A hybrid refinement scheme for intra- and cross- corpora phonetic segmentation

Sixuan Zhao a,⁎, Ing Yann Soon a, Soo Ngee Koh a, Kang Kwong Luke b

a School of Electrical & Electronic Engineering, Nanyang Technological University, 50, Nanyang Avenue, Singapore 639798, Singapore
b School of Humanities & Social Sciences, Nanyang Technological University, 50, Nanyang Avenue, Singapore 639798, Singapore

Received 21 September 2013; received in revised form 6 July 2014; accepted 13 August 2014
Available online 23 August 2014

Abstract

This paper proposes a hybrid refinement scheme for more accurate localization of phonetic boundaries by combining three different post-processing techniques, including statistical correction, fusion, and predictive models. A statistical method based on state-level correction is proposed to improve the segmentation results. Effects of search ranges on the statistical correction process are studied and a state selection scheme is used to enhance the correction results. This paper also examines the effects of time resolution, i.e., stepsize, of acoustic models on the accuracy of segmentation. A multi-resolution fusion process is proposed to further refine the statistically corrected results. Finally, predictive models are designed to improve the segmentation accuracy by incorporating various acoustic features and searching around the preliminary boundary with a smaller stepsize. By applying the hybrid refinement scheme on a well-known corpus, significant improvements of segmentation results in terms of segmentation accuracy with different tolerances, mean absolute error (MAE), and root-mean-square error (RMSE) can be observed. Furthermore, a scenario of cross-corpora segmentation is examined in generating the segmentation results for a new corpus with a small set of labeled data. Experimental results show that the proposed refinement procedure can generate segmentation results comparable to those given by well-trained acoustic models obtained from the new corpus.

© 2014 Elsevier Ltd. All rights reserved.

Keywords: Phonetic segmentation; Statistical correction; Multi-resolution; Predictive model; Cross-corpora segmentation

1. Introduction

In many state-of-the-art speech processing techniques, corpus-based methodologies play an important role. In order to perform accurate speech recognition, speech synthesis, or other applications, an appropriate and informative speech corpus is indispensable. To facilitate the development of speech tools, the corpus should provide information about the content (labeling) as well as the time alignment (segmentation) of speech signals (Campbell and Black, 1997; Narayanan and Alwan, 2005).
Labeling at the phone level is the most popular and desirable way of segmentation, as phone is the smallest unit of speech, with which larger units (e.g., syllable, word, etc.) can be constructed (Ladefoged and Johnstone, 1982). With accurate phonetic labeling and time alignment, a number of tasks can be performed. The most well-known topic related to automatic phonetic segmentation is the concatenative text-to-speech (TTS) system (Chappell and Hansen, 2002; Hunt and Black, 1996), which requires reliable time alignment to generate separate speech units. Phonetic segmentation can also enhance accent conversion systems (Felps et al., 2009; Felps and Gutierrez-Osuna, 2010; Huckvale and Yanagisawa, 2007) which require time alignments to enable the modification of prosodic features such as pitch and duration produced by non-native speakers, contributing to computer-aided language learning (CALL).

Although manual labeling can generate the most accurate segmentations, it is very costly and labor-intensive, requiring a lot of time and effort (Cosi et al., 1991; Ljolje et al., 1997). In addition, manual segmentation could introduce labeler subjectivity and may not be able to maintain labeling consistency (Pellom and Hansen, 1998; van Hemert, 1991). As a result, automatic segmentation is often required to facilitate the acquisition of phone boundaries. A number of attempts have been made to enhance the phonetic segmentation process in the past. These can be classified into implicit methods, which purely rely on acoustic features and spectral information without prior knowledge about phonetic sequences, and explicit methods, which incorporate both phonetic transcriptions and acoustic features to produce phone boundaries (van Hemert, 1991).

Recent studies on implicit phonetic segmentation are presented in Almpanidis et al. (2009), and Khanagha et al. (2011). The approach given by Almpanidis et al. (2009) models two speech segments with either two separate Gaussian models (M1) or one single Gaussian model (M0). The posterior probabilities given by M1 and M0 are then compared to determine whether a phone boundary should exist between segments or not. In addition to this approach, the study in Khanagha et al. (2011) applies singularity exponents to detect dynamics in the speech signal for the identification of phonetic boundaries. However, implicit phonetic segmentation generally underperforms the explicit one which utilizes linguistic information. Furthermore, linguistic contents are available in most popular applications of phonetic segmentation, such as TTS or accent conversion system. Therefore, more focus should be given to explicit segmentation which is linguistically constrained.

Current mainstream explicit segmentation methods are based on hidden Markov model (HMM) speech recognizers, using forced alignment based on Viterbi decoding (Brugnara et al., 1993; Malfrière et al., 2003; Namnabat and Homayounpour, 2006; Toledano et al., 2003; Yan et al., 2006). Unlike implicit segmentation which mainly relies on spectral information, this kind of method involves both transcription information and the training of acoustic models. The representative features, normally Mel-frequency cepstral coefficients (MFCCs) plus its delta and acceleration, are input to the trained acoustic models along with the phonetic level transcriptions to detect phone boundaries. Normally, the speech utterances to be segmented are considered to have been pronounced correctly, i.e., they exactly match the given transcriptions. This kind of baseline segmentation system using forced alignment can generate reasonable segmentation accuracy within a certain range (30–50 ms). However, further studies on phonetic segmentation are still necessary to improve the accuracy and various approaches have been proposed in recent years. Recent studies on explicit segmentation can be categorized into two groups: enhancement of acoustic modeling, which modifies the structure of acoustic models, and post-processing method, which refines the preliminary boundaries generated by an existing system. This paper focuses on the latter one, intending to improve the segmentation results obtained from an existing system conveniently.

