Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-Specific Correction

Jindřich Matoušek, Daniel Tihelka, Josef Psutka

Department of Cybernetics
University of West Bohemia in Pilsen, Czech Republic
jmatouse@kky.zcu.cz, dtihelka@kky.zcu.cz, psutka@kky.zcu.cz

Abstract

This paper deals with the problems of automatic segmentation for the purposes of Czech concatenative speech synthesis. Statistical approach to speech segmentation using hidden Markov models (HMMs) is applied in the baseline system. Several improvements of this system are then proposed to get more accurate segmentation results. These enhancements mainly concern the various strategies of HMM initialization (flat-start initialization, hand-labeled or speaker independent HMM bootstrapping). Since HTK, the hidden Markov model toolkit, was utilized in our work, a correction of the output boundary placements is proposed to reflect speech parameterization mechanism. An objective comparison of various automatic methods and manual segmentation is performed to find out the best method. The best results were obtained for boundary-specific statistical correction of the segmentation that resulted from bootstrapping with hand-labeled HMMs (96% segmentation accuracy in tolerance region 20 ms).

1. Introduction

In our previous work, we have designed ARTIC, a new Czech text-to-speech (TTS) system based on concatenation of phone-level speech segments [1, 2]. Generally, the synthetic speech quality of a concatenation-based synthesis system critically depends on the quality of an acoustic inventory. The key task here is the segmentation of the speech corpora the inventories are built from. This paper addresses the problems of the automatic segmentation of speech for the purposes of Czech TTS synthesis.

Traditionally, speech segmentation for concatenative synthesis of speech was performed by *human experts* especially in the field of acoustics or phonetics. Since the quality of resulting synthetic speech to a large extent depends on the accuracy of segmentation of speech into acoustic units, the expert *manual segmentation* was believed to be the only means to guarantee the most exact segmentation. However, the process of hand-labeling is an extremely labor and time-consuming activity. Moreover, if a large amount of speech is to be segmented by hand, keeping the boundary placements consistent may be very difficult, especially when two or more labelers are involved [3]. Such a manual segmentation is reasonable to be performed on a small portion of data only. In the last decade, the accession of corpus-based techniques brings about the need of the segmentation of large speech corpora (usually up to several tens of hours of speech). It is almost impossible to ensure a consistent manual segmentation of so many speech data. It is evident that there is a need of a reliable automatic speech segmentation technique.

In fact, there are two automatic methods extensively used for the segmentation of speech. The first one uses a *dynamic time warping* (DTW) technique to align a to-be-segmented speech with the corresponding segmented speech generated by a speech synthesis system [4]. The second method uses a *statistical approach* with *hidden Markov models* (HMMs), a dominant technique successfully applied in the automatic speech recognition (ASR) systems. In the last decade, statistical approach (see Figure 1 for an illustration) became very popular also for the automatic segmentation of speech in the context of concatenative speech synthesis, especially thanks to very fast and consistent segmentation of large speech corpora (the segmentation tends to be the same in similar speech contexts). The statistical approach is reported to outperform DTW (e.g. [5]). It can be viewed as a speaker dependent speech recognition system running in the so called *forced alignment mode* (in fact no recognition is performed at all, because the phonetic transcription of a sentence is a priori known – the speech recognition system is used just to find the best alignment of HMMs and feature vectors representing the speech data of the sentence) [3]. Additionally, some postprocessing either automatic (e.g. spectral correction [6] or explicit boundary modeling [5, 7]) or manual verification can be applied on the resulting segmentation to improve its accuracy. There are also some drawbacks of this approach. The segmentation accuracy is limited by the principle of HMM-based approach itself: HMMs are built to *identify* phonetic segments, not to *obtain precise phonetic boundaries* [7]. So, the requirement of the most accurate boundary placements is not part of the optimization criteria. There is also a limited resolution for boundary detection given by the speech parameterization mechanism. The solution of these shortcomings is not primarily discussed in this paper, although the boundary-specific statistical correction method proposed in Section 3.3 can correct some of the shortcomings.

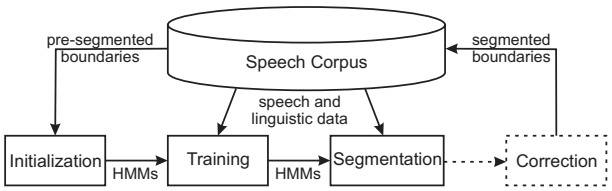

Figure 1: *A simplified illustration of the statistical approach to speech segmentation with optional correction.*

In this paper some improvements on the underlying statistical approach implemented in our baseline Czech speech segmentation system are proposed. The influences of several HMM initialization strategies on the segmentation accuracy are stud-

ied. A boundary-specific statistical correction of the automatic segmentation based on hand-labeled bootstrapping HMM initialization is implemented to minimize the segmentation errors.

