PHONETIC SEGMENTATION USING STATISTICAL CORRECTION AND MULTI-RESOLUTION FUSION

Sixuan Zhao[1], Ing Yann Soon[1], Soo Ngee Koh[1], Kang Kwong Luke[2]

[1] School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore
[2] School of Humanities & Social Sciences, Nanyang Technological University, Singapore

ABSTRACT

This paper focuses on the generation of accurate phonetic segmentations. Statistical methods based on absolute and relative correction are discussed and experimented on both monophone and biphone models to improve the segmentation results. The influence of search range on the statistical correction process is studied and a state selection technique is used to enhance the correction results. This paper also explores the influence of resolution (stepsize) of HMMs and proposes a multi-resolution fusion process to further refine the statistically corrected results. Improvements of segmentation results in terms of segmentation accuracy, mean absolute error (MAE), and root mean square error (RMSE) can be observed by applying the proposed refinement methods.

Index Terms: phonetic segmentation, statistical correction, state selection, multi-resolution

1. INTRODUCTION

Phonetic segmentation refers to the process of identifying the phone level boundaries of an utterance. The labeling information provided by phonetic-level segmentation can provide valuable cues to various applications of speech technologies. For example, Accurate phonetic labeling is necessary for the concatenative text-to-speech (TTS) systems [1] to obtain phone-level speech segments which can be used in the concatenation based speech synthesis process. In addition, phonetic segmentation is required by accent conversion methods [2] which provide informative feedback to non-native English learners in the language learning process. Also, many linguistic studies can use phonetic segmentations to analyze the specified corpus.

Some recent studies on automatic phonetic segmentation have been performed [3-7]. Most of those works are based on hidden Markov models (HMMs) using forced alignment, while [7] uses a large margin classifier. In [3], the concept of statistical corrections on phonetic segmentation results is proposed and the experimental results demonstrate a significant improvement in segmentation accuracy. However, it neither studies the performance of this method on context independent (i.e., monophone) models nor explores the influence of different search ranges on statistical correction. In [4-6], fusion methods are used to combine a number of different HMMs to refine the phonetic segmentation results. One drawback of these kinds of methods is that the segmentation process can be massively time consuming and is hard to be applied for real-time implementations, because segmentation results must be generated by many different HMMs and then merged together. A new large margin algorithm which uses a framework similar to a simplified support vector machine (SVM) is proposed in [7] to do phonetic segmentation, with performances comparable to HMMs method.

Some papers also propose the text-independent phonetic segmentation as in [8, 9], which detect phone boundaries only using acoustic features. However, our main target applications are those in which linguistic information is available, such as TTS or accent conversion. Furthermore, the text-independent phonetic segmentation methods generally underperform the text-dependent methods which utilize linguistic information. Therefore, this paper focuses on the text-dependent segmentation method using HMMs.

This paper relies on statistical corrections and multi-resolution fusion to improve the segmentation results. Although fusion methods have been used in previous works like [4-6], none of them studies the influence of different stepsizes of HMMs on phonetic segmentation. Different from [3], the search range of statistical corrections is considered and the state selection step is proposed to take the benefits of different search ranges, improving the segmentation results.

2. STATISTICAL CORRECTION OF AUTOMATIC PHONETIC SEGMENTAITON

To improve the phonetic segmentation results, an analysis of the segmentation errors using a baseline HMM system is necessary. The HMMs models in this paper are trained using HTK [10]. The training process of the baseline HMMs uses left-context dependent biphone HMMs with 4-state, single mixture, 5 ms stepsize, and 25 ms frame size. The feature vectors consist of 12 MFCCs and normalized energy plus their delta and acceleration. The training and testing are based on the TIMIT English corpus. The phonetic segmentations obtained by the trained HMMs are then compared with the manual segmentations provided by TIMIT to calculate the segmentation error distribution histogram as in Fig. 1.