Among different modifications on acoustic modeling schemes, a large margin algorithm which uses a framework similar to support vector machine (SVM) is proposed in Keshet et al. (2007) to achieve phonetic segmentations, producing more accurate boundaries compared with those given by the baseline method in Brugnara et al. (1993). A modified HMM scheme using artificial neural network (ANN) is also proposed in Hosom (2009). Both phone and phone-transition probabilities are calculated based on phonetic features to generate more accurate time alignments. In Yuan et al. (2013), a phone boundary based segmentation model is used. In addition to phone HMMs, phone boundaries are also modeled with special 1-state HMMs. These phone boundary models are then combined with phone HMMs to enhance segmentation results. In Adell et al. (2005), dynamic time warping (DTW) is applied in combination with an acoustic clustering method to generate improved phonetic boundaries for TTS systems by aligning the segmented voice with unsegmented ones.

Post-processing of segmentation results is also studied in a number of papers to obtain accurate phone boundaries from an arbitrary segmentation system. Statistical correction and fusion methods have been proposed to improve phonetic segmentation results, as in Jarifi et al. (2008), Jindrich et al. (2003), Mporas et al. (2010), Park and Kim
(2007), Seung Seop and Nam Soo (2006), and Toledano et al. (2003). In Jindrich et al. (2003), and Toledano et al. (2003), the concept of statistical correction on phonetic segmentation using either absolute or relative terms is proposed and the experimental results demonstrate a significant improvement of segmentation accuracy. In Mpors et al. (2010), different fusion methods are studied to combine a number of different HMMs for the refinement of segmentation results. Phonetic boundaries given by 112 baseline segmentation engines (BSEs) are combined to produce the final time alignments. A general fusion framework which incorporates hard and soft supervisions to combine boundaries given by different algorithms is proposed in Jarifi et al. (2008) for improving the segmentation accuracy. Fusion methods are also applied in Park and Kim (2007), and Seung Seop and Nam Soo (2006), which either select the most appropriate boundary from segmentation results given by different automatic segmentation machines (ASMs) or obtain a weighted summation of segmentation results given by various ASMs as the final boundary.

Post-processing of the preliminary phonetic segmentations generated by forced alignment using predictive models are studied in Akdemir and Ciloglu (2010), Cheng-Yuan and Jyh-Shing (2007), Lee (2006), Lin et al. (2005a), and Namnabat and Homayounpour (2006). In Kim and Conkie (2002), a spectral transition measure is applied to identify the maximum spectral distortions and the phone boundary. The method proposed in Lin et al. (2005a) performs feature selection using sequential forward selection and refines the boundary using k-nearest neighborhood (KNN) based on the selected features. A multiple-layer perception (MLP) based refinement process is studied in Lee (2006), training a MLP for each kind of phone transition groups and using the obtained model for correction. The methods presented in Cheng-Yuan and Jyh-Shing (2007), Lo and Wang (2007), and Namnabat and Homayounpour (2006) refine the phone boundary using SVM. The feature vectors combining both spectral and prosodic information are used to represent the frames around the preliminary boundary from forced alignment, and then SVMs are used to identify the most probable boundary. A new HMM topology is presented in Akdemir and Ciloglu (2010) as the refiner and produces improved segmentation results.

Modifications on acoustic models can contribute to the segmentation process by providing alternative structures of segmentation systems as well as improving the accuracy. However, applying such kind of methods requires a re-structuring and re-training of all the acoustic models and thus cannot conveniently refine segmentation results generated by existing systems. What is more, some acoustic modeling techniques may sacrifice the segmentation results in a certain range to improve the overall performance. For example, the segmentation accuracy within 5 ms is reduced in Hosom (2009) after applying the proposed model. Post-processing methods can be applied to any existing system directly, e.g., a trained speech recognizer, providing increased convenience and flexibility for the segmentation process. It also imitates the segmentation process of human labelers: first listening to the speech signal to get a rough boundary (baseline system), then examining the spectrogram or waveform in detail to identify the accurate boundary (post-processing). The main focus of this paper is the development of a hybrid refinement scheme to overcome some gaps in current studies on this kind of methods so as to improve the segmentation results obtained from a baseline system.

2. A hybrid scheme for post-processing of phonetic boundaries

Post-processing methods mentioned above can improve the accuracy of phonetic segmentation to a certain extent. However, these methods contribute to the segmentation process from different aspects, e.g., correction, fusion, additional models, etc. As none in the literature studies the combination of different post-processing methods, it is desirable to investigate a hybrid scheme which takes into account the benefits of various refinement methods for achieving more accurate segmentation results. This paper mainly focuses on two issues: designing a hybrid scheme for text-dependent phonetic segmentation and using this scheme for cross-corpora segmentation.

The first contribution of this paper comprises the modification and combination of three refinement methods to improve phonetic boundaries given by an existing segmentation system. The scheme presented in this paper is an extension of the previous work reported in Zhao et al. (2013). It is a hybrid refinement scheme for phonetic segmentation consisting of three components: (1) statistical correction which addresses systematic biases of acoustic models using local information from the most relevant range (i.e., “state-selection” in Section 3); (2) a fusion method which incorporates complementary effects from acoustic models with various resolutions (i.e., stepsizes — the interval to extract each frame of feature vector) which are demonstrated to affect the results significantly; and (3) predictive models which correct non-systematic segmentation biases using predictive models. The flowchart of this system is shown in Fig. 1. High, middle, and low resolution models refer to acoustic models with different stepsizes (e.g., 5 ms, 7.5 ms, or 10 ms), which are further illustrated in Section 3.
In this hybrid segmentation system, we propose several innovations to address some issues in existing methods so as to improve the segmentation accuracy: (1) In contrast to previous studies (Toledano et al., 2003; Yuan et al., 2013) which correct biases either using a fixed number or searching in a fixed range, this paper uses state-selection to determine different search ranges for different phone boundaries; (2) the fusion method in this paper studies the effects of stepsize, which have not been investigated, and combines segmentations given by different HMMs to improve the accuracy; (3) predictive models with a smaller stepsize as well as a different classification scheme are used to further refine the phone boundaries as described in Section 4. All of these methods are examined on a well-known multi-speaker corpus, i.e., TIMIT (John Garofolo, 1993).

