Accuracy of HMM-Based Phonetic Segmentation Using Monophone or Triphone Acoustic Model

Petr Mizera, Petr Pollak
Czech Technical University in Prague
Faculty of Electrical Engineering
K13131 CTU FEE, Technická 2, 166 27 Prague 6, Czech Republic
Email: {mizerpet,pollak}@fel.cvut.cz

***Abstract*—The paper compares the accuracy of HMM-based automatic phonetic segmentation using various signal representation same as acoustic models of various complexity, i.e. acoustic models of monophones or word-internal triphones with various number of mixtures. The precision of automatic phonetic segmentation was measured on the basis of comparison with manually segmented speech data. The analysis showed that the segmentation with acoustic models of word-internal triphones yielded to a better target accuracy. The best results of automatic phonetic segmentation were attained for acoustic models of word-internal triphones with four mixtures. In this case average values of shift of phone boundaries and change of phone length was about 5.9 ms and 0.2 ms respectively.**

***Keywords*—HMM, ASR, TTS, CMN, Phones labelling.**

I. Introduction

Automated phonetic segmentation is a problem which has possible applications in the variety of systems using speech technology. Starting with a role of helpful tool in the basic phonetic research, we can find further applications, for example in automated segmentation of very large databases where manual segmentation is not possible (e.g. speech recognition training databases [1], [2], long records served as resource of typical speech segments for concatenative synthesis [3], [4], etc.) or in the case of automated phone extraction from given utterance for further analysis (e.g. for pathological speech analysis [5]).

Phonetic segmentation can be realized using various approaches based on correlation analysis [6], Bayesian change point detector [7], etc. The approach which is used most frequently is based on HMM-forced alignment. The popularity of this technique is due to the fact that HMM modelling represents widespread approach used in speech recognition systems as it is documented by many works in this field which were already published [8], [9]. However, the precision of such a system has a limit which is determined by discrete short-time segmentation into particular processing frames but also by the quality of trained acoustic models.

This paper describes the comparison of achieved accuracy of HMM-based phonetic segmentation which was realized using acoustic models with various complexity. We analyzed the dependency of the number of Gaussian mixtures used in HMM acoustic model of monophones on achieved accuracy of phonetic segmentation same as the analogous performance in situations when acoustic models based on tied-state triphones were used.

II. HMM-based phonetic segmentation

As it was already mentioned above, phonetic segmentation analyzed within this paper is based on Hidden Markov Models (HMM), exactly on widely used forced alignment of trained HMM-based acoustic models (AM). It is rather standard and well known procedure so its principal description is not presented here. Details can be found in [10] or [11]. Just the common setup of our segmentation algorithm is summarized below.

A. Feature extraction

Rather standard setup of feature extraction was used in our approach. Mel-Frequency Cepstral Coefficients (MFCC) were used as short-time speech representation. Details of feature extraction are given in the following points:

- preemphasis coefficient of 0.97,
- Hamming window with length of 25 ms,
- window shift of 10 ms,
- cepstral mean normalization (CMN) subtracting an average computed over whole processed utterance,
- filter-bank with 22 overlapping frequency bands with triangular response for 8 kHz and 30 overlapping frequency bands for 16 kHz,
- 12 MFCCs with c_0, completed by dynamic and acceleration coefficients.

B. Acoustic Modelling

Algorithms of HMM-based phonetic segmentation work usually with a modelling of subword units at the level of monophones, i.e. context independent phones, as they represent acoustic realization of particular phonemes. The precision of such modelling is typically increased by higher number of mixtures in emitting functions of HMM states. However, precise localization of phoneme boundaries in various contexts should be described better by context-dependent triphones.

Although this context dependency is not taken into account in the final segmentation result, i.e. boundaries are set just for central phones, context-dependent modelling by triphones improves the precision of boundaries localization.

Acoustic models of monophones or word-internal triphones were created in the following setup.

Monophones:

- 43 Czech monophones, models for silence and short-pause (tee-model),
- left-right HMMs with 3 emitting states without skips,
- around $12-20$ mixture components per each emitting state,
- 1 independent streams for static, dynamic and acceleration coefficients.

Word-internal triphones:

- 9136 different tied-state triphones cloned from Czech monophones,
- left-right HMMs with 3 emitting states without skips,
- around $3-12$ mixture components per each emitting state,
- 1 independent streams for static, dynamic and acceleration coefficients.

Training of HMMs:

- training data were from rather clean office subset of Czech SPEECON [12],
- total length of speech data were approximately 51 hours,
- Baum-Welch re-estimation was used for the training.

