Phone Segmentation Tool with Integrated Pronunciation Lexicon and Czech Phonetically Labelled Reference Database

Petr Pollák[†], Jan Volín[††], Radek Skarnitzl[††]

[†] Czech Technical University in Prague, Faculty of Electrical Engineering
Technická 2, 166 27 Praha 6, Czech Republic
pollak@fel.cvut.cz

[††] Charles University in Prague, Institute of Phonetics,
nám. J. Palacha 2, 116 38 Praha 1, Czech Republic
jan.volin@ff.cuni.cz, radek.skarnitzl@ff.cuni.cz

Abstract

Phonetic segmentation is the procedure which is used in many applications of speech processing, both as a subpart of automated systems or as the tool for an interactive work. In this paper we are presenting the latest development in our tool of automated phonetic segmentation. The tool is based on HMM forced alignment realized by publicly available HTK toolkit. It is implemented into the environment of Praat application and it can be used with several optional settings. The tool is designed for segmentation of the utterances with known orthographic records while phonetic contents are obtained from the pronunciation lexicon or from orthoepic record generated by rules for new unknown words. Second part of this paper describes small Czech reference database precisely labelled on phonetic level which is supposed to be used for the analysis of the accuracy of automatic phonetic segmentation.

1. Introduction

Phonetic segmentation is a task which leads to a number of applications in different speech technology systems. The extraction of phones from an utterance is typically needed during speech identification or verification, in speech synthesis systems, and often also for some training purposes as neural network training, LDA-based classes, or sometimes also for HMM training, etc. The need of such tools is self-evident. Therefore, we have created a basic version of the tool based on standard HMM forced alignment. This tool was implemented in the Praat environment and it was also used for the purposes of automatic pre-segmentation before further manual labelling on phonetic level (Pollák et al., 2007).

During previous activities we also analyzed the accuracy of HMM based phonetic segmentation from different points of view as short-time analysis settings, HMM modelling settings (modelling of some rare phones or modelling with skips over states), using different feature extraction techniques, etc. We observed that the above-mentioned technique produced quite satisfactory average results but for particular situations phoneme boundaries could have been placed with significant errors (Pollák et al., 2005). In the current study we present the small extension of the existing tool which can utilize different parameterization techniques, different input data formats, or different sets of HMMs.

Our segmentation tool is designed for locating phone boundaries in known utterances, i.e. we possess a orthographic record for each utterance and we do not require recognition of the linguistic content. On the other hand, we do not know exact phonetic forms so we work with predicted phonetic contents. In this work we want to present procedure which will maximize correct prediction of real pronunciation of analyzed utterances. For this purpose we have completed large pronunciation lexicon from several sources available for the Czech language.

The second important part of this paper describes the creation of precisely phonetically labelled speech database for the evaluation purposes. The main motivation for this work was the need of an improvement of testing setup for further investigating of different post-processing algorithms for automated corrections of boundaries which are set in the first step by HMM based segmentation algorithm.

2. Segmentation algorithm and tool

As mentioned above, the segmentation procedure is based on forced alignment of trained HMM models. For this purpose we need to realize following steps:

1. the choice of proper features describing speech signal,
2. the training of HMM models,
3. the prediction of real utterance pronunciation,
4. the development of a tool with user friendly interface.

Particular solutions of these objectives were realized during previous research and within this work we present extensions in each of presented tasks.

2.1. Speech features

Generally, the mel-frequency cepstral coefficients (MFCC) are the most frequently used features for recognition purposes. For high-quality data with minimal noise background better results can be achieved using PLP cepstral coefficients. On the other hand, when the data contain higher background noise level, some technique removing this additive noise can be used in parameterization of speech. Frequently, we have to solve the mismatch in speech input channels, i.e. different convolution distortion may be present in training and recognition sets and it is reasonable

to perform normalization. The following speech features were used for better description of speech in the above-mentioned particular situations:

- standard set of MFCC and PLP features as a baseline systems,
- choice of different short-time analysis setups,
- possible elimination of noise by frequency-domain suppression techniques,
- possible usage of cepstral mean subtraction (CMS) for channel normalization,
- work with 8 kHz and 16 kHz speech signals,

2.2. HMM modelling and training of HMMs

A quite standard setup of HMM modelling is used in our tool, i.e. standard left-right 3-emitting state HMMs with no skips over states. Emitting functions contain 32 Gaussian mixtures and models are processed in 3 independent streams for static, dynamic, and acceleration parameters (i.e. for delta and delta-delta features which are always used in all situations).