The second contribution of this paper focuses on the study of a scenario of cross-corpora segmentation. Sometimes it is necessary to perform phonetic segmentation on a new corpus without or only with a small set of labeled data for linguistic analysis or engineering purposes. One way to obtain the required phonetic segmentation is first to train acoustic models on a standard corpus and then perform segmentation on the new corpus using the obtained acoustic models. However, there are some drawbacks of this direct application of pre-trained models. First, the new corpus may have properties different from the corpus used to train the acoustic models. For instance, if the corpus for the training of models is a small vocabulary corpus like TIMIT and the new corpus is for large vocabulary continuous speech recognition (LVCSR) or vice versa, both the transcriptions and reading styles can be different in the two corpora and segmentation results will be affected. Second, the new corpus may be constructed in some English accents different from the corpus used to train acoustic models. Under such circumstances, the pronunciation variations in the two corpora can lead to extra biases in the segmentation process and generate poor results compared with those given by the acoustic models trained on sufficiently labeled data from the new corpus.

Recent studies on speech segmentations are performed on a single corpus with the same accent (Cheng-Yuan and Jyh-Shing, 2007; Lo and Wang, 2007; Mporas et al., 2010; Park and Kim, 2007; Seung Seop and Nam Soo, 2006) and not tested under a cross-corpora scenario. Post-processing scheme is particularly appropriate under this kind of scenario, because the segmentation results given by the standard acoustic models can be incrementally refined using the limited labeled data from the new corpus. To address this issue, the proposed refinement process is applied to another corpus with an English accent different from the training corpus to examine the applicability and effectiveness of cross-corpora segmentation.

The remainder of this paper is organized as follows. Section 3 discusses the proposed statistical correction and multi-resolution fusion methods, while refinements using predictive models are introduced in Section 4. Experiments are conducted to assess the proposed refinement scheme in Section 5 followed by discussions. The cross-corpora segmentations, i.e., training and testing on different corpora, are performed in Section 6. Summary and future works are given in Section 7.
3. Statistical correction on segmentation results & multi-resolution fusion

The baseline segmentation system based on forced alignment suffers from systematic biases, as it seeks to maximize the probability of observations of each phone given the corresponding acoustic model rather than optimizing the model to reduce segmentation errors using boundary information (Kominek et al., 2003). Therefore, we trained a baseline HMM system using the experimental setups described in Section 5 and tested it on the TIMIT testing set to demonstrate this phenomenon. It is shown that the overall statistical distribution of segmentation errors can be fitted by a Gaussian curve, with a mean of −5.5 ms and a standard deviation of 10.8 ms. This observation demonstrates a systematic bias in the acoustic models, as the mean error is away from zero and the errors distribute across a wide range. For a more concise discussion and comparison, this observation is given in Fig. 3(a) in Section 5 and not presented separately.

To correct systematic biases, the state-level segmentation obtained from the forced alignment process is used to calculate correction terms for each phone boundary class. As the automatically detected boundary is defined by the onset of the first state of a phone, it is possible to calculate the correction term as a ratio, i.e., a relative term, of the state-level segmentations around the automatically detected boundaries (Zhao et al., 2013):

$$L_s(i) = \text{mean}_{i_k} \left\{ \max \left( 0, \min \left( 1, \frac{S_{R_{1}}^{AS}(i_k) - S_{L_{m-n_i+1}}^{AS}(i_k)}{S_{R_{1}}^{AS}(i) - SL_{m-n_i+1}^{AS}(i_k)} \right) \right) \right\}$$

(1)

$$R_s(i) = \text{mean}_{i_k} \left\{ \max \left( 0, \min \left( 1, \frac{S_{M}(i_k) - S_{R_{1+1}}^{AS}(i_k)}{S_{R_{1}}^{AS}(i_k) - i_k} \right) \right) \right\}$$

(2)

$$S_{n_i}^{corr} = S_{i}^{AS} + R S(i) \times (S_{R_{1+1}}^{AS}(i_k) - S_{R_{1}}^{AS}(i)) - L S(i) \times (S_{R_{1}}^{AS}(i) - SL_{m-n_i+1}^{AS}(i))$$

(3)

$$n_i = \arg \min_{1 \leq n \leq m} \left| S_{n_i}^{AS} - S_{n_i}^{corr} \right|$$

(4)

where $m$ is the total number of states per phone, $n_i$ is the search range for boundary class $i$, $S_{M}(i_k)$, $S_{L_{m-n_i+1}}^{AS}(i_k)$, and $S_{R_{1}}^{AS}(i_k)$ are the manual segmentation, $j$-th state-level segmentation of the left phone (relative to the boundary), and $j$-th state-level segmentation of the right phone, corresponding to the $k$-th observation of boundary class $i$; $L_s(i)$ and $R_s(i)$ are the error correction ratios on the left and right sides of class $i$. The boundaries are then refined by Eq. (3). Because different phone boundary classes have different phone transitions as well as pronunciation schemes, the search ranges of phone boundary classes should also be different (as in above equations). Therefore, a state-selection process is used to determine the appropriate search range for each phone boundary class. It chooses the search range which minimizes the differences between the corrected and manual segmentations as given in Eq. (4).

In addition to statistical correction, the effects of the resolution of HMMs are also examined, because stepsize defines the minimum movement of a frame and is thus linked to the localization of time alignments. Therefore, segmentations were performed on TIMIT using both CI and CD HMMs with stepsize from 2.5 ms to 12.5 ms, i.e., resolutions from high to low. The experimental setups are the same as those described in Section 5. Segmentation results are shown in Table 1, evaluated as either percentages of the differences between manual and automatic segmentations within a specific range, i.e., segmentation accuracy, or mean absolute error (MAE) and root mean square error (RMSE) between manual and automatic segmentations. Statistical tests, i.e., $t$-tests, show significant differences ($P<0.001$) among different MAEs and RMSEs.

