(ei)} \tag{2}$$

The illustration of the principle is in Fig.1.

In practice, the two utterances X and Y, even if they have the same transcription, may differ significantly in their length and timing. Therefore the relation w (named the DTW path) should have the most general form allowing an arbitrary number of vertical, diagonal and horizontal transitions in the DTW plane. In such case the w is not a function and the interpretation of the equation (2) must be slightly modified, namely for special (though rare) situations when y_{bi} (or y_{ei}) projects itself to more than one x frames, or when more than one boundary points of signal y project to the same x frame. These situations can be solved by applying additional rules that determine the boundary points unambiguously and also respect a minimum phoneme length.

3.2 HMM based method

The HMM segmentation method perform on a very similar principle. Here, instead of the template signal we employ a composite model **M** created by concatenating HMMs of all the symbols occurring in the transcription of the signal to be segmented. This long HMM is matched to the speech signal **X** using the Viterbi algorithm. It will find the most probable alignment (so called Viterbi path) between the signal and the states of the composite model. The Viterbi procedure defines the probability of **X** given model **M** by formula (3):

$$P(\mathbf{X},\mathbf{M}) = \max_{w} \sum_{m=1}^{M} t_{w(m)w(m-1)} p_{w(m)}(x_m) \tag{3}$$

where the w is the Viterbi sequence of model states maximising the P(**X**,**M**), $t_{w(m)w(m-1)}$ is the probability of the transition from the state visited (on path w) in frame m-1 to the state aligned to frame m and $p_{w(m)}(x_m)$ is the probability that vector x_m is emitted by the latter state.

Again, like in the DTW method, we can identify the moments when the Viterbi path enters and leaves individual phoneme models and project them onto the x axis. In this way we obtain the sequence h that determines the most probable phoneme boundaries. The method is illustrated in Fig.2.

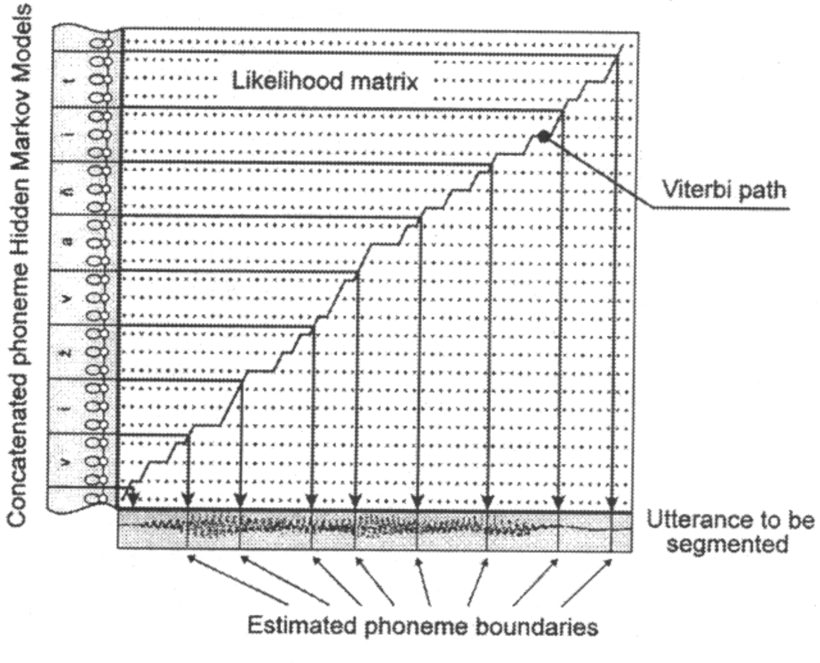

Concatenated phoneme Hidden Markov Models
v
i
ž
v
a
ň
i
t
Likelihood matrix
Viterbi path
Utterance to be segmented
Estimated phoneme boundaries

Fig. 2. Phonetic segmentation based the HMM technique

4. Practical issues

The database segmentation project was performed in the following steps:

