Experiments with Automatic Segmentation for Czech Speech Synthesis[*]

Jindřich Matoušek, Daniel Tihelka, and Josef Psutka

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz, dtihelka@kky.zcu.cz, psutka@kky.zcu.cz

Abstract. This paper deals with the automatic segmentation for Czech Concatenative speech synthesis. Statistical approach to speech segmentation using hidden Markov models (HMMs) is applied in the baseline system [1]. Several experiments that concern various issues in the process of building the segmentation system, such as speech parameterization or HMM initialization problems, are described here. An objective comparison of various experimental automatic and manual segmentations is performed to find out the best settings of the segmentation system with respect to our single-female-speaker continuous speech corpus.

1 Introduction

Accurate segmentation of speech has become very important in the task of concatenative speech synthesis. Corpus-based methods has emerged very popular in the context of speech synthesis. These methods often utilize large speech corpora segmented into acoustic (usually phone-like) units. Traditional human segmentation of such corpora would result in a very tedious and time-consuming work. Moreover, it is also almost impossible to keep the segmentation consistent. So, the need for a reliable automatic segmentation method is obvious. Statistical approach using hidden Markov models (HMMs) adopted from automatic speech recognition tasks (ASR) [4] has become the most successful [3, 6] (see Figure 1). HMM-based segmentation system was implemented in our text-to-speech (TTS) system as well [1].

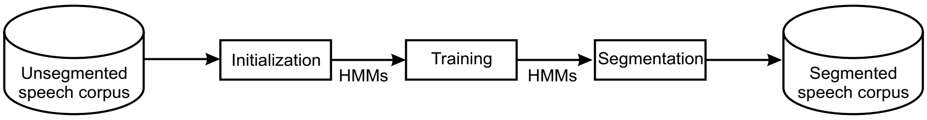

Fig. 1. A simplified scheme of HMM-based approach to speech segmentation.

[*] This research was supported by the Grant Agency of Czech Republic No. 102/02/P134 and the Ministry of Education of Czech Republic, project No. MSM235200004.

In this paper, several experiments with automatic segmentation of Czech speech are presented. The main attention is given to the enhancements of our baseline segmentation system [1] at the same time. Hopefully, an improved quality of the synthetic speech should make use of the efforts dedicated to the research of automatic segmentation techniques described in this paper.

2 Experimental Data

2.1 Training Data

When using HMM-based approach to speech segmentation, there is a need to "train" the segmentation system (i.e. to set up its parameters that are "optimal" for given speech data). Thus, such speech data are often called training data and in context of speech synthesis they are present in a speech corpus from which an acoustic unit inventory is to be created. Training data used in our experiments were primarily designed for our concatenative TTS system ARTIC [1] comprising about 13 hours of continuous speech spoken rather in a monotonous way by a single female speaker [2]. They were spoken in a sentence-by-sentence mode (5.000 sentences are available at the end). Each sentence is represented by its linguistic and acoustic forms. The linguistic information comprises orthography and phonetics. The acoustic form includes speech waveforms and parameterized speech vectors which describes principally the spectral properties of speech.

2.2 Reference Segmentation

To be able to evaluate the results of several automatic segmentation methods described in Section 3, a small portion of the speech data (50 sentences in total) was segmented by hand. The segmentation was performed by a single labeler knowledgeable in Czech acoustics and phonetics. However, this man was not an expert, so the manual segmentation was not supposed to be perceived as absolutely correct. Nevertheless, the reference manual segmentation was supposed to be accurate enough when comparing with the automatic segmentation.

To ensure the reference segmentation to be as correct as possible, the human labeler was asked to mark the boundaries between phones he was not sure about as "unsure" ones. Such suspicious boundaries were not used when evaluating the results of the automatic segmentation. In this way the reference segmentation data was kept as "clean" as possible. The most apparent problems when labeling Czech speech concerned liquids and glides especially in a vocalic context due to similar acoustic properties of both phones.