It is shown that the stepsize should not be greater than 10 ms, because segmentation results given by a stepsize of 12.5 ms deteriorate drastically compared with other stepizes. In addition, stepsizes smaller than 5 ms, e.g., 2.5 ms, can lead to degradation of accuracies across all the ranges in comparison with 5 ms stepsize as well as worse overall performance (higher MAE & RMSE). In addition, it increases the runtime significantly, which is undesirable for the real-time application of an automatic segmentation system. Therefore, the segmentation process should only involve HMMs with stepsizes of 5 ms, 7.5 ms, and 10 ms. From Table 1, it can be observed that stepizes in the range of 5–10 ms can affect the performance of HMMs with a specific pattern: the larger the stepsize, the higher the accuracy for large tolerances and the lower the accuracy for small tolerances. Because the HMMs have different benefits for different tolerances, it is worthwhile to combine the segmentation results generated by them using a regression method so that they can compensate for each other. To further correct segmentation errors, predictive models are also applied as the final step by searching a small region around the preliminary boundary with a smaller stepsize as described in the next
performances where section.

\[
\text{CI-2.5} \quad 86
\]
\[
\text{CD-12.5} \quad 10
\]
\[
\text{CD-7.5} \quad 91.40
\]
\[
\text{CD-12.5} \quad 94.72
\]
\[
\text{CD-2.5} \quad 42.73
\]
\[
\text{CD-10} \quad 87.36
\]
\[
\text{CD-10.5} \quad 87.35
\]
\[
\text{CD-5} \quad 90.21
\]
\[
\text{CD-7.5} \quad 90.20
\]
\[
\text{CD-10} \quad 90.22
\]
\[
\text{CD-12.5} \quad 90.03
\]

\section*{4. Refinement based on predictive model}

Although statistical correction and fusion methods can reduce the segmentation errors to a certain extent, predictive models are able to utilize flexible acoustic features to improve the segmentation accuracy and thus are worthy of consideration. Such kind of methods searches frames around the preliminary boundary, identifying the frame which maximizes the posterior probability obtained from feature vectors and the refinement model as the corrected final boundary:

\[
T_{\text{fin}} = \arg \min_{t_n \in (t_i - r, t_i + r)} \{ P(t_n | o_n) \}
\]  

where \( T_{\text{fin}} \) is the corrected final boundary, \( t_i \) is the \( i \)-th preliminary phone boundary given by forced alignment with statistical correction as stated before, \( r \) is a range around the preliminary phone boundary, \( o_n \) is the observed feature vector, and \( P \) is the probability denoting \( t_n \) as the phone boundary given the trained predictive model and the corresponding feature vector \( o_n \). Both SVM (Burges, 1998) and linear discriminative analysis (LDA) (Yu and Yang, 2001) are considered as predictive models to further refine the phone boundaries corrected by statistical correction and multi-resolution fusion. Although other methods like neural network may also be used, SVM and LDA are more favorable as they require less parameter tuning steps and thus can be more adaptable to various scenarios.

The first step of predictive model based refinement is feature extraction. As in (Cheng-Yuan and Jyh-Shing, 2007; Namnabat and Homayounpour, 2006), 13-order MFCCs are used. Normalized energy, log pitch value, entropy, bisector frequency, and burst degree as proposed in (Lin et al., 2005b) are also included to form a more informative feature vector. According to (Lin et al., 2005b), bisector frequency and burst degree are defined as:

\[
\text{bisector Freq} = \frac{\arg \min_{1 \leq k \leq N} \left[ \sum_{j=1}^{k} A_f - 0.5 \sum_{j=1}^{N} A_f \right]}{N} \times \text{sample Rate} \quad \text{(7)}
\]
\[
\text{burst Degree} = \frac{W_1 \times (1/(\text{avg(\text{localMaxInterval})))) + W_2}{W_1 + W_2} \quad \text{(8)}
\]
where $A_j$ is the amplitude, $N$ is the number of frequency components in the spectrum, $W_1$ and $W_2$ are parameters set as 4 and 1, respectively, and $\text{avg(localMaxInterval)}$ is the average distance between the positions of neighboring local maxima of sample points.

In addition, the dynamic features are also calculated in the form as given in Young et al. (2006):

$$DF(t) = \frac{\sum_{k=-M}^{M} F(t + k)k}{\sum_{k=-M}^{M} k^2}$$

(9)

where $M$ is set to 2 in this case and $F(t)$ is the feature vector at time $t$. The obtained dynamic features are then combined with static features to form a 36 dimensional feature vector for each frame.

For each phone boundary class, a binary classifier is trained, with one class corresponding to frames on the left of the manual phone boundary and the other corresponding to frames on the right of the manual phone boundary. This labeling scheme is different from the previous one as in Lee (2006), Lin et al. (2005a), which labels frames close to and far away from the true boundary as two different classes. They only label a few samples around the true boundary as positive instances (frames close to the true boundary) while the rest are labeled as negative instances (frames away from the true boundary). Therefore, more negative instances than positive instances are involved in the training process. Because classifiers are trained to minimize the number of misclassified instances, the resulted classifier will be biased to predict more instances as negative. To overcome this issue, binary labels $-1$ or $1$ are assigned to each frame to represent frames locating at the left or right of the manual boundary, and the change of labels from $-1$ to 1 is detected as the refined phone boundary. This labeling scheme can assure that the training instances for both classes are the same, thus resolving the dataset imbalance problem. When multiple changes of labels exist, e.g., “$-1 -1 1 1 -1 -1 1 1 1$”, the change closest to the preliminary boundary is selected as the final boundary. The reason is that the phone segmentation error is comparatively small after statistical correction and multi-resolution fusion and thus it is unlikely, or at least only in very rare cases, that the refined boundary is far from the real boundary. Therefore, this refinement scheme using predictive models focuses on the reduction of small segmentation errors which is one of our objectives in this paper.