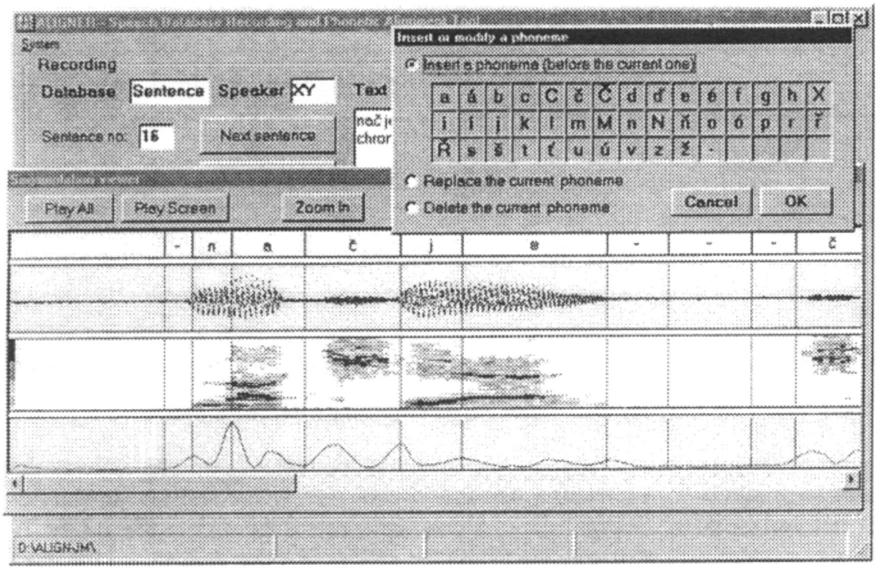
A. A specially designed sentence containing all Czech phonemes was recorded and manually segmented. The most representative instances of each phoneme were stored under the name *ABC set*.

B. The same speaker (MJN - man, one of the authors of this paper) recorded the complete set of the 82 sentences. (This set was named *MJN1*.) Each sentence was automatically pre-segmented using an artificial reference template made by concatenating the parameterised signals of the corresponding phonemes taken from the ABC set. Each phoneme boundary was checked visually and acoustically and some manual corrections were made.

C. A software tool for semi-automatic phonetic alignment and labelling was developed. The tool, named Aligner, helps a user by automating most of the procedures associated with recording and processing the speech database - see subsection 4.1.

D. Ten male and ten female speakers recorded the complete set of the sentences. Their data was segmented by means of the Aligner. Each processed sentence was checked and the labels and their positions were manually corrected if necessary.

E. An initial set of phoneme HMMs (the *HMM1 set* - standard left-to-right mod-



Database: Sentence
Speaker: XY
Sentence no: 15
Text: nač je… chror…

Insert or modify a phoneme
● Insert a phoneme (before the current one)
○ Replace the current phoneme
○ Delete the current phoneme

Fig. 3. A snapshot of the Aligner tool with check and modify windows open

els with 3 states and 4 mixtures) was trained on the available data and used for the fully automatic (unsupervised) segmentation of the sentences recorded by another group of twenty speakers. Both the new and previous sets were employed in training more accurate 3-state-16-mixture models (the *HMM2 set*).

F. All later acquired speech data were automatically segmented using the HMM method and the HMM2 set.

4.1 Aligner - tool for segmentation

The Aligner provides assistance to those who repeatedly record, segment and check speech databases. It simplifies recording by picking up and displaying sentences from the given list, it generates their phonetic transcription, computes speech signal parameters and produces segmentation based on the DTW (or HMM) method.

The segmented and labelled signal can be checked in many ways. The user can listen to each labelled segment, modify its boundaries as well as its label. He can also use visual cues for fast estimation of the phoneme boundaries, such as the signal waveform, the spectrogram and the contour of the spectral variation function (SVF). The latter provides a measure of local changes in the signal spectrum [7]. The Aligner automatically identifies SVF peaks and moves the estimated boundaries towards the nearest SVF peaks provided they are no further than ± 1 frame. The positions of the peaks are marked (by white lines), which helps the user in manual fine-tuning.

5. Segmentation accuracy

In order to evaluate the accuracy of the proposed phonetic alignment procedures we performed a series of experiments. These experiments should have compared manually positioned boundaries with those determined by the two basic (DTW and HMM) methods and various reference sets.