3 Experiments

Since the experiments with automatic speech segmentation described in this paper are primarily dedicated to the task of Czech concatenative speech synthesis, subjective listening tests should be used to assess the segmental quality of the synthetic speech. As there are no reasonable subjective listening tests available for Czech language in the time of writing this paper, more general objective tests were used instead to evaluate the accuracy of the automatic segmentation by comparing it to the manual segmentation. The following statistics were taken into account: absolute mean error between automatic and manual segmentation (|MD|) and standard deviation of this error (SD). The segmentation accuracy is also often expressed as a percentage of automatically detected boundaries which lie within a tolerance region around the human labeled boundary. The tolerance region used to be chosen somewhat arbitrarily. We chose smaller (10 ms) and bigger (20 ms) regions. All experiments were carried out utilizing the hidden Markov model toolkit (HTK) [5].

3.1 The Baseline System

Our baseline speech segmentation system uses the HMM-based approach to align phonetic labels to speech signals (see Figure 1). The very first version of our system was described in [1]. A set of three-state left-to-right single-density state-clustered crossword-triphone HMMs was employed to model context-dependent phone-sized units (triphones) on the basis of the speech corpus described in Section 2.1. The same speech corpus was then segmented using final triphone HMMs. So-called flat-start initialization (see Section 3.5 for details) was used to set up the parameters of HMMs. 12 Mel Frequency Cepstral Coefficients (MFCCs) plus normalized energy together with corresponding delta and acceleration coefficients (39 coefficients in total) were used in the baseline system (MFCC_EDA). MFCCs were computed using 25 ms window and 6 ms shift [1]. The results are shown in Table 1.

3.2 Speech Analysis

During speech analysis stage speech is usually represented by a sequence of feature vectors. Parameters of these vectors typically describe spectral properties of speech in a more compact way than raw speech samples do. Two factors contributes to the quality of speech parameterization: speech parameterization technique and the way the vectors are extracted from speech signal. Since HTK basically supports pitch-asynchronous parameter extraction, window length and positioning should be considered. 25 ms window length and 6 ms window shift (hereafter referred to as a 25/6 coding) were applied in the baseline system. A series of experiments with different MFCC coding schemes were carried out to find the best one for our speaker. As shown in Table 1 (experiment CODING) both 20/4 and 15/4 coding schemes turn up as the most suited for our female speaker. The Figure 2 shows the dependency of various window shifts on the segmentation accuracy.

In our next experiments speech parameterization techniques were also examined. Two most popular ASR speech analysis techniques: MFCC and PLP (perceptual linear prediction) coefficients were tested (see Table 1, experiment PARAM, for the results). Each experiment is described by the analysis technique

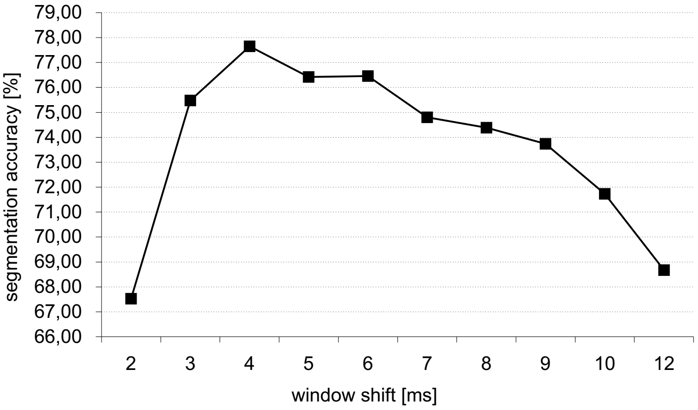

Fig. 2. The dependency of window shift on segmentation accuracy. A fixed window of length 25 ms and segmentation tolerance region of 10 ms were used.