Described phonetic segmentation was implemented by tools from widely used HTK Toolkit [11]. In the case of triphone modelling, simple postprocessing based on backward mapping of triphones to context independent monophones was performed.

III. Evaluation criteria

The accuracy of phonetic segmentation was quantified using the following criteria: *Shift of the Phone Beginning* (*SPB*), *Shift of the Phone End* (*SPE*), and *Change of the Phone Length* (*CPL*) [13] defined as

$$SPB_{ph}[i] = beg_{ph}[i] - beg_{ph,ref}[i], \tag{1}$$

$$SPE_{ph}[i] = end_{ph}[i] - end_{ph,ref}[i], \tag{2}$$

$$\begin{aligned} CLP_{ph}[i] = {} & end_{ph}[i] - beg_{ph,ref}[i] \\ & - end_{ph,ref}[i] + beg_{ph,ref}[i], \end{aligned} \tag{3}$$

where $beg_{ph}[i]$, $end_{ph}[i]$, $beg_{ph,ref}[i]$ and $end_{ph,ref}[i]$ are automatic and reference boundaries respectively. For global evaluation of phonetic segmentation accuracy, these criteria were calculated in the form of simple statistical analysis across all phones, i.e. mean values and standard deviations were computed.

As the occurrence of certain phones can be rather low in smaller testing set, particular phones were grouped into six phone classes, i.e. vowels high (*VH*), vowels non-high (*VNH*), fricatives & affricates (*FAF*), plosives (*PLO*), nasals (*NAS*) and approximants (*APP*). These groups were chosen on the basis of phonetic expertise [14], [15]. These groups of phones with their frequency of an occurrence in testing set are presented below and above mentioned criteria were then computed for these phone groups.

Phone group	*Phones*	Amount
VH	*i, i:, u, u:*	143
VNH	*a, a:, e, e:, o, o:, o_u, a_u, e_u*	297
FAF	*f, v, s, z, S, Z, P\, Q\, x, h\, t_s, t_S, d_Z, d_z*	195
PLO	*p, b, t, d, c, J\, k, g*	201
NAS	*m, F, n, J, N*	92
APP	*r, l, j*	125

TABLE I
Description of phone grouping

IV. Experiments

The accuracy of phonetic segmentation with above described AMs of various complexity was analyzed. Acoustic models were created in two frequency band variants, i.e. for wide-band speech sampled by 16 kHz and for telephone-band speech sampled by 8 kHz. Same data from SPEECON database were used both for the training and the testing, when 8 kHz telephone-band data were obtained by downsampling of 16 kHz utterances. The contribution of CMN channel normalization was also analyzed.

A. Testing speech data

Testing speech subset consisted of the selection from Czech SPEECON database. Selected utterances contained phonetically rich sentences or digit sequences, that were typical representatives of rather clean recording conditions in this subset. Of course, testing data were not part of training subset. In the end, the selection of 32 speech signals, with an approximate duration of about 3 min was used for the evaluation.

B. Acoustic models

The monophone and word-internal triphone AMs were used for our experiments. For both monophone and triphone acoustic modelling, the following four features sets were analyzed: MFCC coefficients for wide-band 16 kHz speech (*mfcc_0_d_a_16k_2510_30*), MFCCs for telephone-band 8 kHz speech (*mfcc_0_d_a_8k_2510_22*), and finally same features utilizing CMN (*mfcc_0_d_a_z_16k_2510_30*, *mfcc_0_d_a_z_8k_2510_22*).

The number of mixtures was different for each AMs used. The optimized setup was chosen on the basis of the best speech recognition accuracy achieved in the task of connected digit recognition, e.g. for *mfcc_0_d_a_16k_2510_30* this value was 19 mixture

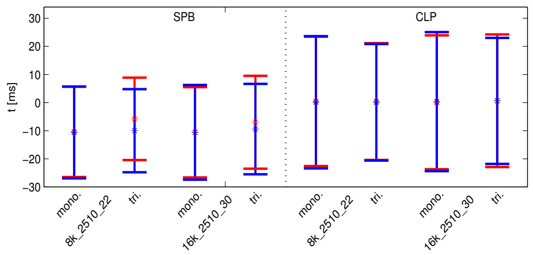

Fig. 1. Comparison of the results of global

for monophone AM and 12 mixtures for triphone AM (marked below as *mono19* and *tri12* respectively). Similar convention was also used for the description of AMs for other features setups.

C. Results

The first group of results was obtained on the basis of analysis realized for all phones together, and the next results were computed across phone grouping defined above.

