Automatic Speech Segmentation with Multiple Statistical Models

Seung Seop Park, Jong Won Shin and Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University, Korea
{sspark, jwshin}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

In this paper, we propose a novel approach to improve the performance of automatic speech segmentation techniques for concatenative text-to-speech synthesis. A number of automatic segmentation machines (ASMs) are simultaneously applied and the final boundary time marks are drawn from the multiple segmentation results. To identify the best time mark among those provided by the multiple ASMs, we apply a candidate selector trained over a set of manually-segmented speech database. The candidate selector defines a mapping from the phonetic boundary to the best ASM index which will output the time mark that may be closest to the manual segmentation result. The experimental results show that our approach dramatically improves the segmentation accuracy.
Index Terms: speech synthesis, unit selection, speech segmentation.

1. Introduction

Nowadays, the unit selection technique [1] has become the most widely used approach in the area of text-to-speech (TTS) synthesis to realize high-quality synthetic speech. In this technique, synthetic speech is generated by concatenating a series of units (i.e. waveform segments) which are selected from a large speech database. For that reason, the quality of synthetic speech critically depends on the quality of the selected units.

The key task for building the database is to mark the boundaries of each speech segment according to the given transcript. Although manual labeling is generally regarded as the most reliable way to get the boundary time marks, it is usually too time-consuming and labor-intensive. Therefore, an automatic method for the segmentation task is considered to be more desirable and practical, especially when a huge amount of speech data are to be segmented.

In the literature, a variety of approaches to automatic speech segmentation have been developed [2]-[8]. Most of the developed approaches are based on the Hidden Markov Model (HMM) which is widely used in the area of automatic speech recognition although some other techniques such as the dynamic time warping (DTW) method [2] are also applied. In the HMM-based framework, the model parameters are trained based on the given speech data with the corresponding transcripts and then the trained HMMs are used to align the training data along the associated transcripts.

It is generally known that various model configurations such as the number of states for each phone, number of mixture components for each state, context-dependency and the feature vectors extracted from the speech waveform produce different segmentation results [3]. Context-dependent HMMs are also known to make some systematic errors (or bias) since they are always trained in the same phonetic context [4]. To compensate the error, some statistical techniques such as the boundary specific correction (BSC) [5] and statistical correction of context dependent boundary marks (SCCDBM) [4] are applied to correct the segment boundaries which are provided by HMM-based alignment. Even after the boundary correction, however, the performance of an automatic HMM-based segmentation technique is usually found insufficient to be directly applied to TTS. In order to alleviate this difficulty, additional post-processing approaches are usually applied to refine the segment boundaries with the use of different features e.g., F0 contour [6] and spectral variation function (SVF) [7]. On the other hand, in [8], multiple independent acoustic models are adopted and the segmentation results are averaged to yield a final result.

In this paper, we propose a novel approach to estimate reliable segmentation boundaries of the speech data when a limited amount of manually-segmented data is available. We apply multiple separate algorithms to obtain a number of segmentation results, and a final decision for each segment boundary is made by combining the multiple segmentation results depending on the robustness of each model against the specific boundary type.

2. Automatic Segmentation by Boundary-Type Candidate Selection (ASBTCS)

2.1. Overview

Let us define an automatic segmentation machine (ASM) to be a system that produces a sequence of boundary time marks $\mathbf{t}^u = \{t_1^u, \ldots, t_{n_u}^u\}$ given an utterance u and its corresponding phonetic labels $\mathbf{p}^u = \{p_0^u, p_1^u, \ldots, p_{n_u}^u\}$ where n_u is the number of phonetic boundaries of u and t_i^u represents the time mark for the phonetic boundary between p_{i-1}^u and p_i^u. An ASM applies an algorithm to align an utterance along its phonetic labels. For that purpose, it may adopt an HMM-based, DTW-based approaches, or the post-processing techniques for boundary refinement. In this paper, our interest lies on how to determine the boundary marks when the segmentation results from various ASMs are given instead of focusing on each specific ASM algorithm.