Frame length of 20 ms and stepsize of 2.5 ms are used in this process to correct the segmentation errors which may be neglected by the forced alignment process. In our system design, statistical correction aims at achieving rough or intermediary results. Predictive models are applied to further refine the boundaries with smaller stepsizes. As only the frames around the intermediary boundaries (given by forced alignment and statistical correction plus fusion) are examined with 2.5 ms, the runtime can be reduced. Classifiers are trained using speech files (data) and manual segmentations (labels) from the training set of the TIMIT corpus. For each observation of the manual phone boundary, 20 frames left to the boundary and 20 frames right to that boundary are extracted, because almost all the detected boundaries are within 50 ms around the real boundary (as shown in Fig. 3(a)). Then, all the frames corresponding to the same phone boundary class are used to train the binary classification model. With the trained classification model, the refinement process is performed by examining 20 frames around the automatic boundary to localize the final boundary. To ensure the inclusion of enough training samples, only those boundaries with more than 10 observations are used to train models and refined in the testing phase with the obtained predictive models.

5. Improving segmentation results with the hybrid refinement scheme

To test the proposed refinement scheme, a number of experiments were conducted on the TIMIT corpus using HTK (Young et al., 2006). TIMIT is a multi-speaker database containing a total of 6300 utterances pronounced by 630 speakers. The 61 phone sets in TIMIT were merged into the classical 48 phone sets as in Lee and Hon (1989), and Mporas et al. (2010). The baseline forced alignment system used HMMs with 4 states, 4 mixtures, 5-ms stepsize, and 25-ms frame length. We used 4 states rather than 3 states to provide more options for the state-selection process. The 4-state HMMs are left-to-right models with no skips. A total of 4 mixtures are used in this paper, because recent studies have shown that 4 mixtures are sufficient for phonetic segmentation (Park and Kim, 2007). Although theoretically higher mixtures may slightly affect the results, they require much more expensive for implementation. In addition, HMMs with 4 mixtures can also achieve good results as shown later. Hamming window and a pre-emphasis coefficient of 0.97 were applied. Feature vectors consist of 12 MFCCs with cepstral mean and energy normalization plus their delta and acceleration, resulting in 39 dimensions for each vector. HMMs were trained on the TIMIT training set excluding
SA utterances, i.e., a total of 3696 utterances. Both context-independent (CI) models, i.e., monophones, and context-dependent (CD) models, i.e., biphones, were studied in this paper. Biphone rather than triphone was used because it can model the transition between two phones, which is more suitable for the detection of phone boundaries between each pair of phones. A decision tree which asks linguistic related questions (e.g., whether the left phone is a vowel, the right phone is a nasal, etc.) was used to group all the phone boundaries into 763 classes. To generate these classes, the outlier threshold, i.e., R0 in HTK, was set as 100 and the cluster log-likelihood threshold, i.e., TB in HTK, was set as 300. Automatic segmentations were performed on the TIMIT testing set excluding the SA utterances, i.e., a total of 1344 utterances, using HMMs trained with conditions mentioned above. A subset, i.e., 1000 sentences, of the training set of TIMIT was used to obtain predictive models based on manual segmentations and acoustic features of speech files, as informal experiments showed that the improvements in accuracies are very limited even when more samples are used for the training of models.

The effects of state-level statistical correction were first studied. Following the approach reported in many of the previous studies (Brugnara et al., 1993; Hosom, 2009; Park and Kim, 2007), we started from HMMs with 5 ms stepsize. Results of phonetic segmentations with/without statistical correction by searching in 1-state range are shown in Table 2. The first two rows are repeated from Table 1 to facilitate comparisons of results. The scheme based on CI models outperforms that based on CD models without statistical corrections, but underperforms CD models with corrections. The underperformance of CD models as compared with CI models stems from the fact that CD models include the transition of two phones and hence cannot provide accurate information for discriminating a phone from its context. In contrast, CI models trained on individual phones have the ability to discriminate a phone (which is fixed) from its context (which varies). Similar observations can also be found in Ljolje et al. (1997), and Mporas et al. (2010). However, both CI and CD models work well after the statistical correction based on relative values, showing that the systematic biases in HMMs are addressed effectively by this step. Also, CD models outperform CI models after corrections. The reason is that statistics used for corrections are trained according to phone boundary classes which are the same as the modeling scheme of CD HMMs, i.e., modeling each phone boundary class with one HMM. As a result, the correction terms are obtained with the same grouping scheme of phone boundaries as that used for the training of CD HMMs. This leads to improved correction of biases, as correction terms are trained particularly for each HMM. In contrast, HMMs in CI models are purely based on each individual phone, not corresponding to the correction terms trained for each phone boundary class. Therefore, the biases of CD HMMs can be corrected in a more precise manner in comparison with those of CI HMMs which are not directly linked to phone boundary classes used to group correction terms. The last two rows list the performances of absolute correction which is used in Jindrich et al. (2003), and Yuan et al. (2013) for comparison. This method corrects the biases using a fixed term obtained from the mean difference between manual and automatic segmentations for each phone boundary class. It can be seen that this method underperforms the state-level correction which uses duration information of each observation to generate adaptive correction terms.

Next, a study on the effects of search range is presented. Statistical corrections with various search ranges are applied on CD models with 5 ms stepsize and the segmentation results are shown in Table 3. From the obtained results, a tradeoff can be observed: a smaller search range (1-state) generates a higher accuracy in the case of a smaller tolerance, e.g., <10 ms, whereas a larger search range (4-state) leads to a higher accuracy in the case of a higher tolerance, e.g., <50 ms. The reason is that the search range determines the states used to calculate the statistics or the portion of the phones...
involved in the training of statistics. Searching in a smaller range will provide the most accurate local information, but it may fail to compensate for larger errors. However, the reverse situation applies for a larger search range. As different phone boundaries are prone to different kinds of biases (e.g., small or big ones) due to their distinctive characteristics, using a single search range is not sufficient. The last row shows the results given by state-selection method which yields both lowest MAE and RMSE. This indicates that an overall reduction in the segmentation errors can be achieved by including the state-selection step. As this step can choose the most appropriate search range for different phone boundary classes and generate correction statistics accordingly, the resulting correction process can be more effective.