We chose 2 speakers (1 man and 1 woman) who recorded the whole sentence list two times. Both the repetitions of each sentence were segmented by an expert with the aid of the Aligner. (The smallest unit for moving the phoneme dividing markers was one frame. i.e. 10 ms.) In a series of experiments we evaluated and compared five different approaches:

1. DTW segmentation using the ABC reference set.
2. DTW segmentation using the MJN1 reference set.
3. DTW segmentation employing the second recording of the same speaker as the reference. (For example, the ZJM1 data was processed using the already segmented ZJM2 data.)
4. HMM segmentation based on the HMM1 model set.
5. HMM segmentation based on the HMM2 model set.

For each test speaker we measured the shift between the phoneme boundaries set by the expert and those deter-

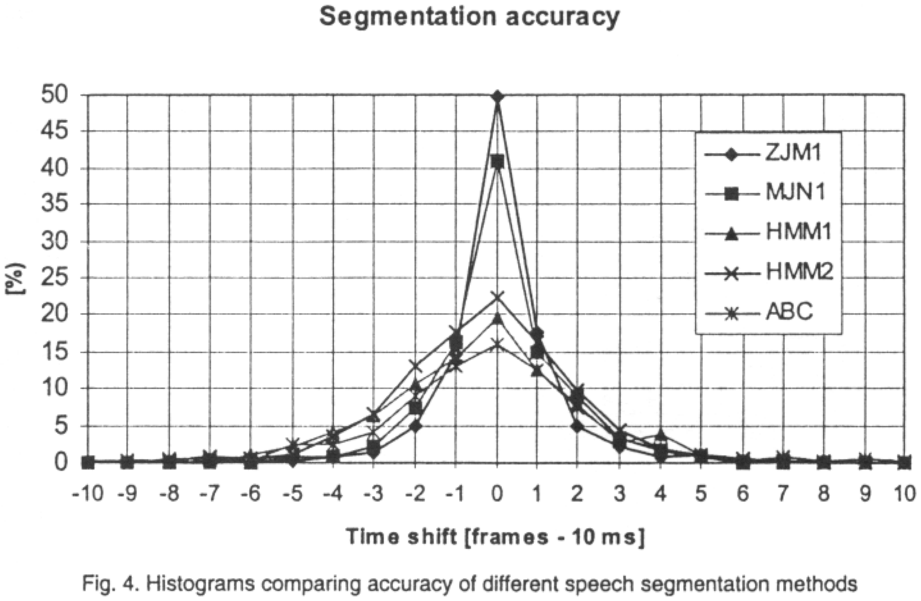


Fig. 4. Histograms comparing accuracy of different speech segmentation methods

mined by the above five methods. The results were given form of histograms that show the frequency of boundary shifts of different lengths. In Fig.4 we can see the histogram for the female speaker (ZJM), the histogram for the male speaker was similar with slightly better values.

From Fig.4 it is evident that the best results were achieved when the labelled data of the same person were used as the reference. In this case about 50 % of the boundaries were placed correctly (precisely speaking: in accordance with the expert) and about 82 % of them were shifted no further than ±1 frame from the correct position. Of course, this very high agreement rate can be achieved only if one speaker provides multiple repetitions of the same sentence. But even in case the segmentation was based on other speaker's (MJN's) data, the DTW method performed quite well. (About 72 % phonemes boundaries lie within the ±1 frame range).

The previously mentioned two approaches represent the ideal case when reference utterances have the same transcription as the labelled ones. The other three approaches cannot benefit from this fact and that is why they achieve a worse performance. On the other side they are database independent because they employ either phoneme templates (the ABC set) or models. Fig. 4 shows that the DTW segmentation with the ABC set yields the worst results (about 39 % and 65 % boundaries within the ±1 and ±3 frame range, respectively). However, they are still good enough to be applied in the initial stage of the phonetic segmentation when no other methods are applicable. On contrary, the HMM based methods perform the better the more already segmented data is available. We can see it in Fig.4 if we compare the segmentation accuracy achieved with the HMM1 and HMM2 sets. In the latter case more than 90 % of all the boundaries lie within the ±3 frame range, which is an acceptable rate, in particular, for speech recognition purposes.

6. Applications

Our work on the spoken data segmentation was motivated mainly by a long-term research in automatic speech recognition (ASR). However, some of the methods we deal in sections 3 and 4 found applications also in other related projects.