As to the acoustic elements, 45 Czech monophones were used for HMM modelling with special effort devoted to training of glottal plosives and schwa, which do not have a phonemic status in standard Czech pronunciation but which appear in colloquial speech, see (Pollák et al., 2007) and (Wells and et al., 2003).

HMMs were trained on large Czech databases collected under different conditions. The training data were from Czech SpeechDat(E), SPEECON, car speech DB, and phonetic DB. Models created from several databases guarantee maximal match of conditions in training and segmentation (recognition) phase. We do not use any adaptation technique, possible mismatch is supposed to be minimized by suitable choice of HMMs.

2.3. Upgrades in Praat tool

Our tool was implemented into the environment of Praat program, see (Boersma and Weenink, 2008). We completed the above-mentioned optional settings of segmentation parameters. Currently we can choose from standard Praat menu proper settings of following parameters:

- sampling frequency (8 kHz and 16 kHz),
- parameterization technique (MFCC, PLP, short-time segmentation, CMS, noise suppression, etc.)
- proper lexica or sublexica.

The setting of these parameters from standard Praat menus provides a simple and clear interface for the user control of automated phonetic pre-segmentation. Tool can be invoked in the moment when two proper objects are selected, i.e. TextGrid and related sound as it can be seen on fig. 1. When Praat script is activated, standard Praat interface is used for the setting of optional parameter, see fig. 2.

Output of the scripts is the TextGrid with time and frequency representation of analyzed sound, see fig. 3. Our TextGrid file contain 4 layers:

1. *layer: "RefPhones"* - manually set phone boundaries,
2. *layer: "AutoPhones"* - generated phone boundaries,
3. *layer: "Words"* - generated word boundaries,