C1) Global values

The global results across of all phone groups were presented in the Tab. II. The values of criteria SPB, CPL had systematically lower both mean values and standard deviation for the case of triphone-based AMs in comparison to equivalent variant of monophone AMs. These results were also presented illustratively in the Fig. 1. Red and blue colors corresponded to the case with and without CMN. However, CMN computed over whole utterance did not improve the accuracy of phonetic segmentation significantly. For more details, histograms of these values were presented in the Fig. 2. The best accuracy (*SPB -5.9 ±14.6, CPL 0.2 ±22.8*) was achieved for the variant of triphone AM with four mixtures without CMN for frequency band of $0-4000$ kHz.

The highest error in boundary localization which was quantified by SPB criterion appeared typically on word boundaries. It was illustratively presented in the Fig. 3. Reference phonetic segmentation created manually was presented by black dashed lines while automatic phonetic segmentation was presented by red color. It was possible to see that phone boundaries in the middle of words were localized usually very precisely in comparison to boundaries on word ends and beginnings.

C2) Group values

Values of above mentioned criteria were computed also for each phone group defined in the Tab. III. Generally, the worst accuracy were achieved for nasals (NAS phone group), it is possible to observe higher values of standard deviation especially for CPL criterion. Exactly, for the case of phones *m, N, N* the error of boundary location was many times higher than 80 ms. On the other hand, the best results were observed for *mfcc_0_d_a_8k_2510_22* features, where the error was below 5 ms in 50% of phone groups.

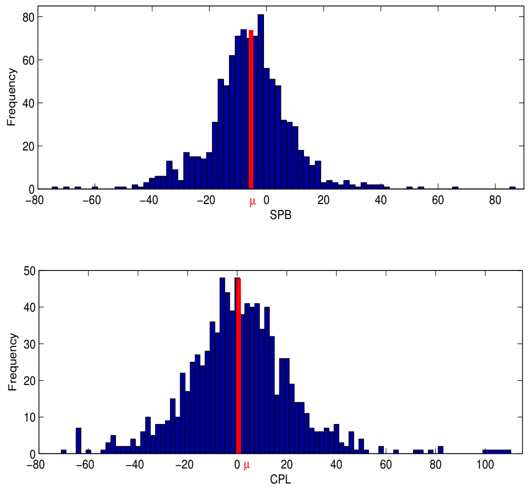

Fig. 2. Histograms of criteria SPB and CPL for setup of phonetic segmentation *mfcc_0_d_a_8k_2510_22*

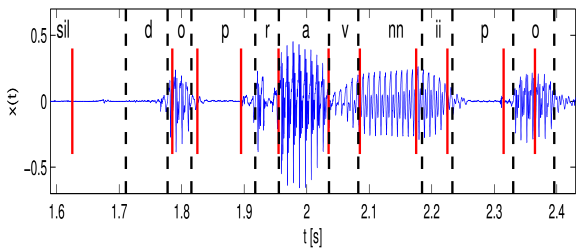

Fig. 3. Illustrative comparison of automatic and manual phonetic segmentation

V. Conclusions

The accuracy of HMM-based automatic phonetic segmentation using various acoustic models and various speech features was analysed. The most important conclusions are presented in the following points.

- The results of segmentation using more sophisticated triphone acoustic models outperformed the results for the case of context independent monophone modelling which is usually used for similar tasks.
- Achieved precision was quantified by criteria quantifying phone boundary shift and the best achieved result was about 5 ms of shift of phone beginning or phone end respectively with rather small standard deviation around 10 ms.
- The application of CMN had rather minor contribution to the improvement of phonetic segmentation accuracy but it was caused mainly by the fact that within this article the analysis was performed with rather clean data with the same convolution distortion.
- Above presented results allowed the usage of this approach for automatic phonetic segmentation within various applications, mainly as a support for basic phonetic research same as for precise and reliable automated phone boundary location in large speech databases for the purpose of neural network training which may be used in other speech technology applications.

	8k_2510_22				*16k_2510_30*			
	mfcc_0_d_a		*mfcc_0_d_a_z*		*mfcc_0_d_a*		*mfcc_0_d_a_z*	
	mono12	*tri4*	*mono15*	*tri3*	*mono19*	*tri12*	*mono20*	*tri5*
SPB [ms]	-10.2±*15.7*	-5.9±*14.6*	-10.9±*16.3*	-9.9±*16.4*	-10.3±*15.6*	-6.8±*15.8*	-10.9±*17.2*	-9.6±*15.6*
CPL [ms]	0.4±*23.1*	0.2±*22.8*	0.1±*22.3*	0.6±*23.5*	0.2±*22.8*	0.6±*22.8*	0.2±*23.9*	0.6±*22.0*