It can be seen that the statistical distribution of segmentation errors can be fitted with a Gaussian curve, whose mean is -6.05 ms and standard deviation is 8.15 ms. Because the mean is different from 0, a systematic bias exists for each trained HMM, degrading the segmentation accuracy. From the observations above, one possible way to reduce the segmentation error is to compensate the systematic bias. A different systematic bias should be used to correct each class of phone boundary, because the acoustic features around different phone boundaries are different from each other.

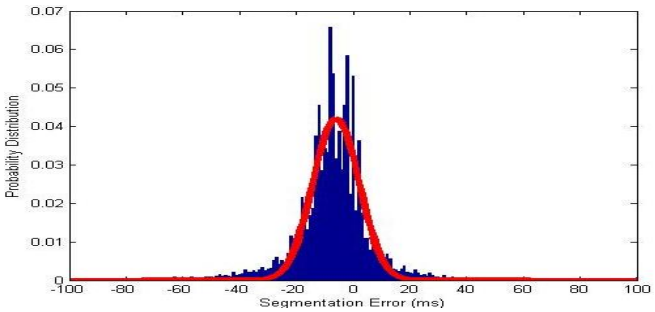

Fig. 1: Segmentation Error Distribution in TIMIT

Two methods are considered to statistically correct the systematic bias. The first method is an absolute method whose correction term is a weighted summation of the error statistics from the statistical correction training dataset:

$$s(i) = \sum_k p_k \times bin[S^{AS}(i) - S^{M}(i)]_k \quad (1)$$

where s is the correction term, i is the index of boundary class, p_k is the probability of segmentation errors falling into the bin ranging from k to k+1 ms, according to the error distribution histogram of the boundary class i calculated from the training data. The $bin[S^{AS}(i) - S^{M}(i)]_k$ indicates the differences between automatic and manual segmentations falling in the k-th bin, having a value of k ms. Since more than 99% of segmentation errors are smaller than 35 ms, the bin size is set as 1 ms, from -40 ms to 39 ms, and k ranges from -40 to 39. The refined phonetic segmentation can then be calculated from the automatic segmentation result and the correction term of the corresponding boundary class i:

$$S^{corr} = S^{AS}(i) - s(i) \quad (2)$$

Although the absolute correction method (i.e., the correction term for each boundary class is a fixed number) can capture the systematic bias of acoustic models, the state-level alignment which is available from the forced alignment is not used. The automatically detected boundary is obtained by the state transition of neighboring HMMs and defined by the onset of the first state of the right phone. Therefore, the statistical correction term can be calculated as a ratio, i.e., a relative term, of the state-level segmentations around the automatically detected boundaries. As the manual segmentation may locate at either side of the automatic boundary, two ratios which account for the errors lying on either side of the automatically detected phone boundary should be calculated:

$$Ls(i) = \underset{i_k}{mean}\{max(0, \min(1, \frac{[SR_1^{AS}(i_k) - S^{M}(i_k)]}{[SR_1^{AS}(i_k) - SL_{4-n+1}^{AS}(i_k)]}))\} \quad (3)$$

$$Rs(i) = \underset{i_k}{mean}\{max(0, \min(1, \frac{[S^{M}(i_k) - SR_1^{AS}(i_k)]}{[SR_{1+n}^{AS}(i_k) - SR_1^{AS}(i_k)]}))\} \quad (4)$$

where n is the search range (the number of considered states around the automatically detected phone boundary, from 1 to 3), $S^{M}(i_k)$, $SL_j^{AS}(i_k)$ and $SR_j^{AS}(i_k)$ are the manual segmentation, j-th state-level segmentation of the left phone, and j-th state-level segmentation of the right phone, corresponding to the k-th boundary of class i; $Ls(i)$ and $Rs(i)$ are the error correction ratios on left and right side of the boundary class i. The scheme is similar to the one used in [3], but it considers correction ratio in different search ranges (1-3 states in a 4-state HMM). The 1st state of the left phone and the 4th state of the right phone are excluded to avoid influences from neighboring boundaries. To enjoy the benefits of different search ranges, we use a state selection method, which calculates the statistics with different search ranges during the training phase and selects the appropriate range for each class i by choosing the one minimizing the mean distance between corrected and manual segmentations. Hence, different classes can use different search ranges to refine phone boundaries. This selection process is shown in Fig. 2 and details will be discussed in section 4.