6.1 Phoneme based speech recognition

The database mentioned in section 2 was processed exactly in the way described in section 4. We obtained a large training material, in which each phoneme was represented by more than 1000 instances. This allowed us to train high-quality models of 42 Czech phonemes (including a model of background noise). Recently we prefer using context-independent (monophone) models with a higher number of mixtures (16 or 32) that are more convenient for real-time applications. The models were tested in speaker-independent isolated-word tasks with vocabularies containing hundreds to several thousands Czech words (e.g. Czech calendar names or city names). The results are shown in Table 1. For details, see [11].

Database (size)	Recog. rate [%]	Recog. time [ms]
Names (360 words)	93.2	210
City names (5346 words)	91.6	845

Table 1. Results from speech recognition tests with 3-state 32-mixture monophone models trained on the phonetically segmented database (June 2000).

6.2 Telephone speech recognition

A telephone signal has characteristics much different and the overall quality significantly lower if compared to the signal acquired in laboratory room conditions by a standard microphone. A special training database containing telephone speech is thus required for building a telephone ASR system. However, recording and post-processing such a database is an expensive job. Therefore we made a test to demonstrate that a database with appropriate characteristics could be created in a more economical way.

We accomplished it by transferring the whole existing database through a telephone line. The data was replayed, sentence by sentence, from a loudspeaker to an adjacent telephone microphone and sent through local and distant switchboards to the computer with a telephone board. In this way we obtained recordings that had most characteristics of true telephone speech. Automatic phonetic segmentation of the new signal has been accomplished using correlation with the previously labelled original signal. After training a new set of models, the recognition results yielded a level that was only 1 - 3 % lower compared to the tests done with the same vocabulary but microphone speech.

6.3 Application in speech therapy and foreign language training

The methods described in this paper can be applied also in education, namely for learning either the native or a foreign language. Several studies proved that training the pronunciation of difficult words or phrases is more successful if the trainee can see - not only hear - his/her speech and compare it with correct reference examples.

We applied this principle when developing a prototype of a teaching aid for hearing-impaired persons. The tool, named VICK (VIsual feedbaCK), is a PC based system that compares speech of the trained subject to speech that is pre-recorded or provided on-line by a therapist and displays both the signals and their spectrograms on the screen. In this way the missing acoustic information is transformed into a visual one. In the VICK's latest design [12] we used the segmentation methods to visualise the signal together with its phonetic transcription. In this way the trainee has a better chance to understand what is on the screen and where (in which part of the utterance, word, syllable or a phoneme) his/her pronunciation or intonation differs most from the reference templates. The latest design of this tool is depicted in Fig.5.

The VICK has been adapted also for learning a foreign language [13]. We have conducted a series of tests in which students used the tool in practicing pronunciation and intonation of English. In this application, the phonetic segmentation was based on the DTW method and a version of the ABC set extracted from a speech of a native English speaking person. For intonation training a set of prerecorded sentences from the same speaker was employed. (See Fig. 5b.)

7. Conclusions

In this paper we focused on the theoretical and practical issues of the phonetic speech segmentation problem. We offer several approaches to solving this problem, from a very simple one that can be used as a starting point in creating an annotated speech database to more complex methods that allow for a fully automated processing of a large speech corpus. For each method we have made a qualitative evaluation of its segmentation accuracy, which shows that even the simpler methods provide results that are acceptable in speech analysis/recognition tasks.

We are aware of the fact that the reverse task, the text-to-speech (TTS) synthesis, requires a more accurate segmentation, in which the maximum allowed shift of phoneme boundaries is measured in milliseconds. On the other side, the amount of data that must be processed to create a phonetic inventory for a TTS system is relatively small and therefore an automatic pre-segmentation followed by manual correction seems to be an adequate solution.

We hope that our experience in building a large phonetically segmented speech database may be useful and inspiring for other people or groups who will face a similar task. Those interested in acquiring the segmentation tool, the Aligner, may contact the first author.

References:

[1] Kvalle K.: Segmentation and Labelling of Speech. PhD Thesis, University of Trondheim, 1993.