To gain an insight as to how the boundaries are determined by an ASM, let us consider the case of the HMM-based technique. The HMM-based ASM depends on a collection of (context-independent or -dependent) phone models which have been constructed through the training procedure. When a speech signal and the corresponding phonetic transcript are given, a sequence of feature vectors such as the mel-frequency cepstral coefficients (MFCCs) are extracted for each frame and then each feature vec-

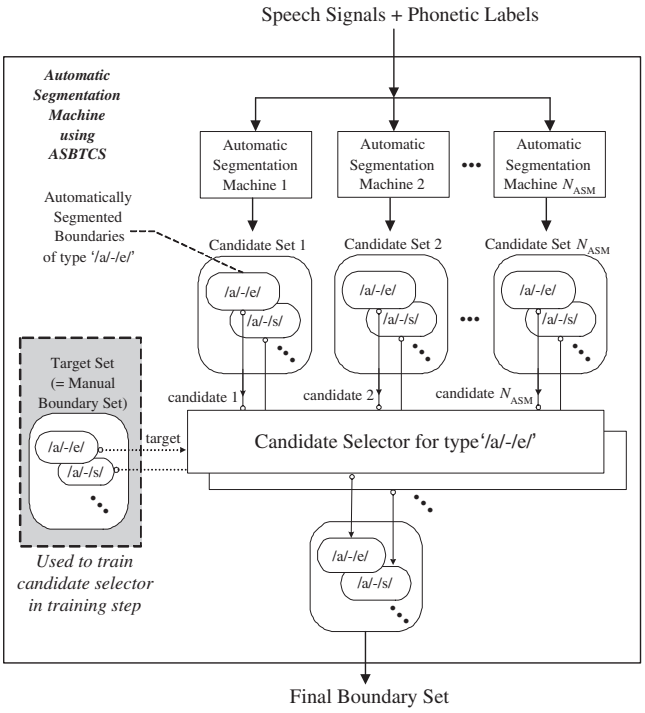

Figure 1: *Overview of the proposed ASBTCS method.*

tor is aligned along the corresponding HMM state by applying the Viterbi algorithm. Strictly speaking, all the HMMs that are used to segment the given utterance contribute to the determination of all the phone boundaries. However, it is reasonable to assume that a phone boundary is dominantly affected by the two adjacent (left and right) phone models. If we define the boundary type, b_i^u, for the time mark t_i^u as the two phonetic identities adjacent to this time mark, i.e. $b_i^u = (p_{i-1}^u, p_i^u)$, the estimation error for t_i^u, based on the above assumption, will vary according to the trained models for both p_{i-1}^u and p_i^u. Therefore, an ASM will show a different performance for each boundary type.

Now, suppose that there are a number of ASMs which use a variety of algorithms of their own. In this case, our goal is to make a final segmentation result by utilizing all the available ASMs. One of the promising ways would be to apply a separate ASM for each boundary type depending on the phonetic characteristics. In this paper, to implement this idea, we propose a novel approach called the *Automatic Segmentation by Boundary-Type Candidate Selection* (ASBTCS). The overall idea of the proposed ASBTCS method is shown in Fig. 1. Firstly, multiple boundary sets denoting the collection of boundary time marks are produced by multiple ASMs which adopt different methods from each other. Then, for each boundary type the candidate selector chooses the best time mark among the boundaries provided by the multiple ASMs. In this respect, the candidate selector defines an one-to-one mapping between the boundary type and the ASM index such that it can identify the ASM which results in the minimum segmentation error for the given boundary type. The candidate selector is constructed using a training procedure where the manually-segmented data are considered the target values for the boundary time marks. For each boundary type observed in the training database, the average time differences between the target time marks and those provided by the ASMs are computed and the ASM with the minimal error is selected as the winner for the given boundary type.

2.2. Training of Candidate Selector

A set of manually-marked boundaries are required for the training of candidate selector in the ASBTCS method. Since it is usually cumbersome to conduct manual segmentation over the entire speech database, the candidate selector is trained based on a limited portion of the whole utterances. Let us denote the manual segmentation result as $\mathrm{M} = \{t_i^{\mathrm{M}}\}_{i=1,\ldots,N_{\mathrm{M}}}$, where N_{M} is the total number of manually-segmented boundaries. If the number of ASMs considered in the candidate selector is N_{ASM}, we have N_{ASM} candidate boundary sets $\{\mathrm{C}_{(1)}, \ldots, \mathrm{C}_{(N_{\mathrm{ASM}})}\}$ where $\mathrm{C}_{(k)} = \{t_i^{\mathrm{C}_{(k)}}\}_{i=1,\ldots,N_{\mathrm{M}}}$ represents the time marks produced by the k-th ASM. We should note that the boundary type (p_{i-1}, p_i) for t_i^{M} is the same to that for each candidate time mark $t_i^{\mathrm{C}_{(k)}}$.