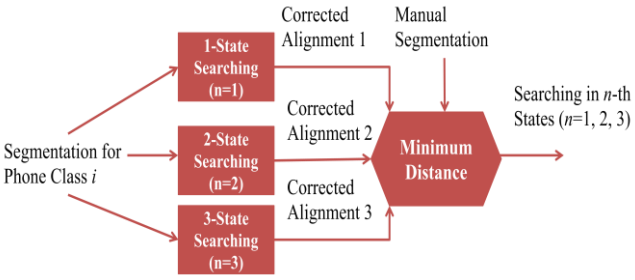

Fig. 2: State Selection for Relative Statistical Correction

Once the relative correction ratios for each boundary class are obtained, the phonetic boundaries can be refined according to the class index i of the current phoneme and the correction terms:

$$S^{corr} = S^{AS}(i) + Rs(i) \times (SR_{1+n}^{AS}(i) - SR_1^{AS}(i)) - Ls(i) \times (SR_1^{AS}(i) - SL_{4-n+1}^{AS}(i)) \quad (5)$$

3. MULTI-RESOLUTION FUSION OF SEGMENTATION RESULTS

The resolution, or stepsize, of the HMMs used for phonetic segmentation is important, because it can decide the minimum resolution of the segmentation results. The current segmentation methods always use a uniform stepsize such as 5 ms or 10 ms to train the HMMs. Even in systems using fusion methods like [5, 6], the differences among HMMs are only related to the number of states or mixtures, without considering the stepsize of HMMs. However, different stepsizes may affect the segmentation results. Although HMMs with 5 ms stepsize produces a high accuracy due to its high resolution, HMMs with 10 ms stepsize may generate a more accurate boundary in certain situations, as shown in Fig. 3:

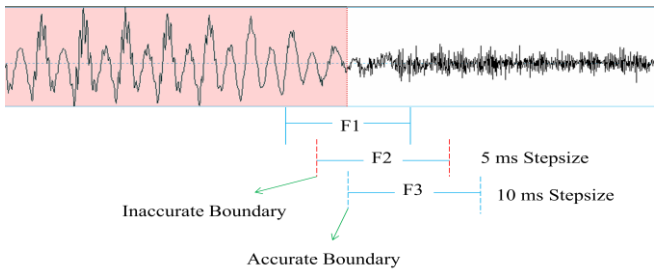

Fig. 3: Influences of HMMs Stepsize

In Fig. 3, F1 refers to the previous frame which locates around the boundary between the voiced and the unvoiced segments. Therefore, the feature vector extracted in such a frame can still represent the voiced segment, i.e., the left part of the phone boundary. If the stepsize is 5 ms, the next frame (F2) will obviously contain more information about the unvoiced segment, and the HMMs will indicate a transition from voiced segment to unvoiced segment, i.e., the phone boundary, based on the

probability calculated from features of F2. This phone boundary, however, is not accurate because the detected phone transition, which is the beginning of F2, locates at the left side of the real phone boundary. In contrast, HMMs with 10 ms stepsize can generate the next frame (F3) whose left side is very close to the real phone boundary. In other words, the larger stepsize help the segmentation system "skip" the ambiguous part in such a situation to obtain a more accurate phone boundary. Although theoretically the errors are determined by the stepsize of HMMs, the inaccuracy of acoustic models can lead to much larger errors (e.g., >20 ms). Considering this issue, HMMs with different stepsizes may contribute differently to the segmentation results.