After statistical corrections, segmentation results from HMMs with different stepsizes were fused using SVR, which was implemented using LIBSVM (Chang and Lin, 2011) with radial basis function (RBF) kernel. All the features were normalized to the range of \([-1, 1]\) to facilitate the classification process using LIBSVM. A grid search as proposed in (Hsu et al., 2003) was used to set the SVR parameters \(C\) and \(g\) by using 500 utterances and their labels from the TIMIT training set. Three groups of HMMs with different resolutions (5 ms, 7.5 ms, 10 ms) were used to generate phone boundaries. Bigger stepsizes were not considered as HMMs with a stepsize of 12.5 ms already resulted in worse performance across all the criteria as in Table 1, showing significantly degraded segmentation results when stepsize is greater than 10 ms.

The statistically corrected segmentations with state selection approach given by HMMs in different resolutions and the fusion method are shown in Fig. 2, where “Fusion (CD)” means fusion results using only CD models and “Fusion (all)” means fusion results incorporating both CI and CD models. The segmentation results given by the fusion method are consistently better than those generated by HMMs with various resolutions in terms of different criteria like segmentation accuracy, MAE, and RMSE. Although HMMs with a stepsize of 10 ms can generate the highest MAE, it performs relatively well for large tolerances like 40 ms and thus contains complementary information which can be incorporated by the fusion scheme to improve the overall performance. Segmentations given by CI models are also included in the fusion process to enhance the results. It can be observed that the fusion of HMMs with different resolutions makes the main contribution and the inclusion of CI models achieves some incremental improvements in the results, i.e., slight reductions in MAE and RMSE. Therefore, the stepsize of HMMs does play an important role in the segmentation process and the multi-resolution fusion scheme enables HMMs with different resolutions to compensate for the biases of each other for improving segmentation results.

After the fusion process, a group of predictive models were applied as the final refinement step. This step attempts to provide further refinements with a stepsize smaller than before, i.e., 2.5 ms, to resolve small segmentation errors. A classifier was trained for each phone boundary class and the same phone boundary classes generated from the decision tree as mentioned before were used to train the classification models. SVM and LDA were implemented using LIBSVM and Matlab, respectively. The results are shown in Table 4. In addition, segmentation results given by recent studies on phonetic segmentation using TIMIT are shown in the last four rows of Table 4 for comparison.

It can be observed that by using predictive models, certain improvements in the segmentation results can be achieved, and the SVM method outperforms the LDA method. Compared to LDA, SVM is a more discriminative classifier which can maximize the margin between two classes and thus maintains the generality of the developed classification model. The RBF kernel used by SVM also maps acoustic features to a high-dimensional space which can effectively address
non-linearly separable data, such as the acoustic features of speech signals. Compared to statistical correction which mainly addresses the systematic biases using global statistics, predictive models reduce segmentation errors based on local acoustic features around the preliminary boundary. It can be observed that segmentation accuracy with 5 ms tolerance is improved to the most extent by this step, which results from the searching scheme of predictive models, e.g., the use of small stepsize and the preference to small shift from the preliminary boundaries (the final boundary is identified as the change of labels closest to the preliminary boundary). In addition to the contribution of predictive

<table>
<thead>
<tr>
<th>Table 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refinements with predictive model.</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Original</td>
</tr>
<tr>
<td>Stats + fusion</td>
</tr>
<tr>
<td>With LDA</td>
</tr>
<tr>
<td>With SVM</td>
</tr>
<tr>
<td>Mporas et al. (2010)</td>
</tr>
<tr>
<td>Keshet et al. (2007)</td>
</tr>
<tr>
<td>Hosom (2009)</td>
</tr>
<tr>
<td>Yuan et al. (2013)</td>
</tr>
</tbody>
</table>

Bold values are the best results (highest accuracy or lowest MAE/RMSE) in the corresponding column or sub-column.
models, the appropriate selection of search range which provides more relevant local information during the statistical correction and the complementary effects of HMMs with differences in stepsizes, as shown in Table 3 and Fig. 2, also reduce errors in this small tolerance.

The error distributions before/after the refinement process and the reduced MAE for each individual phone in TIMIT are given in Fig. 3. According to Fig. 3(a), the error distribution given by the baseline system as mentioned in Section 3 shows a mean value away from zero and a relatively big variance. By applying the proposed refinements, the error distribution has a mean value much closer to 0 and is more concentrated around the mean value. Overall, both the systematic and non-systematic biases are addressed by the hybrid refinement scheme. In addition, reductions in the mean segmentation errors for all individual phones can be observed in Fig. 3(b). The reduced MAEs are plotted for each phone whose onset is defined by the detected boundary. It is shown that the boundaries of most of the phones in TIMIT are identified more precisely. T-tests show significant reduction ($t<0.005$) in MAE in all the experiments above.
The incremental improvements by incorporating each of the proposed methods are demonstrated in Fig. 4. The left and right vertical axes indicate the MAE and RMSE in ms and the accuracy in percentage, respectively. All the proposed refinement steps contribute to the improvements of segmentation results in terms of accuracy and MAE/RMSE. Compared with the previously reported work in Zhao et al. (2013), the results presented in this paper are improved due to the use of isolated-unit training to obtain improved baseline models as suggested in Donovan (1996), and Yuan et al. (2013), the inclusion of both CI and CD models in the fusion process, and the application of predictive models for refinements. As presented in Table 4, the proposed scheme exhibits higher accuracies on TIMIT as compared with recent studies. The achieved MAE and RMSE are lower than those reported in Mporas et al. (2010) which involves a fusion method for post-processing with the same phone sets and experimental conditions on TIMIT. Compared with results in Hosom (2009), and Mporas et al. (2010), the reported segmentations demonstrate improved accuracies across different tolerances. The results also outperform those in Yuan et al. (2013) in terms of segmentation accuracies. In addition, this hybrid scheme leads to more improvements for small segmentation errors, i.e., those within 5 ms and 10 ms as shown in Fig. 4 and Table 4.

It should be noted that this study focuses on a post-processing scheme to refine phone boundaries given by a baseline segmentation system in comparison with those reported in Hosom (2009), Keshet et al. (2007), and Yuan et al. (2013) which change the acoustic modeling scheme in various ways. Although these reported schemes can contribute to the segmentation accuracy significantly, the proposed scheme involves a concatenation of several different refinement methods to improve segmentation results. The inclusion of each step, i.e., statistical correction, fusion, and refinements with predictive models, can be determined according to the availability of training data and the constraint of implementation time, allowing for increased flexibility and a tradeoff between efficiency and accuracy.