(MFCC or PLP) and the number of static coefficients. E means (normalized) energy, D delta, A acceleration and T third differential coefficients. 0 stands for 0-th MFCC (a replacement of energy). The 20/4 coding scheme was applied. Our experiments revealed the superiority of MFCCs over PLPs. There were small differences for different MFCC coding schemes. Measuring the segmentation accuracy within the given tolerance region 10 ms, MFCC_12_EDAT showed to be the best coding scheme. When minimum |MD| is preferred, MFCC_12_EDA achieved the best results.

3.3 Context Dependency

In our next research the influence of the context dependency in HMM modeling on the segmentation accuracy was investigated. Four experimental sets of HMMs were taken into account: context-independent monophone HMMs (MONO), context-dependent state-clustered triphones (SCTRI), context-dependent not-clustered triphones (TRI), and monophone HMMs with context of phone groups (GRP) with acoustically similar phones in each group. The results in Table 1 (experiment CTX) show the superiority of triphone HMMs over monophone ones. “Pure” monophone HMMs performed most badly. On the other hand, the best results were achieved for not-clustered triphone HMMs. A very good performance was also observed for phone groups in HMM contexts. Unfortunately, these two methods cannot be applied in our TTS system directly, since the current version uses HMMs not only to segment speech but also to generate acoustic units that represents all sounds available in Czech language for our statistical approach to speech synthesis as well (by clustering similar HMMs as in SCTRI). Nevertheless, these techniques can be utilized when just the segmentation of speech is needed (e.g. when some other corpus-based techniques with unit selection are involved to synthesize speech).

3.4 Multiple Mixtures

As our baseline segmentation system uses single Gaussian HMMs only, another set of experiments was run to find out the influence of multiple mixture components on the segmentation accuracy. The motivation for this step was that ASR systems perform better when employing multiple mixtures (multiple mixture components were shown to get the best results in a Czech telephony ASR [4]). When incrementing the number of mixtures for our speaker-dependent triphone HMMs, the segmentation accuracy was decreasing (for 8 mixtures the accuracy was about 75% in tolerance region 10 ms). If many speaker-dependent data are available, single Gaussian density is assumed to be good enough to model output probability distribution when precise context modeling as with triphone HMMs is provided. Somewhat different behavior was observed for monophone HMMs where the best segmentation performance was obtained for 4-8 mixtures. The next mixture incrementing (up to 64 mixtures were tested) did not improve the segmentation accuracy (see Figure 3).

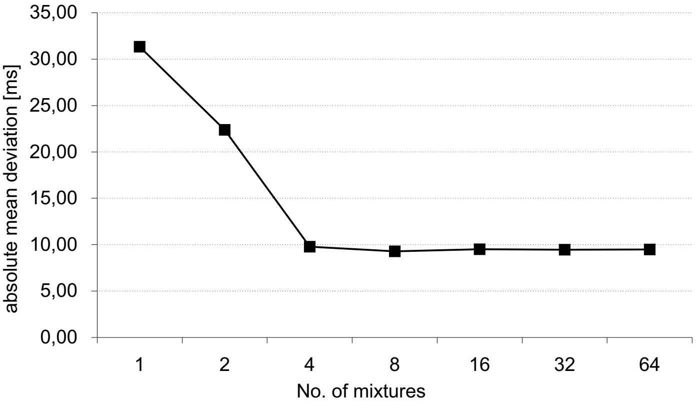

Fig. 3. Absolute mean deviation of automatic and manual segmentation for different number of mixtures in the monophone HMM segmentation system.

3.5 HMM Initialization

Since the principle of HMM-based approach consists of statistical refining the estimates of each HMM (starting from rough estimates and ending with more precise estimates in each estimation cycle), the initial estimates of HMM parameters play an important role. Good initial estimates can ensure that the local maximum is as close as possible to the global maximum of the likelihood function. Two strategies to initialize HMMs are extensively used. If no information about the boundaries between phones is available, flat-start initialization (FS) is usually performed to set up all HMMs with the same data. Such an initialization does not require any human intervention.