To study the benefits of HMMs with different resolutions, segmentations are performed on TIMIT corpus using biphone HMMs with 5ms and 10ms stepsize. Results are shown in Table 1:

Table 1 HMMs Performance with Different Stepsizes

	< 10 ms	<20 ms	<30 ms	MAE (ms)	RMSE (ms)
5ms Stepsize	**57.83%**	**84.72%**	92.36%	**12.13**	19.77
10ms Stepsize	52.11%	82.37%	**93.00%**	12.52	**18.40**

From the table, HMMs with 5ms stepsize has a higher accuracy with smaller tolerances (i.e., 10 ms & 20 ms) and a lower MAE. However, HMMs with 10 ms stepsize can generate a higher accuracy with the large tolerance (i.e., 30 ms) and a lower RMSE, which gives higher weights to big errors. Therefore, it seems that smaller stepsize can reduce small errors while larger stepsize can reduce large errors, showing that different resolutions contribute to segmentation results in different ranges. In order to take the benefits of HMMs with different resolutions, a fusion method can be used to integrate the segmentation results from HMMs with different resolutions. In this paper, the support vector regression (SVR) is used to combine HMMs with different resolutions.

4. EXPERIMENTAL RESULTS & DISCUSSIONS

Experiments based on TIMIT corpus are conducted to test all the proposed methods. The 3696 utterances in TIMIT training set, eliminating all the SA sentences which are common for all the speakers, are used to train HMMs models. The testing corpus consisting of 1344 utterances are used for the training of statistical correction terms and testing of the results. A total of 48 phonemes as proposed in [11] are used for modeling. HMMs have 4-state and single mixture, with a frame size of 25 ms and an initial stepsize of 5 ms. Feature vectors consist of 12 MFCCs and normalized energy plus their delta and acceleration (39-dimensional). The 5-fold cross validation is used for each case on the TIMIT testing set. Both monophone and biphone models are used for experiments. Biphone rather than triphone is used here because biphone can model the transition between two phones, which is more relevant to the detection of phone boundaries between the two phones. For monophones, the boundary class is determined by the phone whose onset is given by the segmentation, leading to a total of 48 classes. For biphones, a decision tree asking linguistic questions classifies the 1652 biphones into 763 different biphones (or boundary classes). The correction statistics are calculated for each boundary class. The results of phonetic segmentation with and without statistical correction are shown in Table 2, with the relative method searching in one neighboring states (n=1).

Table 2 Segmentation Results Using Mono-/Bi- phone Models with / without Statistical Correction

	< 10 ms	<20 ms	<30 ms	MAE (ms)	RMSE (ms)
Mono-Original	**65.47%**	**86.86%**	**93.06%**	**10.79**	**18.40**
Bi-Original	57.83%	84.72%	92.36%	12.13	19.77
Mono-Absolute	69.83%	87.98%	93.36%	10.16	17.53
Bi-Absolute	71.85%	88.94%	93.45%	9.41	16.94
Mono-Relative	70.99%	88.14%	93.55%	10.01	17.42
Bi-Relative	**73.00%**	**89.98%**	**94.23%**	**9.22**	**16.31**

It is found that the scheme based on monophones outperforms that based on biphone models without statistical corrections. The reasons is that biphone models include the transition of two phones and cannot provide accurate information to discriminate one phone from its context, while monophone models trained on individual phones have the ability to discriminate the current phone (which is fixed) from its context (which varies) and generate more accurate results. However, biphone models outperform monophone models after statistical corrections, which stems from the context-dependent statistics involved in biphone models. Compared to the monophone models with only 48 groups of statistics, the biphone models provide 763 classes which generate correction statistics for each phone with different contexts, providing more detailed information about the phone boundary.

Table 2 shows that the relative method generally outperforms the absolute method, which may result from the variation of phone durations. The variation of phone durations due to different speaking styles or sentence patterns may degrade the performance of the absolute correction, e.g., the fixed correction bias can be too small for a longer phone or too big for a shorter phone. On the other hand, the relative method reflects the bias as a ratio of the neighboring state durations, obtaining a more accurate correction. This comparison shows that the statistical correction must take into account the phone duration information as in the relative method.