6. Cross-corpora phonetic segmentation

A cross-corpora segmentation study was conducted in this section. In addition to TIMIT, the corpus of Cambridge wall street journal (WSJCAM0) (Robinson et al., 1995) was also involved for experiments. It is a British version of wall street journal (WSJ) corpus recorded in Cambridge University, incorporating utterances produced by 140 speakers. Due to the differences in pronunciations and speaking styles between American and British English, it can be expected that the segmentation results on TIMIT given by acoustic models obtained from WSJCAM0 are likely to be worse as compared with those given by acoustic models trained on TIMIT.

To study cross-corpora segmentation, the British acoustic models (BAM) were trained by WSJCAM0 corpus and tested on TIMIT with/without the proposed refinement scheme, because TIMIT is a standard corpus widely used for segmentation studies. The training scheme of BAM is the same as that for the training of American acoustic models (AAM) with TIMIT, i.e., using the standard Baum–Welch algorithm to optimize parameters of HMMs. The same
Table 5
Cross-corpora segmentation results on TIMIT.

<table>
<thead>
<tr>
<th></th>
<th>&lt;10 ms (%)</th>
<th>&lt;20 ms (%)</th>
<th>&lt;30 ms (%)</th>
<th>&lt;40 ms (%)</th>
<th>&lt;50 ms (%)</th>
<th>MAE (ms)</th>
<th>RMSE (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline BAM</td>
<td>57.67</td>
<td>79.60</td>
<td>88.11</td>
<td>93.06</td>
<td>95.03</td>
<td>14.33</td>
<td>24.23</td>
</tr>
<tr>
<td>BAM + stats</td>
<td>66.84</td>
<td>85.74</td>
<td>91.91</td>
<td>95.26</td>
<td>96.52</td>
<td>11.50</td>
<td>19.51</td>
</tr>
<tr>
<td>BAM + stats + fusion</td>
<td>68.15</td>
<td>87.28</td>
<td>93.31</td>
<td>95.96</td>
<td>97.38</td>
<td>10.29</td>
<td>17.94</td>
</tr>
<tr>
<td>BAM + stats + fusion + SVM</td>
<td>70.04</td>
<td>88.36</td>
<td>93.82</td>
<td>96.35</td>
<td>97.56</td>
<td>9.82</td>
<td>17.17</td>
</tr>
<tr>
<td>AAM</td>
<td>69.36</td>
<td>87.80</td>
<td>93.59</td>
<td>96.27</td>
<td>97.64</td>
<td>9.97</td>
<td>17.36</td>
</tr>
<tr>
<td>AAM200</td>
<td>53.95</td>
<td>76.87</td>
<td>86.25</td>
<td>91.62</td>
<td>93.75</td>
<td>16.71</td>
<td>27.52</td>
</tr>
<tr>
<td>AAM200 + refine200</td>
<td>64.12</td>
<td>83.61</td>
<td>90.12</td>
<td>94.28</td>
<td>95.68</td>
<td>12.86</td>
<td>21.65</td>
</tr>
</tbody>
</table>

Bold values are the best results (highest accuracy or lowest MAE/RMSE) in the corresponding column or sub-column.

window type (Hamming window), pre-emphasis coefficient (0.97), and features (39-dimensional) as those described in Section 5 were applied. The HMMs also had 4 states and 4 mixtures to facilitate the refinement process. Similar to Section 5, the initial stepsize was 5 ms and HMMs with other stepsizes (7.5 ms and 10 ms) were trained for the fusion process. The phonetic transcriptions provided in WSJCAM0 were used for the training of acoustic models. As WSJCAM0 has much more speech data than TIMIT, only a subset of WSJCAM0 data was used for a fair comparison. All the 5013 utterances in CD1 of WSJCAM0 were used as the training set to train British acoustic models which were used to obtain phone boundaries from the testing set of TIMIT. These utterances were produced by 46 different speakers with each one producing approximate 110 utterances. Compared with the training set of TIMIT (new corpus in this case) which consists of 2.68 h of speech signals, the duration of training samples in WSJCAM0 (standard corpus) is 9.62 h. It can be observed that the duration of training samples in WSJCAM0 is longer than that in TIMIT. This situation is acceptable and common in the application of cross-corpora segmentation, as it is expected that the standard corpus (developed for general purposes like the training of standard ASR system) is more complete than the non-standard new corpus (normally developed for a specific purpose like linguistic analysis or adaptation).

Besides, it should also be noted that the number of phones in TIMIT is bigger than that in WSJCAM0. Closures like “cl”, “vcl” and “epi” are not labeled in WSJCAM0 and there seems to be no similar patterns in WSJCAM0 to replace these labels. As the focus of this paper is to resolve the effects of different accents or speaking styles on cross-corpora segmentations rather than dealing with transcription discrepancies, we therefore eliminated these phones in cross-corpora experiments and evaluated the results based on the rest of phones.

Under the cross-corpora segmentation task, BAM which was obtained from WSJCAM0 were used to generate segmentation results for TIMIT testing set. The segmentation results obtained by BAM were then refined by 200 utterances extracted from the training set of TIMIT. These 200 utterances were selected to incorporate a number of different transcriptions in the experiment. As all the sentences in TIMIT are designed to be phonetically rich ones in John Garofolo (1993), these selected utterances can be considered as phonetically rich samples. Again, TIMIT was considered as the new corpus here and only these utterances were taken as labeled in this cross-corpora task. All the refinement parameters were obtained from the 200 utterances from the TIMIT training set while the refinement process is performed on the testing set. Segmentation results are compared with those generated by AAM which is obtained from the whole TIMIT training set. Similar as before, SA sentences in both the training and testing sets were excluded. It should be noted that the biphone structures of the two corpora are quite distinct due to different training transcriptions and accents. In addition, refinements using CI models can also generate significant improvements in segmentation accuracies. Therefore, CI rather than CD models were used here and the correction statistics were trained for individual phones to resolve the discrepancies and enable the cross-corpora segmentation. The segmentation results using BAM, AAM, and BAM plus refinements are shown in Table 5.