For each boundary type, the candidate selector outputs the index of the ASM which produces the minimum error with respect to the manual segmentation results. Let θ and S_θ denote a boundary type and the corresponding output of the candidate selector, respectively. Then,

$$S_\theta = \underset{k \in \{1,2,\ldots,N_{\mathrm{ASM}}\}}{\operatorname{argmin}} \xi(\mathrm{C}_{(k)}\langle\theta\rangle, \mathrm{M}\langle\theta\rangle) \qquad (1)$$

where $\mathrm{M}\langle\theta\rangle$ is a subset of M to which only the time marks associated with the boundary type θ belong and $\mathrm{C}_{(k)}\langle\theta\rangle$ is defined in the same way. In (1), ξ represents a cost function specified in terms of the two boundary sets. Since the cost function ξ should account for some distance metric, it can be defined, for example, as the absolute errors between the two boundary sets given as follows:

$$\xi(\mathrm{C}_{(k)}\langle\theta\rangle, \mathrm{M}\langle\theta\rangle) = \sum_{i:\, t_i^{\mathrm{M}} \in \mathrm{M}\langle\theta\rangle} \left| t_i^{\mathrm{C}_{(k)}} - t_i^{\mathrm{M}} \right|. \qquad (2)$$

2.3. Clustering of Boundary Types

It is important for robust candidate selection that there are sufficient amount of manually-segmented data for each boundary type. Since, however, the size of the manually-segmented data is usually small, training of the candidate selector does not guarantee a robust estimate for all the boundary types. Furthermore, some boundary types are not even present in the manually-segmented data. In order to complete our segmentation technique, we should also have candidate selection rules for those unseen boundary types.

To alleviate this difficulty, we apply a decision tree [9] to cluster the boundary types. The decision tree is built as follows: First, all boundaries of the manually-segmented data are pooled together in the root node of the tree. Then, this pool is subsequently split up into two child nodes according to phonetically-motivated questions, such as the place of articulation, the voicing of phone, and the preceding and following phonetic context of the boundary. The split at each node is made such that the sum of two child nodes' absolute errors could be minimized when the ASM with minimum mean absolute error is selected at each node. The stopping criterion is to ensure at least δ data points in each leaf node. After the decision tree has been built, the candidate selector is trained such that it can define an one-to-one mapping between each leaf node of the tree and the ASM index.

3. Experimental Results

In order to evaluate the performance of the proposed ASBTCS method, we applied 36 candidate ASMs, all were built based on the HMM approach. The speech database used in our experiment consisted of 5000 Korean utterances (286082 phones) which were spoken by a professional female narrator in a studio environment and were recorded in 16-bit precision with 16 kHz sampling frequency. In the speech database, manual segmentation results were available for 2000 utterances among which a maximum of 1600 utterances were used for the training of candidate selector and the remaining 400 utterances were reserved for performance evaluation.

To train the HMM-based ASMs, a feature vector was extracted for each frame with 24 ms window length and 3 ms frame shift. The feature vector was composed of 12 MFCCs, normalized log energy, and their first and second order delta components (39-dimension in total). The basic structure of the phone HMMs was a left-to-right type without any state skipping. In addition, the observation distribution specified in each state was characterized by the Gaussian mixture model with a finite number of mixture components. The 36 ASMs were established by varying the number of states for each phone HMM, the number of mixture components per each state, and the manner of incorporating context dependency. The number of states for each phone model was allowed to vary from three to five and 1∼6 Gaussians were used to represent the observation distribution of each state. Both the context-independent monophone and context-dependent triphone models were trained for each HMM structure configuration resulting in total 36 ASMs. All the candidate ASMs were trained over the 4600 utterances (excluding the evaluation data) without any manual segmentation information, and no post-processing techniques for boundary refinement were employed. Training of the HMMs was carried out with the use of HTK [10] software where state tying was applied to estimate the parameters of the triphone models.