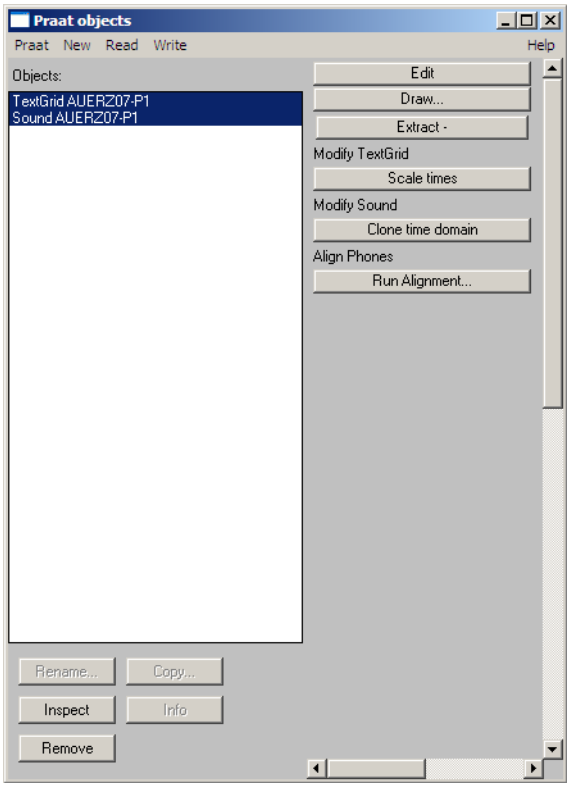

Figure 1: Praat object with phonetic segmentation tool

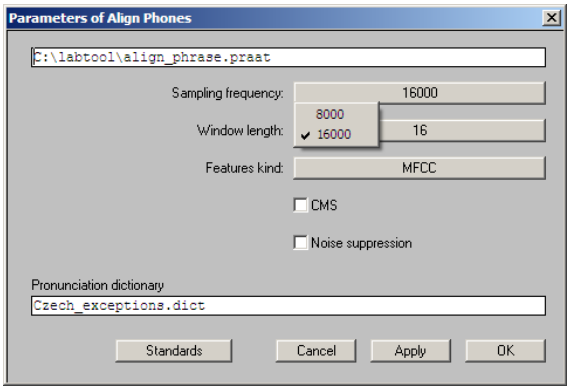

Figure 2: Praat script parameters for phonetic segmentation

4. *layer: "Phrase"* - input orthographic transcription.

Layers No. 2 and 3 are always automatically generated by the segmentation algorithm from the layer No. 4. When the layer No. 1 is empty or identical as the layer No. 2, both these layers are created commonly. Otherwise the original content of layer No. 1 remains unchanged. It is typically in situations when phone boundaries were already manually adjusted.

2.4. Prediction of pronunciation - creation of large lexicon

Phonetic content related to known orthographic record is created for each segmented utterance on the basis of following three steps:

1. the word is searched in large pronunciation lexicon which contains also possible variants of word pronunciation,

2. for new unknown words, rules based tool is used for generation of the regular word pronunciation (Pollák and Hanžl, 2002),
3. finally, for words with exceptional pronunciations (mainly words of foreign origin) which are not in the lexicon yet, irregular pronunciation can be specified manually by special syntax, i.e. (word/pronunciation), see (Pollák et al., 2007) and (Pollák and Hanžl, 2002).

The key role in the prediction of utterance pronunciation is played by the lexicon mentioned in the first item. We have created very large pronunciation lexicon containing reasonable amount of the most frequent words of Czech with possible multiple pronunciations. In this lexicon data from three very large database collections are included, i.e. from lexica of Czech SpeechDat(E) and SPEECON databases, see details in (Černocký et al., 2000) and (Pollák and Černocký, 2003) and from major part of currently created Czech lexicon within LC-StarII project (Moreno, 2008). Currently our lexicon contains more than 100,000 lexical items which means reasonable coverage for our purposes. However, not all word forms are present in the lexicon for each lemma (which might not be sufficient for LVCSR). Given pronunciations in source lexica were extended by possible pronunciation variants derived on the basis of inter-word context dependency and also with respect to fast and more colloquial pronunciation.
Due to limited license we are not able to distribute within the tool this large pronunciation lexicon. But the lexicon covering the most important irregularities in pronunciation is publicly distributed within the tool. This restriction in used lexicon in public version does not limit the functionality of the tool as observed pronunciation irregularities can be marked interactively by the syntax mentioned above.
Also other lexicon can be used as second solution. As the labelling tool uses standard tools from HTK toolkit (Young and et al., 2005), used pronunciation lexicon should have HTK format using standardized SAMPA symbols for Czech phones according to (Wells and et al., 2003). Different pronunciation lexicon can be specified commonly with other options of used Praat script.

3. ALIGN1CS - Czech phonetically labelled reference database

Within the research in the field of automated phonetic segmentation, a reference phonetically labelled database is required for evaluation purposes. As another important result of this work we have created such database with particular subsets containing the data collected under different conditions.

3.1. Data blocks in ALIGN1CS

Particular subsets of ALIGN1CS were carefully selected to guarantee sufficient coverage of phonetic contents, different quality of speech data, and different noise backgrounds. Selected subsets should guarantee statistical significance of tests realized with this DB.

3.1.1. Wide-band data from real environments

The first subset consists from phonetically balanced material and digits collected within SPEECON project in different environment types. Our subset contains signals collected by high quality head-set microphone. Sampling frequency is 16 kHz in this case.
Data are organized into two particular blocks, i.e. utterances with rather small level of background noise are in block 'HEAD0' and slightly more noisy utterances are in block 'HEAD1'.