The segmentation accuracies of TIMIT trained by AAM are slightly different from those in Table 2, as several phones mentioned above are not assessed. It can be seen that the baseline cross-corpora segmentation results using BAM underperform the baseline segmentation results using AAM which is the well-trained model from the new corpus. Therefore, the direct application of the standard acoustic models (AAM in this case) on the new corpus cannot generate satisfactory segmentation results. The significant differences between segmentation results based on AAM and BAM show that English accents and speaking styles do affect the phonetic segmentation process. Hence, there are...
certain biases caused by the different properties of corpora under this scenario when the forced alignment approach is applied directly. To compensate for these biases, a refinement process is necessary in order to yield improved phonetic boundaries.

Significant improvements in segmentations can be observed if the proposed statistical correction is applied using the small group of labeled utterances from the TIMIT training set. With the fusion methods applied to combine results given by HMMs with stepsizes from 5 ms to 10 ms and refinements done by predictive models, the accuracies can be further improved. The refined segmentation results slightly outperform those obtained by AAM in terms of most of the tolerances and MAE/RMSE. This experiment demonstrates that the proposed refinement process trained on a small set of labeled utterances from the new corpus can produce segmentation accuracies at least comparable to those achieved by well-trained acoustic models obtained from the new corpus, i.e., TIMIT, using the whole labeled training set (3696 labeled utterances).

We also conducted two additional experiments using acoustic models trained by insufficiently labeled American utterances shown in the last two rows of Table 5. The second last row “AAM200” shows the segmentation results...
on TIMIT given by the acoustic models trained on the 200 American English utterances which were used to refine results produced by BAM as in the first part of Table 5. The last row “AAM200 + Refine200” indicates segmentation results given by acoustic models trained by the same 200 American English utterances and then refined by additional 200 American English utterances from the TIMIT training set using “Stat + Fusion + SVM”. It can be observed that these results underperform those given by BAM plus refinements on the 200 American utterances. This phenomenon may result from the insufficient training data for acoustic models. Because the training samples, i.e., labeled American utterances, are limited in the cases of the last two rows, the trained acoustic model “AAM200” is not able to discriminate the probability distributions of different phones effectively. This situation is similar to an undertrained classifier or regressor. For the same reason, these models are also not stable enough to be refined by the proposed steps, as the start point is already not reliable by itself. Therefore, it is not surprising that the segmentation results given by “AAM200 + Refine200” are inaccurate as compared with those from “BAM + Stat + Fusion + SVM”.

The improvements in terms of the error distribution for the whole database and the reductions on MAE for individual phones in TIMIT under this cross-corpora scenario are also plotted in Fig. 5. Compared with the error distribution generated by the baseline system, the kurtosis and biases are corrected with the error distribution more concentrated around 0 after applying the hybrid refinement scheme. In addition, it can be observed that the reduction on MAE is significant for most of the phones after using the hybrid refinement scheme. Particularly, vowels like “IH”, “AE”, “ER”, “AY”, “AH”, “EY”, and “EH” receive high reductions in error, corresponding to their distinctions between British and American English in terms of both spectral and prosodic features as presented in (Yan and Vaseghi, 2002, 2010; Yan et al., 2007). Such kind of distinctions can result from both pronunciation differences and context variations around these vowels across two accents. These observations show that systematic biases in this cross-corpora segmentation scenario, especially those related to the differences between the two English accents in the training and testing sets, are reduced effectively.

Based on the experiments above, the refinement method proposed in this paper provides a convenient and effective way to achieve improved segmentation results for a new corpus by using existing acoustic models and a small group of labeled data from the new corpus, with accuracy comparable to that generated by a forced alignment system as if it were trained by sufficiently labeled data from the new corpus. Besides, this approach can also facilitate the segmentation system in case of the existence of multiple accents by using a standard set of acoustic models plus refinement parameters for each accent rather than incorporating several groups of acoustic models for different accents. The proposed study can facilitate the analysis or application of a new speech corpus containing accented utterances.

7. Discussion & summary

This paper proposes a hybrid refinement scheme for explicit phonetic segmentation. This hybrid scheme combines statistical correction, multi-resolution fusion, and predictive models for improving segmentation results. To correct phonetic boundaries statistically, both CI and CD models are used for experiments. A state selection method is also used to improve the statistical correction and leads to increased segmentation accuracy. Experiments were also conducted to study the effects of time resolutions of acoustic models and showed that the variation of this property can affect segmentation results with a specific pattern. The proposed multi-resolution fusion embraces the benefits of HMMs with various resolutions, which were not studied in the past. Predictive models were also applied with a different labeling scheme and a small stepsize to further refine the detected boundaries. Each of these steps can contribute to the segmentation results, detecting phone boundaries in a more precise manner. This refinement scheme does not require retraining of acoustic models and can thus be applied to any existing segmentation system to post-process the obtained phone boundaries. It is also not constrained to a specific segmentation model and can still gain benefits from advanced modeling techniques of segmentation systems in future.

In addition, this paper studies a cross-corpora segmentation scenario which is not examined before. Automatic phonetic segmentation is performed on a new corpus with a different accent using the segmentation system trained on a standard corpus as well as the proposed refinement scheme. Experimental results demonstrate that the differences of the characteristics and accents between corpora can affect the segmentation accuracy. However, the proposed refinement scheme can be applied, by using a small set of labeled data from the new corpus, to improve this cross-corpora segmentation so that the results are comparable to those generated by acoustic models trained on large volumes of labeled data from the new corpus.
In future studies, more advanced predictive models may be designed to further improve the refinement process. In addition, extensive experiments on the cross-corpora segmentation task can be performed by collecting accented English data, specifically accented English from Asian speakers, to construct a corpus whose characteristics are more distinct from a standard English corpus like TIMIT.

References