When some pre-segmented speech data are available, so-called bootstrap can be used to initialize each phone HMM individually. In this case, each HMM is initialized using the phone-specific data. In fact, there are two possibilities of obtaining some pre-segmented speech data. Ideally, a large amount of training sentences would be labeled by hand (preferably by an expert in acoustic phonetics). However, the manual segmentation is a very labor and time-consuming process. In our experiments 50 sentences were labeled by hand and used for hand-labeled HMM bootstrapping (HL, see Section 2.2 for details).

An alternative way to hand-labeling is to use speaker-independent (SI) ASR system to pre-segment training speech sentences. The advantage of this so-called SI HMM bootstrapping is that the labor process of manual segmentation is not needed any more. Moreover, all available training data can be used, resulting in more robust initial estimates of HMM parameters. An extended version of Czech SI continuous-speech ASR system [4] was employed for the bootstrapping (SI).

The influence of different HMM initialization strategies on our segmentation system with MFCC_EDAT and 20/4 coding is shown in Table 1 (experiment INIT). The best results were achieved for hand-labeled bootstrapping. Somewhat worse results were observed for SI HMM bootstrapping. Nevertheless, SI can be used as a reasonable compromise to segment Czech speech for concatenative speech synthesis when the tedious manual work is aimed to be eliminated or when no hand-labeled speech data are available.

3.6 Iterative Training

Iterative HMM training was proposed in [6] to get more accurate segmentation results. This approach is motivated by the assumption that more accurate initial estimates of the HMM parameters produce more accurate segmentation results. In this way a feedback is introduced because the segmentation from a previous iteration is used as the input for HMM initialization and re-estimation. Moreover, the influence of subjective hand-labeling for HMM bootstrapping is minimized since all training data can be used to initialize HMMs in the next iterations. More robust results are thus expected. The results for our Czech speech data after 10 iterations are shown in Table 1 (experiment ITER). The influences of each iteration on the results are illustrated in Figure 4. There were slightly worse results after 10 rounds of iterations when comparing to corresponding segmentation systems with no iteration.

4 Conclusion

In this paper we presented some experiments with automatic segmentation for the use in Czech concatenative speech synthesis. Different speech parameterizations were taken into account to find the best suited one for our female speaker speech corpus. The influence of context modeling and multiple mixture density on the segmentation accuracy were also considered. Several HMM initialization strategies were also taken into account. Iterative training was proposed as

Table 1. Summary of our experiments with segmentation of Czech speech: Baseline system was described in Section 3.1, CODING denotes experiments with various MFCC coding schemes (see Section 3.2), PARAM means experiments with different parameterization types (Section 3.2), CTX represents tests with context modeling (Section 3.3), INIT describes our research in HMM initialization strategies (Section 3.5), and ITER denotes a set of experiments with iterative HMM training (Section 3.6). |MD| express absolute mean error between automatic and manual segmentation, SD is standard deviation of the error. Segmentation accuracy was measured within the tolerance regions 10 ms and 20 ms.

Experiment	Descript.	|MD|	SD	Accuracy [%]	
				10ms	20ms
Baseline System		9.31	17.56	76.46	91.52
CODING	15/4	8.38	15.97	78.61	92.05
	15/6	8.65	16.97	78.49	92.08
	20/4	8.65	16.41	78.66	92.07
	20/6	9.05	17.40	77.32	91.81
	25/4	8.99	16.71	78.65	91.75
	25/6	9.31	17.56	76.46	91.52
	30/4	9.54	17.80	74.52	91.23
	30/6	9.69	18.15	74.50	91.26
PARAM	MFCC_11_EDA	8.72	17.19	78.81	92.08
	MFCC_12_EDA	8.65	16.41	78.66	92.07
	MFCC_12_0DA	8.73	16.30	77.85	91.39
	MFCC_12_EDAT	9.20	22.98	79.19	91.91
	MFCC_13_EDA	8.72	16.70	78.55	91.86
	PLP_11_EDA	10.09	23.27	75.74	90.10
	PLP_12_EDA	9.49	18.13	76.16	90.18
	PLP_12_EDAT	9.79	22.55	75.62	90.81
CTX	MONO	30.35	215.85	75.05	90.21
	GRP	9.04	20.32	79.26	91.22
	SCTRI	9.20	22.98	79.19	91.91
	TRI	9.21	23.31	79.68	91.69
INIT	FS	9.20	22.98	79.19	91.91
	SI	7.02	12.21	81.15	94.32
	HL	6.77	12.57	82.24	95.23
ITER	FS	9.20	22.86	79.01	91.94
	SI	7.22	12.21	81.30	94.35
	HL	6.94	13.79	81.80	94.95