3.1.2. Telephone speech data

Second subset contains telephone speech data sampled at 8 kHz and utterances containing phonetically rich material and digit sequences are chosen to this selection. The source of this data is Czech SpeechDat database. Similarly as above mentioned SPEECON data, this subset is organized in two parts according to SNR, i.e. rather clean data are in block 'TELE0' and more noisy data are in block 'TELE1'.

3.1.3. High quality speech for phonetic research

This subset contains the material selected from the Prague Phonetic Corpus. It contains high quality 32 kHz recordings of a text read by 20 university students. It involves a short meaningful text (each about 220 phones) describing an interaction of a schoolboy with his grandmother. The recordings were made in a soundproof booth under identical conditions. No noisy data are supposed to be recorded so only one block FUPE0 is in the database for this subset.

3.2. Phone statistics of selected data

We have chosen 50 phonetically rich sentences, 40 phonetically rich words, and 10 digit sequences into each block HEAD0, HEAD1, TELE0, and TELE1. Phonetically rich material should guarantee well coverage of all phones, especially sufficient appearance of rare phones. Digit sequences are used for the representation of utterances with longer inter-word pauses.
For selected data we have evaluated achieved appearance rates for all particular phones and for phones organized in particular groups. From the point of view requirements of phonetic research following two categorization of phones are used:

Variant 1 of phone grouping

vowels: "a, a:, e, e:, i, i:, o, o:, u, u:, o_u, a_u, e_u, @",

fricatives: "f, v, s, z, S, Z, P\, Q\, x, h\",

affricates &plosives:
"t_s, t_S, d_Z, d_z, p, b, t, d, c, J\, k, g, ?",

sonorants: "m, F, n, J, N, r, l, j"

Variant 2 of phone grouping

vowels high: "i, i:, u, u:",

vowels non-high: "a, a:, e, e:, o, o:, o_u, a_u, e_u, @",

fricatives & affricates:
"f, v, s, z, S, Z, P\, Q\, x, h\, t_s, t_S, d_Z, d_z ",

plosives: "p, b, t, d, c, J\, k, g, ?",

nasals: "m, F, n, J, N"

approximants: "r, l, j"