The paper is organized as follows. In Section 2 our baseline system for the segmentation of Czech speech is introduced. Section 3 describes several proposals how to improve the speech segmentation accuracy. In Section 4 we present the results of the various automatic speech segmentation methods. Finally, Section 5 contains the conclusion and outlines our future work in this field.

2. The Baseline System

Our baseline speech segmentation system (FS0) uses the statistical approach to align phonetic labels to speech signals (see Figure 1). The hidden Markov model toolkit (HTK) was utilized to perform the segmentation [8]. The very first version of our system was described in [1]. A set of three-state left-to-right single-density crossword-triphone HMMs was employed to model context-dependent phone-sized units (triphones) on the basis of a large single-female-speaker continuous speech corpus. The same speech corpus was then segmented using final triphone HMMs. So-called *flat-start* initialization (see Section 3.2 for details) was used to set up the parameters of HMMs. The corpus was designed very carefully to contain phonetically balanced sentences [2]. Nowadays, there are 5.000 sentences (about 13 hours of speech) in the updated corpus. From speech segmentation point of view, the corpus comprises both linguistic and signal representations of speech. As for linguistics, both orthographic and phonetic transcriptions of each sentence are used. Speech signals are represented by their waveforms and their spectral properties are described by vectors of Mel Frequency Cepstral Coefficients (MFCCs) calculated using 20 ms windowed speech signal with 4 ms shift. In the current system 12 MFCCs plus normalized energy together with corresponding first, second and third differential coefficients (52 coefficients in total) are used.

3. Experiments

The speech segmentation accuracy using our baseline system (FS0) is shown in Section 4. The results are not good enough (about 60% in tolerance region 10 ms). Surprisingly, the quality of synthetic speech produced by our TTS system that uses the speech unit database built from this not very accurate segmentation is supposed to be very good [1, 2]. It is assumed the relatively good quality is obtained thanks to the consistency in HMM-based segmentation (in other words: the segmentation system always makes the same mistakes, so they tend to cancel during concatenative speech synthesis). Nevertheless, it is believed that the approaching of the automatic segmentation to the manual segmentation should lead to the better quality of synthetic speech imitating the quality of expert segmentation while maintaining the HMM-based consistency.

To improve the speech segmentation abilities of our system, a series of experiments were carried out. The enhancements on the baseline system are discussed in the next subsections.

3.1. Shifting the HTK boundaries

After analyzing the results of the baseline segmentation system we found a nearly constant forward shift between each automatically and manually segmented phone boundary. When studying the manner HTK works with speech data, we learned that

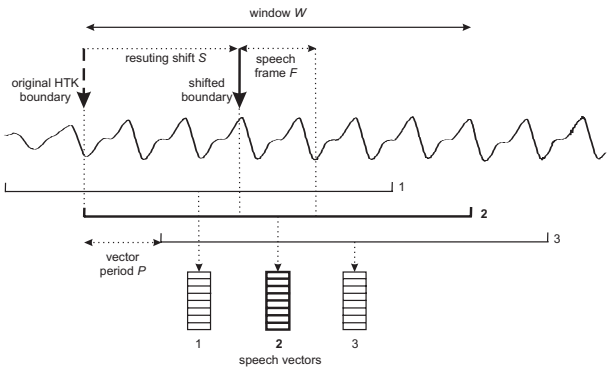

Figure 2: *Shift between automatic and manual segmentation. The shift is illustrated for speech vector No. 2.*

the shift had been imposed by HTK speech parameterization tool. As shown in Figure 2 each parameterized speech vector is associated with the start of an analysis window W from which it is computed (the length of the window $L_W = 20$ ms in our case). But in fact, the vector describes the signal around the middle of the window most exactly, so there is a shift between speech vector and waveform times. The shift between successive windows determines the period P between each parameter vector ($P = 4$ ms in case of our parameterization). Recalling the manner HMMs model the speech production, speech vectors are supposed to describe the properties of speech frame F of length $L_F = P$ in the best way. Therefore the resulting forward shift S is given by centering the speech frame F described by its speech vector around the middle of the analysis window W ($S = 8$ ms in our case):

$$S = \frac{L_W - L_F}{2}. \quad (1)$$

The accuracy of the segmentation system with shifted boundaries (FS1) is shown in Table 1. There was nearly 20% improvement when comparing with the baseline system, so the shift was implemented in other experiments described further as well.

3.2. Initialization of HMMs

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 is usually performed to set up all HMMs with the same data. Such an initialization does not require any human intervention and thus was used in our systems described above (FS0 and FS1).

