Phone Boundary Detection Using Sample-Based Acoustic Parameters

You-Yu Lin¹, Yih-Ru Wang¹ and Yuan-Fu Liao²

¹ Institute of Communication, National Chiao Tung University, Hsinchu, Taiwan, ROC
² Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan, ROC
rossi0927.cm97g@g2.nctu.edu.tw, yrwang@mail.nctu.edu.tw, yfliao@ntut.edu.tw

Abstract
A sample-based phone boundary detection algorithm is proposed in this paper. Some sample-based acoustic parameters are first extracted in the proposed method, including six sub-band signal envelopes, sample-based KL distance and spectral entropy. Then, the sample-based KL distance is used for boundary candidates pre-selection. Lastly, a supervised neural network is employed for final boundary detection. Experimental results using the TIMIT speech corpus showed that EERs of 13.2% and 15.1% were achieved for the training and test data sets, respectively. Moreover, 43.5% and 88.2% of boundaries detected were within 80- and 240-sample error tolerance from manual labeling results at the EER operating point.

Index Terms: Speech segmentation, speech analysis.

1. Introduction
Automatic phonetic segmentation is a historic and basic problem in speech signal processing. Although a lot of researches had been done in the past [1], an automatic phonetic segmentation algorithm with high accuracy and precision is still a state-of-the-art work. Without knowing the text of the speech signal, it becomes a phone boundary detection problem which is more difficult than the phone boundary alignment problem. An accurate phone boundary detector is important and essential for speech processing engineering and linguistics.

The most popular approach to automatic phonetic segmentation, when knowing the text of the speech signal, is the HMM-based time-alignment method. The performance is usually evaluated by using the percentage of detected boundaries with errors smaller than 20ms as a figure of merit. In [2], the minimum boundary error (MBE) criterion was used in the training of the HMM method. For TIMIT acoustic-phonetic continuous speech corpus, the minimum MBE-trained HMMs can identify 79.75% of human-labeled phone boundaries within a tolerance of 10ms, compared to 71.23% achieved by the conventional ML-trained HMMs. Moreover, there are only 7.89% of automatically labeled phone boundaries having errors larger than 20ms. Besides, some other techniques, like SVM, fuzzy logic and neural network [3], were also proposed to refine the HMM aligned boundaries in order to get more accurate boundary positions.

2. The proposed sample-based phone boundary detection algorithm
In the automatic boundary detection problem, without knowing the text of the speech signal, the rate of acoustic signal change is the most important cue for decision making. In [4], the spectral transition measure, which is in fact the norm of delta MFCC, was used to find the phone boundaries. 15.4% miss detection (MD) and 22.0% false alarm (FA) rates were achieved for the TIMIT training data set. In [5], the model selection technique, DISTBIC, was used to perform the phone boundary detection. The DISTBIC first used the Kullback-Leibler (KL) distance to find the boundary candidates, and then employed the Bayesian information criterion (BIC) to further verify these candidates. 25.7% MD and 23.3% FA rates were achieved for the NTIMIT database.

In those past studies, they adopted the frame-based approach to use acoustic features like MFCCs in phone boundary detection. The main drawback of the approach lies in the incapability of the frame-based features to model the rapid spectral changes in speech signal. Besides, the time resolution of the frame-based approach is too coarse for phone boundary detection. This motivates us to propose a sample-based phone boundary detection algorithm in this paper.

The paper is organized as follows. Section 2 presents the proposed sample-based phone boundary detection algorithm. Its performance is examined by simulations discussed in Section 3. Some conclusions are given in the last section.
2.1. Sample-based acoustic parameters for phone boundary detection

It is known that the spectrum of a speech signal is an effective cue for phone boundary detection. In this study, six sub-band signal envelopes are used. Fig.1 displays the schematic diagram of these envelope detectors. The input speech signal firstly passes through six band-pass filters with cutoff frequencies shown below:

- 0.0 – 0.4 KHz
- 0.8 – 1.5 KHz
- 1.2 – 2.0 KHz
- 2.0 – 3.5 KHz
- 3.5 – 5.0 KHz
- 5.0 – 8.0 KHz