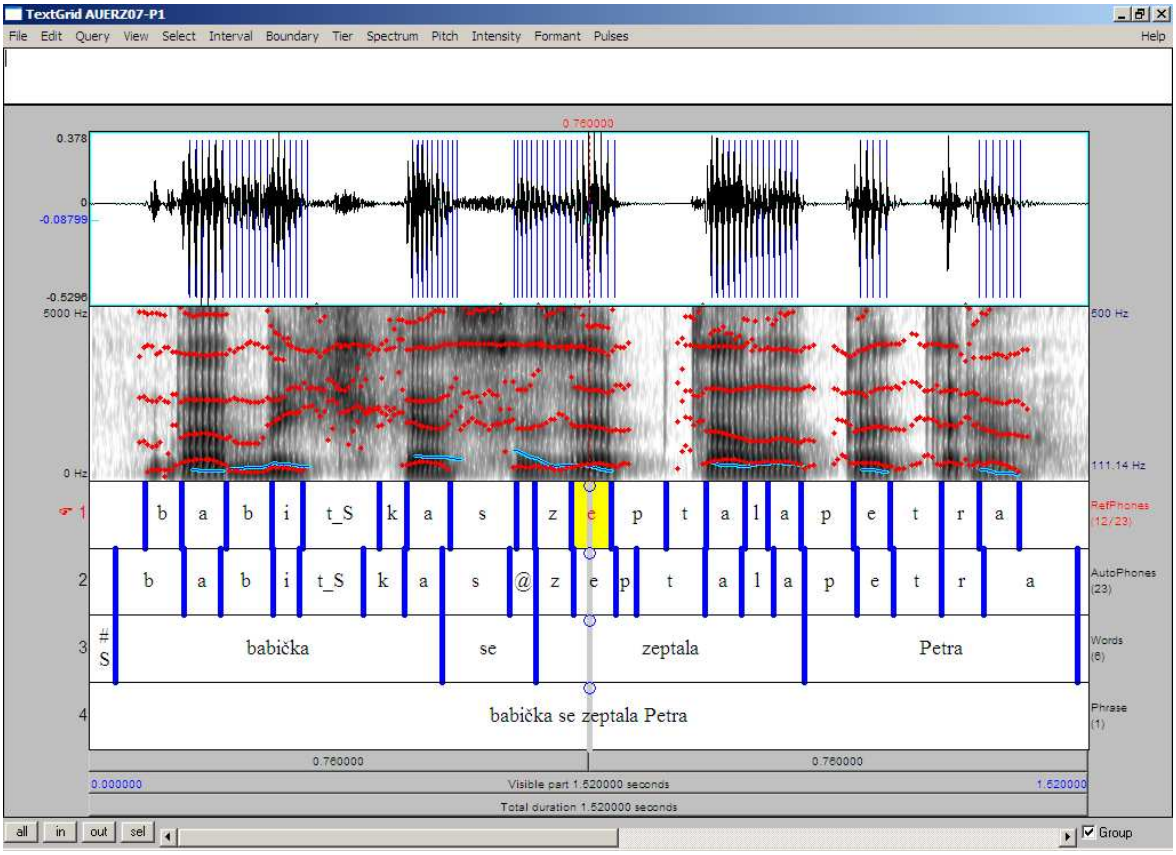

Figure 3: Example of resulting window with phonetic labels

SUBSET	HEAD0	HEAD1	TELE0	TELE1	FUPE0
phones	**2842**	**2882**	**2985**	**3011**	**4240**
affricates & plosives	622	633	618	642	1060
fricatives	423	440	476	487	480
sonorants	605	620	657	650	900
vowels	1192	1189	1234	1232	1800

Table 1: Phone rates grouped according to variant 1

SUBSET	HEAD0	HEAD1	TELE0	TELE1	FUPE0
phones	**2842**	**2882**	**2985**	**3011**	**4240**
approximants	310	304	330	307	440
fricatives & affricates	538	550	573	579	560
plosives	507	523	521	550	980
nasals	295	316	327	343	460
vowels high	406	376	419	444	400
vowels non-high	786	813	815	788	1400

Table 2: Phone rates grouped according to variant 2

The statistics of all phone appearances in particular subsets are saved in database structure as files PHSTATS.TXT. For general overview, the statistics for particular groups of phones defined above are presented in tables 1 and 2.

3.3. SNR of selected signals

Also the information about noise level in signals from subsets HEAD0, HEAD1, TELE0, and TELE1 is extracted from original databases and it is saved in the files SNRTABLE.TXT. Presented Signal-to-Noise Ratios (SNRs) were estimated during database collection. The same value of SNR in particular subsets may represent slightly different real noise level as SpeechDat and SPEECON data has slightly different quality and also slightly different algorithms of SNR estimation were used, for details see (Černocký et al., 2000) and (Pollák and Černocký, 2003). But this small inconsistency does not influence the grouping of data according to noise level for our purposes and the overview about the noise level in particular data blocks is presented in figures 4 and 5.

3.4. Labelling on phonetic level

All utterances were precisely labelled on phonetic level with maximal effort to specify precisely both correct phonetic contents of the utterance and the placement of phone boundaries.

The information is saved in Praat TextGrid-file and also in HTK formated lab-file. Praat TextGrid file is supposed to be used preferably within an interactive manual analysis of given speech data. HTK lab-files are supposed to be used mainly for the classification of automated phonetic segmentation accuracy.