Each of these ASMs was applied to segment the utterances, and the obtained results were taken as candidate sets for the ASBTCS method. For the training of candidate selector, 400 manually-segmented utterances were used. There were 949 boundary types observed in the training database, while 1218 boundary types existed in the entire database. To cope with the unseen boundary types and to select the candidates in a robust way, a decision tree was built based on the training database such that there should be at least δ data points for each leaf node of the tree, yielding 642 leaf nodes for $\delta = 5$. Finally, the candidate selector was trained to define an one-to-one mapping from each leaf node to the ASM index.

In order to investigate the effectiveness of applying multiple ASMs, we evaluated the segmentation performance of the proposed method by varying the number of candidate boundary sets included. The performance was evaluated by measuring the mean of absolute time differences between the manual time marks and those obtained from the automatic segmentation techniques. The curves in Fig. 2 demonstrate the performance improvement of the ASBTCS method as more ASMs contributed to the boundary determination. For the purpose of comparison, we also plot the performance of each single ASM which was newly incorporated in the ASBTCS approach. From the result, it is noted that the performance of the ASBTCS approach improved even though the newly incorporated single ASM performed worse than the other ASMs. This phenomenon somewhat confirms that it is advantageous to apply different ASMs depending on the given boundary type.

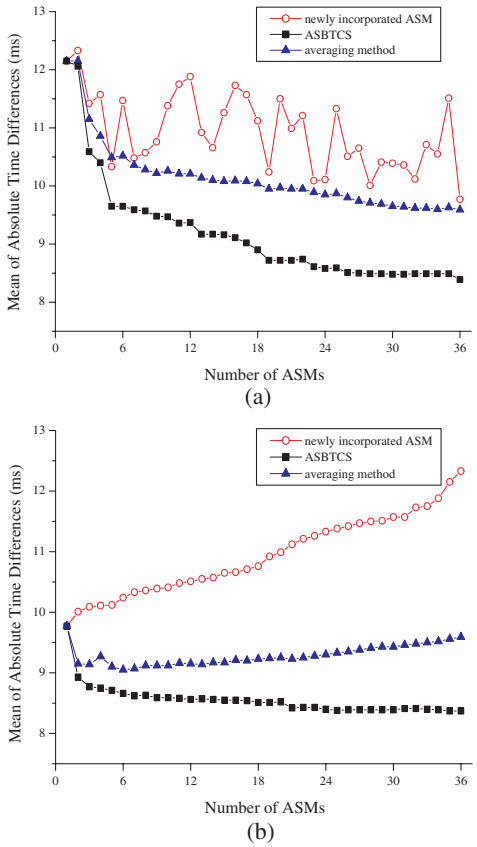

Figure 2: *Performance of the ASBTCS method (400 training data for candidate selector, δ=5) as the number of ASMs incorporated is varying. The ASMs are added (a) in random order (b) in the order of decreasing performance.*

To compare the proposed method with other previous multiple-model-based approaches, we also evaluated the performance of the algorithm presented in [8] where the final segmentation result is obtained through simple averaging and the result is shown in Fig. 2, where we can see that the ASBTCS method outperformed the simple averaging scheme. In Fig. 2(a), both the ASBTCS and the averaging method showed a decreasing trend of error as more ASMs participated in each method. On the other hand, Fig. 2(b) shows that when ASMs were added in the order of decreasing performance the performance of the averaging method became degraded while that of the proposed method did not. Therefore in the ASBTCS method, adding more ASMs seems to be rather safe and is expected to have some positive effect on the final performance even in case each ASM's performance is not so good.

For a more quantitative analysis, we compared the performances of the ASBTCS approach, the simple averaging method, and the single ASM which had achieved the best overall performance among the 36 ASMs. The best ASM chosen was the one with context-independent, 5-state and 1-mixture HMM. This time, we counted the relative number of time marks which lied within

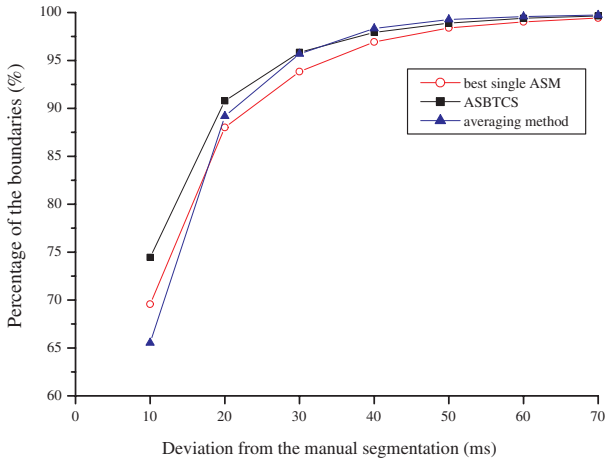

Figure 3: *Percentage of the boundaries within some tolerances with respect to the manual time marks.*