well to get more stable initial HMM estimates. Roughly said, with regard to our Czech TTS system the best results of our speech segmentation experiments were achieved when single-Gaussian state-clustered triphone HMMs with hand-labeled bootstrapping were used to model speech parameterized with vectors of 12 MFCCs plus normalized energy and the first, second and third differential coefficients extracted from 20 ms long windows with 4 ms shift.

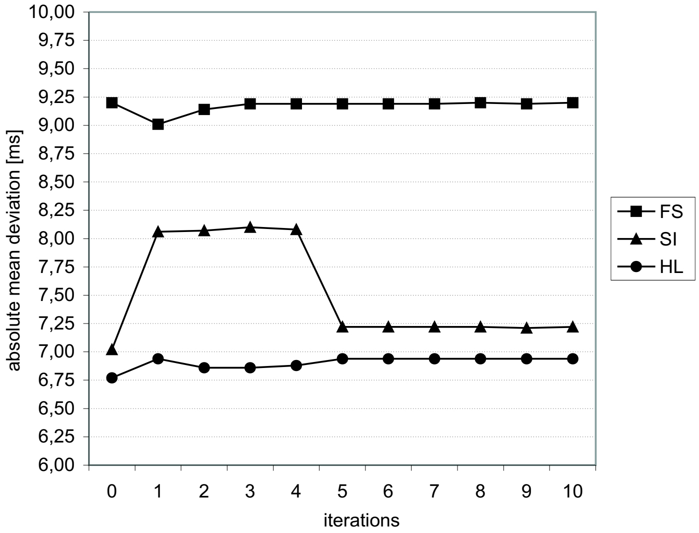

Fig. 4. The influence of iterative training on segmentation accuracy.

In our future work we will build acoustic inventories for our Czech TTS system with respect to our findings described in this paper. We are convinced that the improved segmentation methods should lead to a better quality of the synthetic speech. Nevertheless, listening tests will be also proposed to pick up the best segmentation method with respect to the quality of the synthetic speech.

References

1. Matoušek, J., Psutka, J.: ARTIC: a New Czech Text-to-Speech System Using Statistical Approach to Speech Segment Database Construction. Proceedings of ICSLP 2000, vol. IV. Beijing (2000) 612–615.
2. Matoušek, J., Psutka, J., Krůta, J.: On Building Speech Corpus for Concatenation-Based Speech Synthesis. Proceedings of Eurospeech2001, vol 3. Ålborg (2001) 2047–2050.
3. Ljolje, A., Hirschberg, J., van Santen J. P. H., Automatic Speech Segmentation for Concatenative Inventory Selection. Progress in Speech Synthesis. Springer (1996) 305–311.
4. Psutka, J., Müller, L., Psutka, J. V.: Comparison of MFCC and PLP Parameterization in the Speaker Independent Continuous Speech Recognition Task. Proceedings of Eurospeeech 2001. Ålborg, (2001) 1813–1816.
5. Young, S., et al.: The HTK Book (for HTK Version 3.2). Cambridge University Press. Cambridge, UK. (2002).
6. Kim, Y.-J., Conkie, A.: Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction. Proceedings of ICSLP 2002. Denver (2002) 145–148.