[2] Karjalainen M., Altosaar T, Huttunen M.: An Efficient Labeling Tool for the Quicksig Speech Database. Proc. of Int. Conf. on Spoken Language Processing (ICSLP'98), Sydney, Dec. 1998, pp. 1535-1538

[3] Malfrere F., Deroo O., Dutoit T.: Phonetic Alignment: Speech Synthesis Based vs. Hybrid HMM/ANN. Proc. of Int. Conf. on Spoken Language Processing (ICSLP'98), Sydney, Dec. 1998, pp. 1571-1574.

[4] van Santen J.P.H, Sproat R.W.: High-Accuracy Automatic Segmentation. Proc. of Eurospeech'99, Budapest, Sept. 1999, pp.2809-2812

[5] Husson J.L.: Evaluation of a Segmentation System Based on Multi-Level Lattices. Proc. of Eurospeech'99, Budapest, Sept. 1999, pp.471-474

[6] Pollák P., Vopička J., Hanžl V., Sovka P.: CAR2 - Czech Database of Car Speech. Radioengineering, Dec.1999, Vol.8, No.4, pp.1-6

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

[8] Hájek D.: A System for Continuous Speech Recognition. MSc thesis (in Czech). Technical University of Liberec, May 1998.

[9] Deller J.R, Proakis J.G., Hansen J.H.L.: Discrete-Time Processing of Speech Signals. Macmillan, New York, 1993.

[10] Psutka J.: Komunikace s počítačem mluvenou řečí. Academia. Praha, 1995.

[11] Nouza J.: A Czech Large Vocabulary Recognition System for Real-Time Applications. Proc. of the Int. Workshop on Text, Speech, Dialogue. Springer-Verlag Heidelberg, 2000, pp. 217-222.

[12] Nouza J: Training Speech Through Visual Feedback Patterns. Proc. of 5th Int. Conference on Spoken Language Processing (ICSLP'98). Sydney, Dec. 1998, pp.3293-3296

[13] Nouza J.: Computer-Aided Spoken-Language Training with Enhanced Visual and Auditory Feedback. Proc. of Eurospeech'99, Budapest, Sept. 1999, pp.183-186

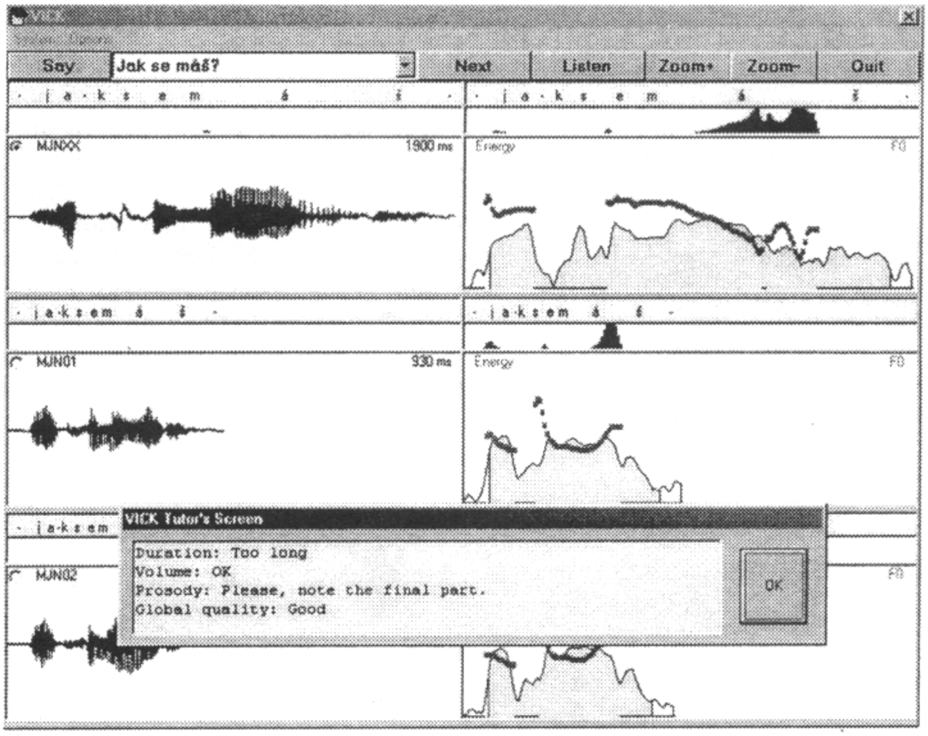
Acknowledgments

The research reported in this paper got support from Grant Agency of the Czech Republic (102/96/KO87), Czech Ministry of Education, through research goal project MSM 242200001 and by project Kontakt no. ME293.