TABLE II
THE RESULT OF GLOBAL CRITERIA $[ms]$

	8k_2510_22				*16k_2510_30*			
	mfcc_0_d_a		*mfcc_0_d_a_z*		*mfcc_0_d_a*		*mfcc_0_d_a_z*	
	mono12	*tri4*	*mono15*	*tri3*	*mono19*	*tri12*	*mono20*	*tri5*
SPB [ms]								
VH	-8.4 ±*19.1*	-3.8 ±*17.1*	-10.8 ±*16.8*	-9.9 ±*19.5*	-8.7 ±*15.6*	-4.6 ±*19.7*	-9.2 ±*19.7*	-6.7 ±*19.1*
VNH	-4.2 ±*12.3*	-2.3 ±*12.3*	-3.9 ±*14.6*	-7.0 ±*12.1*	-1.7 ±*13.1*	-3.0 ±*13.3*	-2.3 ±*13.5*	-7.3 ±*12.7*
FAF	-14.4 ±*14.6*	-8.4 ±*12.9*	-15.5 ±*15.1*	-12.6 ±*13.2*	-17.8 ±*12.5*	-10.5 ±*13.7*	-18.4 ±*13.2*	-14.0 ±*13.4*
PLO	-11.2 ±*11.2*	-6.0 ±*12.5*	-12.8 ±*10.4*	-6.8 ±*17.2*	-10.8 ±*11.6*	-3.5 ±*12.5*	-11.5 ±*11.7*	-6.2 ±*12.8*
NAS	-12.6 ±*11.8*	-4.9 ±*14.6*	-12.2 ±*15.0*	-9.9 ±*14.7*	-13.8 ±*12.9*	-5.7 ±*14.5*	-13.2 ±*14.2*	-6.9 ±*16.0*
APP	-16.6 ±*17.7*	-13.2 ±*14.2*	-16.4 ±*18.5*	-17.7 ±*17.0*	-17.0 ±*16.8*	-17.9 ±*14.1*	-19.0 ±*19.2*	-19.2 ±*13.8*
SPE [ms]								
VH	-8.1 ±*14.5*	-4.4 ±*13.2*	-10.2 ±*14.5*	-7.3 ±*14.1*	-9.7 ±*14.7*	-5.3 ±*14.6*	-11.6 ±*14.5*	-9.3 ±*13.0*
VNH	-14.7 ±*13.1*	-9.2 ±*11.3*	-15.5 ±*12.8*	-13.0 ±*12.7*	-17.1 ±*12.1*	-10.5 ±*13.2*	-17.5 ±*14.2*	-11.5 ±*13.1*
FAF	-10.4 ±*13.4*	-8.2 ±*12.1*	-10.6 ±*14.5*	-12.6 ±*17.4*	-8.2 ±*13.3*	-9.4 ±*12.7*	-9.2 ±*13.8*	-11.8 ±*11.9*
PLO	-13.0 ±*12.7*	-9.4 ±*9.2*	-13.2 ±*14.9*	-13.4 ±*9.7*	-11.2 ±*13.0*	-9.9 ±*12.3*	-11.8 ±*14.0*	-15.0 ±*13.0*
NAS	2.0 ±*16.8*	4.9 ±*14.9*	0.0 ±*14.0*	-2.1 ±*13.9*	0.3 ±*14.4*	1.6 ±*14.9*	0.0 ±*13.0*	-1.4 ±*15.5*
APP	-5.9 ±*12.3*	1.9 ±*14.0*	-5.5 ±*15.5*	-1.5 ±*17.0*	-4.1 ±*14.4*	3.5 ±*15.1*	-4.0 ±*21.9*	0.0 ±*11.7*
CPL [ms]								
VH	1.6 ±*20.0*	0.1 ±*19.4*	0.6 ±*19.2*	2.4 ±*19.2*	-1.2 ±*21.0*	0.9 ±*20.1*	-3.0 ±*19.8*	-1.6 ±*19.2*
VNH	-9.9 ±*17.8*	-6.1 ±*16.9*	-11.0 ±*19.6*	-5.5 ±*17.7*	-15.0 ±*18.4*	-7.0 ±*19.6*	-14.8 ±*19.3*	-3.5 ±*18.5*
FAF	4.9 ±*17.4*	-0.2 ±*18.1*	5.3 ±*18.6*	-1.2 ±*17.5*	9.9 ±*16.1*	1.4 ±*17.9*	9.6 ±*17.6*	2.5 ±*17.1*
PLO	-1.5 ±*17.3*	-3.0 ±*16.1*	-0.3 ±*17.4*	-6.1 ±*20.8*	-0.5 ±*16.8*	-5.8 ±*17.4*	-0.3 ±*17.7*	-8.1 ±*19.0*
NAS	14.9 ±*22.4*	9.6 ±*21.7*	12.5 ±*20.1*	7.3 ±*22.1*	14.0 ±*17.7*	6.9 ±*19.6*	13.4 ±*18.9*	4.7 ±*20.0*
APP	9.7 ±*20.5*	14.8 ±*17.4*	10.1 ±*20.3*	15.6 ±*17.7*	12.7 ±*19.0*	22.0 ±*19.3*	14.8 ±*21.0*	19.2 ±*16.4*