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 (HLB, see Section 4.1 for details).

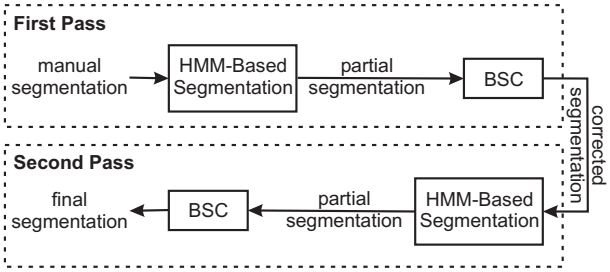

Figure 3: *Automatic two-pass speech segmentation using boundary-specific statistical correction (BSC).*

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 [9] was employed for the bootstrapping (SIB).

Again, when using both kinds of bootstrapping in HTK system the same shift as described in Section 3.1 (but in the reverse direction this time) should be used to ensure HMMs are initialized with the right data.

3.3. Boundary-Specific Statistical Correction

When the reference manual segmentation is available (as in HLB), it can be used to correct the final segmentation. After HLB segmentation had been performed, the errors of the automatic segmentation were statistically evaluated. In the current version simply the average deviation $\bar{d}_i$ of the automatic segmentation comparing to manual segmentation was computed for *each type of boundary* b_i (*boundary-specific statistical correction* – BSC). Then, more precise estimate $\hat{b}_i$ of each individual boundary b_i was computed by shifting the boundary b_i with respect to the boundary-specific average deviation $\bar{d}_i$:

$$\hat{b}_i = b_i + \bar{d}_i. \quad (2)$$

To minimize the influence of subjective hand-labeling on the results, *two-pass* segmentation scheme was implemented (see Figure 3). After the first pass all boundaries in all training data were shifted as shown in equation 2. Such corrected segmentation was used as the input for the second pass of automatic segmentation. The statistical refinement is justifiable because the aim of the segmentation for concatenative speech synthesis is to get the most accurate boundaries for *given training data*. To make more robust estimates for the statistical correction, Czech phones were divided into 10 groups: short (VOS) and long vocals (VOL), diphthongs (DIP), voiced (STV) and unvoiced stops (STU), voiced (FRV) and unvoiced fricatives (FRU), nasals (NAS), affricates (AFR) and liquids+glides (L+G). The groups reflect phonetic and acoustic properties of phones (especially the manner of articulation and voiced/unvoiced characteristics). Each boundary was then specifically described by its left and right context represented by a corresponding group of phones.

4. Results

The evaluation of the segmentation accuracy can be generally performed in two ways: objective and subjective. The subjective assessment usually depends on what the segmentation is used for. In our case the time-aligned labeling serves for the purposes of building a speech unit database in tasks of Czech text-to-speech synthesis. Therefore, the subjective evaluation should concern the segmental quality of the synthetic speech. As far as we know there are no reasonable subjective listening tests available for Czech language in the time of writing this paper. So, more general objective tests were used instead to evaluate the accuracy of the automatic segmentation by comparing it to the manual segmentation.

4.1. Manual 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 without any a priori information (unlike e.g. [10] where the reference segmentation was obtained by a manual correction of an existing automatic segmentation). 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 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.

4.2. Objective Comparison of Automatic and Manual Segmentation

To evaluate the segmentation methods described in Section 3, statistics of the deviation between the automatic and manual segmentation were computed (see Table 1). 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 (see Figure 4 for results).

Table 1: *Statistics of the comparison of automatic and manual segmentation. Mean deviation (MD), standard deviation (SD), absolute mean deviation (|MD|) and absolute maximum deviation (|MaxD|) are introduced in ms.*

Method	MD	SD	|MD|	|MaxD|
FS0	5.98	22.98	12.04	526.72
FS1	-2.02	22.98	9.20	518.72
SIB	-0.35	12.21	7.02	299.20
HLB	0.30	12.57	6.77	315.37
BSC	0.00	11.12	5.78	318.43

The results show the superiority of hand-labeled bootstrapping methods. Indeed, the more accurate HMM initialization, the more accurate segmentation results were obtained. Absolutely the best performance (96% in tolerance region 20 ms) was achieved when the boundary-specific statistical correction was applied, resulting in a roughly half average absolute devi-