About authors...

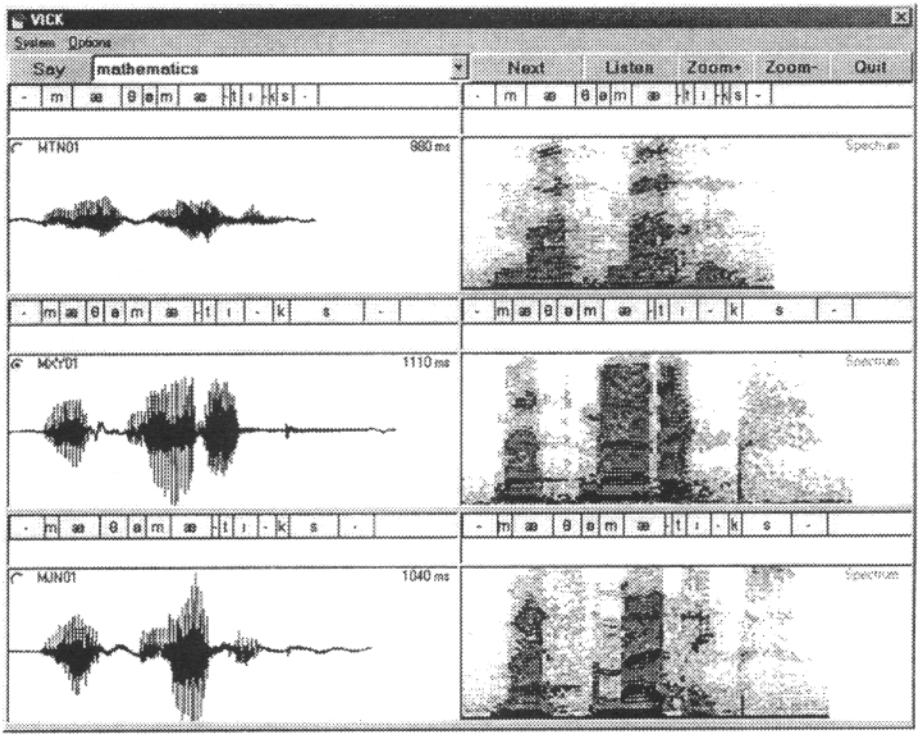
Jan NOUZA was born in Ústí n.L. in 1957. He received his master degree (1981) and doctor degree (1986), both in radioelectronics, at the Czech Technical University (Faculty of Electrical Engineering) in Prague. In 1986 he became a lecturer at the Technical University of Liberec (TUL). Currently he is a professor at the Department of Electronics and Signal Processing and vice-dean of the Faculty of Mechatronics. His major research field is speech processing with a special focus on its recognition. He has founded the Speech Processing Laboratory (SpeechLab) at the TUL. He is a member of IEEE (Signal Processing Society) and International Speech Communication Association (ISCA).

Martin MYSLIVEC was born in Hořice v Podkrkonoší in 1973. In 1998 he received master degree at the Technical University of Liberec (TUL) and joined the SpeechLab team as a PhD student. His research work is focused on continuous speech recognition of Czech language.



Duration: Too long
Volume: OK
Prosody: Please, note the final part.
Global quality: Good

a)



b)

Fig. 5. The VICK (VIsual feedbaCK) tool developed at Technical University of Liberec utilises speech segmentation methods for computer assisted language learning. It can be applied either for speech therapy of hearing impaired persons or for improving pronunciation and intonation in a foreign language.

In Fig.5a the application in speech therapy is demonstrated. A sentence spoken by a subject (in the top panel) is visually compared with the same sentence spoken by reference speakers (the lower two panels). The subject can immediately see in which parameters his attempt differs most, e.g. in sentence duration and timing, in volume and in intonation curve. Fig. 5b offers a snapshot of a foreign language practicing session. An English word „mathematics“ spoken by a Czech student (above) is compared to the same words recorded by native speaker.)