TABLE III
AVERAGE VALUES AND STANDARD DEVIATIONS OF PARTICULAR CRITERIA FOR PHONE GROUPS

ACKNOWLEDGEMENTS

Research described in the paper was supported by internal CTU grant SGS12/143/OHK3/2T/13 “Algorithms and Hardware Realizations of Digital Signal Processing”.

We would like to appreciate also the work of Zdenek Patc who has created reference manual segmentation of testing data.

REFERENCES

[1] P. Pollak, J. Volin, and R. Skarnitzl, *HMM-Based Phonetic Segmentation in Praat Environment.* In The XII International Conference Speech and Computer - SPECOM 2007. Moscow, 2007, p. 537-541.

[2] J. Volin, R. Skarnitzl and P. Pollak, *Confronting HMM-based Phone Labelling with Human Evaluation of Speech Production.* In Interspeech Lisboa 2005 [CD-ROM]. Grenoble, 2005, vol. 1, p. 1541-1544. ISSN 1018-4074.

[3] J. Matousek, D. Tihelka, and J. Psutka, *Automatic segmentation for Czech concatenative speech synthesis using statistical approach with boundary-specific correction.* In Proc. Eurospeech,Geneva, Switzerland, 2003, p. 30-304.

[4] J. Matousek, *Automatic Pitch-Synchronous Phonetic Segmentation with Context-Independent HMMs.* In Text, Speech and Dialogue, proceedings of the 12th International Conference TSD 2009, Lecture Notes in Artificial Intelligence, p. 178-185, Springer, Brno, 2009.

[5] M. Novotny, J. Rusz, R. mejla, *Automatic segmentation of phonemes during the fast repetition of (/PA -/TA//KA/) syllables in a speech affected by hypokinetic dysarthria.* In Lekar a technika. 2012, vol. 42, no. 2, p. 81-84.

[6] Z. Xie, P. Niyogi, *Robust Acoustic Based Syllable Detection.* In Proceedings of Interspeech 2006, Pittsburgh, Pennsylvania, 2006.

[7] R. Cmejla, J. Rusz, P. Bergl, and J. Vokral, *Bayesian changepoint detection for the automatic assessment of fluency and articulatory disorders.* In Speech Communication, 55(1): 178-189,2013.

[8] J. Adell, and A. Bonafonte, *Towards Phone Segmentation For Concatenative Speech Synthesis.* In Proceedings of the 5th ISCA Speech Synthesis workshop, p. 139 -144, 2004.

[9] K. Sjolander, *An HMM-based system for automatic segmentation and alignment of speech.* In Proceedings of Fonetik 2003, p. 93-96, 2003.

[10] L. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition.* Proceedings of the IEEE , vol.77, no.2, p.257,286, Feb 1989.

[11] HTK speech recognition toolkit. [Online]. Ver. 3.3. July 2005. Available at: http://htk.eng.cam.ac.uk/

[12] P. Pollak and J. Cernocky, *Czech SPEECON adult database*. Technical report, Nov. http://www.speechdat.org/speecon, 2003.

[13] P. Pollak, J. Volin, R. Skarnitzl, *Influence of HMM's Parameters on the Accuracy of Phone Segmentation - Evaluation Baseline*. In Proceedings of the 16th Conference Joined with the 15th Czech-German Workshop ”Speech Processing”. Prague, 2005, vol. 1, p. 302-309.

[14] P. Machac, R. Skarnitzl, *Fonetická segmentace hlásek*. Praha: Nakladatelstvi Epocha, 2009. (In Czech language, translated title: Phonetic segmentation of phones)

[15] P. Pollak, J. Volin, and R. Skarnitzl, *Phone Segmentation Tool with Integrated Pronunciation Lexicon and Czech Phonetically Labelled Reference Database.* In 6th International Conference on Language Resources and Evaluation. Marrakech (Morocco), 2008, vol. 1, p. 1-5.