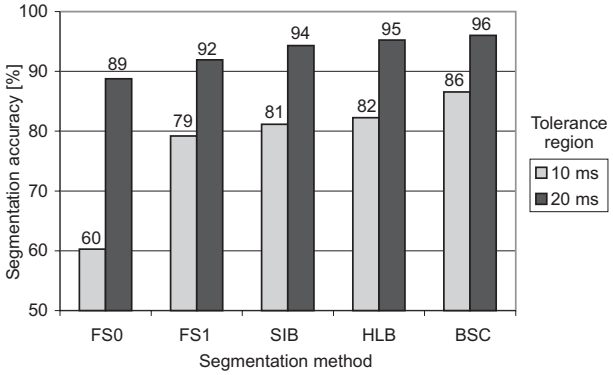

Figure 4: *Speech segmentation accuracy for various automatic segmentation methods described in Section 3. The tolerance region was taken 10 ms (left) and 20 ms (right).*

ation when compared to the baseline system. Somewhat worse results were observed for SI HMM bootstrapping (SIB). Nevertheless, SIB 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.

Table 2 shows the segmentation statistics of BSC method for different types of phones. To make the statistics of the comparisons of automatic and manual segmentation results more readable, similar groups of phones as defined in Section 3.3 were used. Similarly as for human labeling the worst results were obtained for liquids and glides. Sometimes it was also difficult to detect some other phones, specifically fricatives (especially the start of voiced fricatives) and the start of affricates. Some problems also occurred for the ends of long vocals.

Table 2: *Absolute average label start and end BSC segmentation errors |MD| in ms and segmentation accuracy (Acc) for tolerance region 10 ms.*

Phone group	Start		End	
	|MD|	Acc [%]	|MD|	Acc [%]
VOS	5.89	87.32	5.28	88.51
VOL	5.62	87.90	6.45	82.69
DIP	4.93	92.86	5.12	85.11
NAS	4.59	91.98	5.53	86.69
L+G	8.36	76.11	9.06	68.16
STV	5.66	86.78	4.36	92.55
STU	6.31	87.47	4.97	91.37
FRV	8.81	75.95	5.57	86.42
FRU	6.80	85.02	7.87	82.59
AFR	7.22	82.83	2.99	95.52

5. Conclusion

In this paper we proposed some enhancements on the automatic segmentation of speech in the context of Czech speech synthesis by concatenation. Since HTK system was used to implement HMM-based approach, a speech parameterization-dependent shift was applied to the output segmentation to reflect speech analysis implemented in HTK. Several HMM initialization strategies were also taken into account. Hand-labeled HMM bootstrapping (HLB) was evaluated as the best initialization method. Boundary-specific statistical correction was then applied together with HLB in a two-pass segmentation system and the best segmentation results were achieved (96% segmentation accuracy in tolerance region 20 ms). The segmentation accuracy was improved up to 26% (in tolerance region 10 ms) when compared to the baseline system.

In our future work we will build speech unit databases for our Czech TTS system using the enhanced segmentation methods 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 confirm our hypotheses and to evaluate the segmentation methods with respect to the quality of the synthetic speech. Other experiments (e.g. some spectral corrections) are also planned to get even more accurate segmentation results (with emphasis on the problematic phones mentioned above – liquids and glides).

6. Acknowledgment

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.

7. References

[1] Matoušek, J., Psutka, J., “ARTIC: A New Czech Text-to-Speech Synthesis System Using Statistical Approach to Speech Segment Database Construction”, Proceedings of ICSLP2000, vol. IV, Beijing, 2000, pp. 612–615.

[2] Matoušek, J., Psutka, J., Krůta, J., “Design of Speech Corpus for Text-to-Speech Synthesis”, Proceedings of Eurospeech2001, vol. 3, Ålborg, 2001, pp. 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, pp. 305–311.

[4] Black, A. W., Lenzo K. A., “Building Synthetic Voices”, <http://festvox.org>.

[5] Sethy, A., Narayanan, S., “Refined Speech Segmentation for Concatenative Speech Synthesis”, Proceedings of ICSLP 2002, Denver, 2002, pp. 149–152.

[6] Kim, Y.-J., Conkie, A., “Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction”, Proceedings of ICSLP 2002, Denver, 2002, pp. 145–148.

[7] Torre Toledano, D., Rodríguez Crespo, M. A., Escalada Sardina, J. G., “Trying to mimic Human Segmentation of Speech Using HMM and Fuzzy Logic Post-Correction Rules”, Proceedings of ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, Australia, 1998, pp. 207–212.

[8] Young, S., et al., “The HTK Book (for HTK Version 3.2)”, Cambridge University Press, Cambridge, UK, 2002.

[9] 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, pp. 1813–1816.

[10] Makashay, M. J., Wightman, C. W., Syrdal, A. K, Conkie, A., “Perceptual Evaluation of Automatic Segmentation in Text-to-Speech Synthesis”, Proceedings of ICSLP 2000, Beijing, 2000, pp. 431–434.