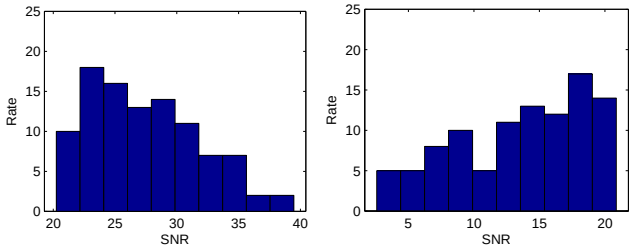

Figure 4: SNRs in HEAD subsets of ALIGN1CS database

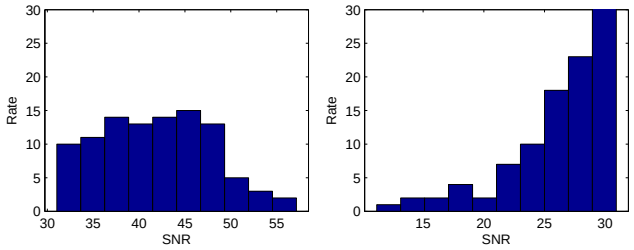

Figure 5: SNRs in TELE subsets of ALIGN1CS database

3.5. Database structure

The database ALIGN1CS has very simple structure based on separation of signals into particular blocks. As we are working generally with different sampling frequency and as for wide-band data some down-sampling can be assumed, the data are structured also according to sampling frequency. The label files which are independent on sampling frequency are saved in the directory LAB. Current structure of our database is as follows.

```
ADULT1CS
  |-- HEAD0
  |      |-- 16K
  |      |-- LAB
  :      :
  |-- TELE0
  |      |-- 8K
  |      |-- LAB
  :      :
  |-- FUPE0
         |-- 32K
         |-- LAB
```

4. Conclusions

In this paper we presented the new developments in our phonetic segmentation tool in Praat environment together with the description of reference database supporting either further more precise testing of automated segmentation algorithms or general phonetic research. The most important contributions of this work can be summarized in following points:

- New version of our Praat-based tool is presented where settings of several optional parameters can be chosen. The control is very simple, user friendly, and in compliance with standards used in Praat environment. The tool works with good precision and it is publicly available via our WEB site *http://noel.feld.cvut.cz/speechlab* in the section *Download*.
- The pronunciation lexicon is an important part of our segmentation tools. Presently, public distribution contains the most important lexical items with possible irregular pronunciation. User defined pronunciation lexicon can be used specifying it in Praat script options. It is convenient especially when user has available larger pronunciation lexicon.
- The reference database for testing of accuracy of automated phonetic segmentation were created. Speech data were selected from the existing speech databases, but all selected utterances were precisely manually relabelled. This database is also publicly available via our WEB-page *http://noel.feld.cvut.cz/speechlab*.

Acknowledgement

The activities of the first author from Czech Technical University in Prague were supported by grant GACR 102/08/0707 and by research activity MSM 6840770014 “Perspective Informative and Communications Technicalities Research”.
The work of co-authors from Charles University in Prague was supported by research activity MSM 0021620825 “Variability of acoustic features in language and speech: The sources and limits from communicative viewpoint”.

5. References

P. Boersma and D. Weenink. 2008. Praat: Doing phonetics by computer. http://www.fon.hum.uva.nl/praat/.

A. Moreno. 2008. LC-StarII. Lexica and corpora for speech-to-speech translation components. http://www.lc-star.org.

P. Pollák and V. Hanžl. 2002. Tool for Czech pronunciation generation combining fixed rules with pronunciation lexicon and lexicon management tool. In *Proc. of LREC’02, Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands - Spain, May.

P. Pollák and J. Černocký. 2003. Czech SPEECON adult database. Technical report, Nov. http://www.speechdat.org/speecon.

P. Pollák, J. Volín, and R. Skarnitzl. 2005. Influence of HMM’s parameters on the accuracy of phone segmentation - evaluation baseline. In *ESSP2005, Electronic Speech Signal Processing*, Prague, Sep.

P. Pollák, J. Volín, and R. Skarnitzl. 2007. HMM-based phonetic segmentation in Praat environment. In *Proc. of SPECOM 2007*, Moscow.

J. Černocký, P. Pollák, and V. Hanžl. 2000. Czech recordings and annotations on CD’s - Documentation on the Czech database and database access. Technical report, SpeechDat(E), Nov. Deliverable ED2.3.2, workpackage WP2.

J. C. Wells and et al. 2003. Czech sampa home page. http://www.phon.ucl.ac.uk/home/sampa/czech-uni.htm.

S. Young and et al., 2005. *The HTK Book, Version 3.3*. Cambridge.