The energies of the above six sub-band signals were shown to be effective in speech landmark detection [6]. In the sample-based approach, the envelopes of those sub-band speech signals are extracted instead of their energies. Each sub-band signal envelope detector, shown in Fig. 1, is realized by first generating a Hilbert transformed signal from the input speech signal and then passing it through a low-pass filter. The Hilbert transformed signals can be produced by:

\[ y_i[n] = x_i[n] \otimes h[n] \quad \text{for } i = 1, \ldots, 6 \]  

where

\[ h[n] = \begin{cases} 0, & n \text{ is even and } n < 0, n \geq N \\ \frac{1}{n\pi}, & n \text{ is odd} \end{cases} \]

The envelope of the i-th sub-band signal is denoted by \( e_i[n] \). Beside, the envelope of the original speech signal, \( e_0[n] \), is also extracted. The cutoff frequency of the low-pass filter is set to 30 Hz.

Spectral entropy is commonly used in measuring the flatness of a speech power spectrum in a frame-based system [7]. In this study, it is extended to the sample-base spectral entropy defined by:

\[ H[n] = \sum_{i=1}^{6} \left( E_i[n] \right) \log \left( E_i[n] \right) \]  

where \( n \) is the sample index. Its value will be small in the fricative/affricate and nasal parts of speech signal.

The normalized sub-band signal envelope, sample-based KL distance and spectral entropy and their delta terms are effective parameters for modeling the short-term spectral changing rate. They are used as the input features of the phone boundary detector.

2.2. Sample-based boundary detection framework using MLP neural network

A boundary candidate pre-selection procedure is first used to reduce the number of data needed to be processed in the boundary detector. The selected boundary candidates are those samples having larger speech signal changing rate. Thus, the sample-based KL distance is employed for boundary candidates pre-selection. A simple peak picking method with threshold is used to select all samples which satisfy the following constrains as candidates:

\[ d_{k}[n] > d_{k}[n-1], \quad d_{k}[n] > d_{k}[n+1], \quad \text{and} \quad d_{k}[n] \geq Th_d \]  

where \( Th_d \) is a threshold. The sequence of boundary candidates is denoted as \( \{c_j; \ j = 1 \ldots, N_c\} \). For the TIMIT database, the miss rate of the pre-selection procedure is 2.9%.

Since the speech signal is segmented into pieces by the candidates found by the pre-selection procedure, segmental parameters can also be used as the input features of the boundary detector. The average normalized sub-band envelope of the stable parts of adjacent segments, \( e_{k\rightarrow k+1} \) and \( e_{c_i \rightarrow c_{i+1}} \), as shown in Fig. 3, can be calculated by:

\[ ES_{k\rightarrow k+1} = \left( \frac{\sum_{n=0}^{N_c} E_k[n]}{c_{k+1} - c_k} \right) / \left( \frac{c_{k+1} - c_k}{c_{k+1} - c_k} \right) \]  

Then, a 38 dimensional feature vector is constructed for each boundary candidate. For the k-th candidate, at time \( c_k \), its feature vector includes the following acoustic parameters:

1. Features from the current boundary candidate:
   \( d_{k}[c_k], H[c_k], \Delta H[c_k] \); \( j = k - 1, k, k + 1, \)
   \( E_k[c_k], \Delta E_k[c_k] \); \( i = 1, \ldots, 6 \), \( \Delta E_k[c_k] \)

   where, \( \Delta H[c_k] \) and \( \Delta E_k[c_k] \) are the delta terms of spectral entropy and the i-th sub-band normalized envelope.

2. Features from the adjacent segments of the current candidate:
(3) Two indicators used to indicate the first and last candidates in an utterance.

Unlike the previous studies using a simple threshold value for acoustic parameter [4] or using a model selection criterion [5], a supervised multi-layer perceptron (MLP) is employed for the phone boundary detection in the proposed system. A problem to be solved is how to decide the phone boundary targets for training the MLP classifier. Conventionally, phone boundary targets are manually labeled. But, those manually labeled boundaries may not be consistent with the boundaries found from an objective measure such as the sample-based KL distance used in the pre-selection procedure.