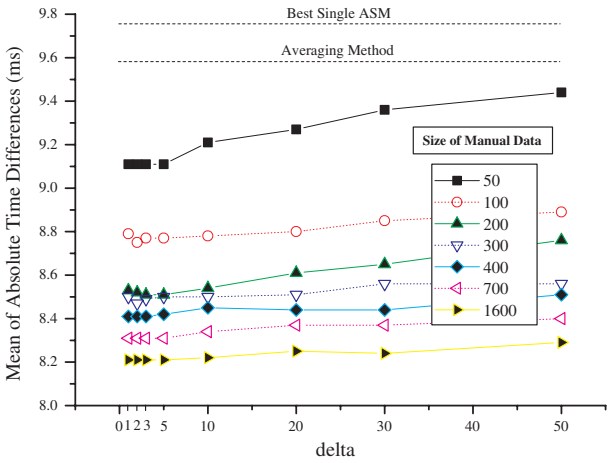

Figure 4: *Performances of the ASBTCS method as the number of training data and δ are varying.*

a specified distance from the corresponding manual boundaries. The result is given in Fig. 3 where we can see that the ASBTCS approach made more boundary time marks placed closer to those of the manually-segmented data than the best single ASM for all tolerances and than the averaging scheme in the regions below 40 ms.

Fig. 4 shows how the segmentation performance of the ASBTCS approach was affected by varying δ, which specifies the minimum number of data kept in each leaf node of the decision tree. In this experiment, we also varied the amount of training utterances from 50 to 1600. For the purpose of comparison, we also display the results of the simple averaging technique and the best single ASM. In the figure, the performance of the proposed method using only 50 utterances was better than the averaging method. The performance was seen not sensitive to δ and a variety of values for δ worked quite similarly.

4. Conclusions

In this paper, we proposed a new approach called ASBTCS to improve the performance of automatic speech segmentation for concatenative speech synthesis. It has been found beneficial to select the best ASM among a variety of ASMs depending on the boundary type. The experimental results have shown that the proposed method remarkably improves the accuracy of the automatic segmentation technique.

5. Acknowledgements

This work was partly supported by IT R&D Project funded by Korean Ministry of Information and Communications.

6. References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, pp. 373-376, 1996.

[2] S. Paulo and L. C. Oliveira, "DTW-based phonetic alignment using multiple acoustic features," in *Proc. Eurospeech*, Geneva, Switzerland, pp. 309-312, 2003.

[3] H. Kawai and T. Toda, "An evaluation of automatic phone segmentation for concatenative speech synthesis," in *Proc. ICASSP*, vol. I, Montreal, Canada, pp. 677-680, 2004.

[4] D. T. Toledano, L. A. H. Gómez, and L. V. Grande, "Automatic phonetic segmentation," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 617-625, Nov. 2003.

[5] J. Matoušek, D. Tihelka, and J. Psutka, "Automatic segmentation for czech concatenative speech synthesis using statistical approach with boundary-specific correction," in *Proc. Eurospeech*, Geneva, Switzerland, pp. 301-304, 2003.

[6] T. Saito, "On the use of F0 features in automatic segmentation for speech synthesis," in *Proc. ICSLP*, vol. VII, Sydney, Australia, pp. 2839-2842, 1998.

[7] C. D. Mitchel, M. P. Harper, and L. H. Jamieson, "Using explicit segmentation to improve HMM phone recognition," in *Proc. ICASSP*, vol. I, Detroit, USA, pp. 229-232, 1995.

[8] J. Kominek and A. W. Black, "A family-of-models approach to HMM-based segmentation for unit selection speech synthsis," in *Proc. ICSLP*, Jeju, Korea, 2004.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. New York: Chapman & Hall, 1984.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2)*. Cambrige University Engineering Department, 2002.