Since the relative correction method works better, a detailed study of the search range (or the number of states involved for correction) is performed as shown in Table 3. The first three rows show a tradeoff of the search range: the smaller search range (1 neighboring state) performs higher accuracy with smaller tolerances and a lower MAE, whereas the larger search range (3 neighboring state) performs higher accuracy with a higher tolerance and a lower RMSE. The reasons are as follows: searching in a smaller range will provide the most accurate local information and thus generate a lower MAE (as most of errors are around -6 ms according to the error distribution in section 2), but may fail to compensate for larger errors; In contrast, searching in a larger range will be able to compensate for the large errors such as

incorrect phone recognition, but may degrade the ability to correct small errors because states far from the preliminary boundary may not be able to provide the most relevant local information.

Table 3 Relative Correction with Different Search Ranges

	< 10 ms	<20 ms	<30 ms	MAE (ms)	RMSE (ms)
1-State (n=1)	**73.00%**	88.98%	94.23%	**9.22**	16.31
2-State (n=2)	70.43%	89.13%	95.05%	9.31	15.75
3-State (n=3)	71.93%	**90.04%**	**95.32%**	9.51	**15.40**
State Selection	**73.64%**	89.79%	**95.58%**	**8.78**	**15.17**

The segmentation performance using state selection is shown in the last row of Table 3. Although the accuracies in terms of 20 ms tolerance are not the highest, it is comparable to the best one of the first three rows using fixed search ranges. However, both the MAE and RMSE are significantly reduced by the state selection method, indicating an overall reduction of the segmentation errors and demonstrating the effectiveness of the state selection step.

After the statistical correction, results generated by HMMs with different resolutions can be combined by SVR to further improve the performance. The SVR is implemented by LibSVM [12]. Three groups of HMMs with different resolutions (5 ms, 7.5ms, 10 ms) trained using the TIMIT training set are used. Higher stepsize is not considered as HMMs with a stepsize of 12.5ms result in both highest MAE (11.49 ms) and highest RMSE (18.21 ms) after statistical correction, showing significantly degraded segmentation results when stepsize is greater than 10 ms. The 5-cross validation segmentation results are shown in Fig. 4.

It can be observed from Fig. 4 that the segmentation results from the 5ms stepsize HMMs have the lowest MAE (8.78 ms) but a relatively high RMSE (15.17 ms), demonstrating lower accuracy in terms of large errors. On the other hand, 10 ms stepsize HMMs have the lowest RMSE (14.00 ms) but a higher MAE (9.12 ms), demonstrating a lower accuracy in terms of small errors. The 7.5 ms stepsize HMMs generally performs in the middle of the other two in terms of both MAE (8.91 ms) and RMSE (14.52 ms). However, the combined results show the lowest MAE (8.17 ms) and lowest RMSE (13.12 ms) compared with all the individual HMMs with different resolutions. Therefore, fusing HMMs with different resolutions can enhance segmentation results.

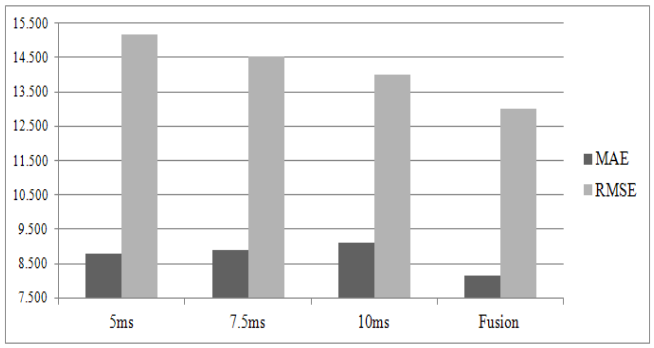

Fig. 4: Segmentation Results of HMMs with Different Resolutions and Multi-resolution Fusion

In all the experiments above, T-tests show significant differences ($t<0.001$) of absolute segmentation errors, i.e., differences between manual and automatic segmentations, for different segmentation setups. Overall improvements by including all the proposed methods are demonstrated in Fig. 5:

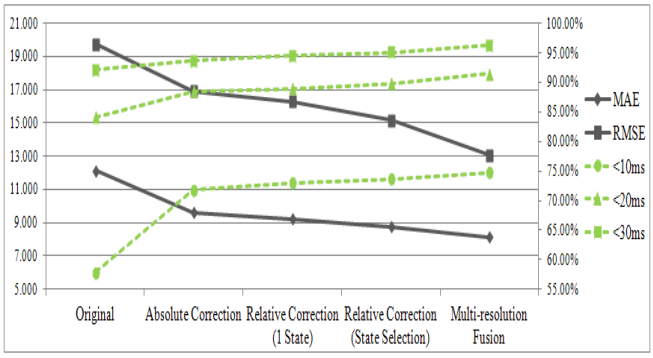

Fig. 5: Overall Improvements of Segmentation Results

The left and right vertical axis indicates the MAE & RMSE in terms of ms and the accuracy in terms of percentage, respectively. It can be found that statistical correction, state selection and multi-resolution fusion all contribute to the improvements of segmentation results in terms of both accuracy and MAE & RMSE. The achieved segmentation results outperform that in [5], which reports MAE of 10.01 and RMSE of 17.15 with the same training and testing sets of the TIMIT corpus..

5. SUMMARY

This paper proposes a refinement process for HMMs based text-dependent automatic phonetic segmentation. Corrections based on both absolute and relative statistics are compared using monophone and biphone models, and the results show that the relative method with biphone models can generate the highest accuracy. A state selection method, which is not considered in previous research using statistical correction [3], is used to refine the relative statistical correction. The proposed multi-resolution fusion step embraces the benefits of HMMs with different stepsizes, which are not studied in other fusion methods [4-6]. Each of the proposed refinement steps can contribute to the segmentation results, detecting phone boundaries more accurately.

In future studies, finer modeling of correction statistics as well as statistical models based voice activity detection techniques should be included to achieve better segmentation results.

6. ACKNOWLEDGMENT

The authors would like to acknowledge the Ph.D. grant from the Institute for Media Innovation, Nanyang Technological University, Singapore.

REFERENCE

[1] A. J. Hunt, and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996, pp. 373-376.

[2] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation

training," *Speech Communication,* vol. 51, no. 10, pp. 920-932, 2009.

[3] D. T. Toledano, L. A. H. Gomez, and L. V. Grande, "Automatic phonetic segmentation," *Speech and Audio Processing, IEEE Transactions on,* vol. 11, no. 6, pp. 617-625, 2003.

[4] P. Seung Seop, and K. Nam Soo, "Automatic Speech Segmentation Based on Boundary-Type Candidate Selection," *Signal Processing Letters, IEEE,* vol. 13, no. 10, pp. 640-643, 2006.

[5] I. Mporas, T. Ganchev, and N. Fakotakis, "Speech segmentation using regression fusion of boundary predictions," *Computer Speech & Language,* vol. 24, no. 2, pp. 273-288, 2010.

[6] S. S. Park, and N. S. Kim, "On using multiple models for automatic speech segmentation," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, no. 8, pp. 2202-2212, 2007.

[7] J. Keshet, S. Shalev-Shwartz, Y. Singer *et al.*, "A Large Margin Algorithm for Speech-to-Phoneme and Music-to-Score Alignment," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, no. 8, pp. 2373-2382, 2007.

[8] G. Almpanidis, M. Kotti, and C. Kotropoulos, "Robust detection of phone boundaries using model selection criteria with few observations," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 17, no. 2, pp. 287-298, 2009.

[9] V. Khanagha, K. Daoudi, O. Pont *et al.*, "Improving text-independent phonetic segmentation based on the microcanonical multiscale formalism," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 4484-4487.

[10] S. Young, G. Evermann, D. Kershaw *et al.*, "The HTK book," 1997.

[11] K. F. Lee, and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 37, no. 11, pp. 1641-1648, 1989.

[12] Chih-Chung Chang, and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001.