An iterative procedure that integrates target selection and MLP training is proposed in this study for the phone boundary detector. For each manually labeled phone boundary, \( m_i \), a target for MLP detector is chosen from one of the candidates in the interval \([m_i - \delta, m_i + \delta]\), where \( \delta \) denotes the allowable tolerance between the manually and automatically detected phone boundaries. The procedures of the training algorithm is listed in detail in the following:

1. Select initial targets from the candidates with the largest sample-based KL distance \( d_{KL}(c_i) \) where \( c_i \in [m_i - \delta, m_i + \delta] \)
2. Train the MLP-based phone boundary detector using the given targets;
3. Select the candidate with largest output of the MLP detector, i.e. the most probable candidate in \([m_i - \delta, m_i + \delta]\), as the new target.
Repeat (2) and (3) until convergence.

**3. Experiments and results**

The TIMIT speech corpus was used to evaluate the effectiveness of the proposed sample-based phone boundary detection algorithm. The numbers of phone boundaries in the training and testing parts of TIMIT were 172460 and 62465, respectively. The total numbers of samples were \( 2.27 \times 10^8 \) and \( 8.29 \times 10^7 \) for training and testing data sets, respectively. Compared with the performance of [4], which is 15.4% MD and 22% FA rates for the TIMIT training data set, about 20% EER reduction was achieved. Besides, it is worthy to note that [4] is a frame-based approach. Since it can't distinguish the boundaries within the same frames, the actual FA of [4] will be a little higher.

A threshold for detector output was then used to decide the phone boundaries. The curves of MD (Miss Detection) rate vs. FA (False Alarm) rate for the training and test data sets were shown in Fig. 4. The definition given in [5], FA was defined as (number of false alarms)/(number of actual boundaries + number of false alarms). 13.2% and 15.1% EERs were achieved for the training and test data sets, respectively. Compared with the performance of [4], which is 15.4% MD and 22% FA rates for the TIMIT training data set, about 20% EER reduction was achieved. Besides, it is worthy to note that [4] is a frame-based approach. Since it can't distinguish the boundaries within the same frames, the actual FA of [4] will be a little higher.

The normalized cumulative histogram of the absolute deviation between the automatically detected boundaries and the manually labeled ones was shown in Fig. 5. As shown in the figure, 43.5% and 88.2% of boundaries detected were within 80- and 240-sample error tolerance from the manual labeling results at the EER operating point. While in [4], only 27% boundaries detected were within the same frame of the manually labeled boundaries and 70% were within one-frame (10-ms) tolerance. In our method, after converting the answers into frame-based resolution, 42% boundaries detected were within the same frame of the manually labeled boundaries and 87% were within one-frame tolerance. This shows that the proposed sample-based boundary detection method performed better.

A preliminary error analysis was done. First, some phone boundaries were essentially difficult to detect because the acoustic signals surrounding those boundaries are changing slowly. At the EER operating point, 28.6% of the boundaries between \{vowels, semivowels, glides\} and \{vowels, semivowels, glides\} are not detected, i.e., misses. Basically, the filter band [6] used in this paper was designed for pronunciation manner and landmark detection. The MD rates for other boundaries between phones with the same pronunciation manner were also higher than average. The MD rate for fricative-fricative boundaries was 30.5%; and 51.8% for nasal-nasal boundaries. Finally, lots of short silences in front of fricative and semivowel, usually less than 5 ms, were not labeled in TIMIT, but those boundaries were correctly detected in the proposed method.
4. Conclusions

In this paper, a sample-based phone boundary detection algorithm was discussed. The effectiveness of the proposed algorithm was checked on the TIMIT speech corpus. The experimental results show that both the accuracy and precision were better than the previous studies using the frame-based approach. 13.2% and 15.1% EERs were achieved for the training and test data sets, respectively. 43.5% and 88.2% of boundaries detected were within 80- and 240-sample error tolerance from manual labeling results at the EER operating point.

Fig. 5. The normalized cumulative histogram of the absolute deviation between automatically detected boundaries and manually labeled result.

5. Acknowledgement

This work was supported by the National Science Council, Taiwan, ROC, under the project with contract NSC 97-2221-E-009-080-MY3.

6. References