Automatic Phonetic Segmentation by Using a SPM-based Approach for a Mandarin Singing Voice Corpus

Cheng-Yuan Lin and J.-S. Roger Jang.

Department of Computer Science
National Tsing Hua University, Taiwan
{gavins, jang}@wayne.cs.nthu.edu.tw

10.21437/Interspeech.2006-589

Abstract

This paper proposes a score predictive model (SPM) based approach to integrate two segmentation results obtained by HMM and DTW for a Mandarin singing voice corpus. The SPM can predict the score of a boundary according to its corresponding 14 dimensional feature vector. In order to verify the performance of the proposed method, several experiments were performed. The experimental results demonstrate the feasibility of the proposed approach.

Index Terms: automatic phonetic segmentation, boundary refinement, score predictive model

1. Introduction

Corpus-based speech synthesis systems are becoming increasingly popular due to the high degree of fluency and the natural feel of the generated speech. Recently, the corpus-based approach was also applied for the synthesis of the singing voice [3][9]. However, these systems require a large amount of human effort to label the phonetic boundaries of the corresponding corpus. As a result, it has become quite important to design an efficient approach for automatic phonetic segmentation especially when the size of the corpus is very large. There are many studies concerning automatic segmentation to be found in the literature [2][4][7][8]. Generally, these methods involve two steps: first perform a rough phonetic segmentation by forced alignment of the Viterbi search using a hidden Markov model (HMM), and then apply a boundary correcting postprocessor to refine the results obtained by the HMM. Consequently, it should be feasible to perform segmentation of singing voice corpora by employing the same scheme. Unfortunately, the initial HMM-based segmentation of singing voice corpora does not perform as well as that of speech corpora. This is probably caused by several aspects of the physical differences between the singing voice and speech. For example, the pitch range variation of a singing voice is much wider than that of speech; the average singing rate is generally slower but has a higher variance. Furthermore, there is no HMM-based recognizer specialized for the singing voice available. If the initial segmentation results are not reliable enough, then the corresponding postprocessor will refine the boundaries inefficiently. Park et al. [4] also concluded that it is very difficult to cope with the problem of large labeling errors by using a boundary refinement postprocessor. Therefore, there is a need to improve the performance of the HMM-based recognizer. However, this is quite difficult, even if the HMMs were obtained by employing an embedded-reestimation procedure as observed in [7].

In view of the above, an alternative method is to use melody information (notes and tempos). This is because the recording artist is required to sing a song by following the corresponding melody information when the singing voice corpus is being collected. Therefore, it seems feasible to use DTW (dynamic time warping) instead of HMM for performing segmentation according to the corresponding melody information. In our previous study concerning the automatic singing voice rectifier [1], DTW was used to perform the segmentation tasks and its performance was found to be acceptable. In other words, other segmentation results can be obtained by using DTW. Nevertheless, there is no guarantee that the performance of DTW will be better than that of HMM, or vice versa. Consequently, the aim of this work is to develop a practical approach to coordinate the efforts of DTW and HMM. In this paper, a score predictive model (SPM) based approach has been proposed which can predict the score of a boundary according to a set of essential acoustic features. Subsequently, several experiments have been carried out to verify the performance of the proposed method.

The remainder of this paper is organized as follows. Section 2 explains the SPM-based approach that can integrate the two results obtained by DTW and HMM. In Section 3 we present the experimental results and analysis. Finally, we draw our conclusions in Section 4 and indicate potential future work.

2. A SPM-based approach

In this section, we propose a SPM-based approach that involves several essential procedures. The details are discussed in the following subsections.

2.1 The phonetic category transitions in Mandarin

Since the synthesis units are syllable-based in most Mandarin TTS systems or Mandarin singing voice systems, the primary task of phonetic segmentation is focused on how to precisely refine the boundaries between two consecutive syllables. There are 22 distinct consonants and 38 distinct vowels in Mandarin Chinese. That is, in theory there are a total of 836 (38 x 22) possible phoneme combinations for a boundary. In order to avoid the influence of insufficient data coverage, six primary types for consonants and nine primary types for vowels are classified in advance according to their acoustic characteristics. This suggests the use of a reduced set composed of 54 (9 x 6) possible phonetic category transitions for each boundary. Therefore, a total of 54 corresponding models for these transitions shall be constructed, which is addressed in the following subsections. Table I lists the types of consonants and vowels (using Hanyu Pinyin) used in this paper.

Table I. (a). Six types of consonants

1	m, n, l, r, "null"	2	h, x, sh	3	b, d, g
4	j, zh, z	5	p, t, k	6	q, ch, c, f, s

Table I. (b). Nine types of vowels

1	"null"	2	a, ya, wa	3	o, wo
4	e, er	5	ê , ye, yue	6	ai, iai, wai, yi, ei, wei
7	ao, yao, wu, yu, ou, you	8	an, yan, wan, yuan, en, yin, wen, yun	9	ang, yang, wang, eng, ying, weng, yong

2.2 The score function definition

Generally speaking, the positions of two labeled boundaries are not necessarily close to each other even though their corresponding features might be similar. Here, the corresponding features are extracted from a frame (ex. a rectangular window for each labeled boundary, shown in Fig. 1) located around this labeled boundary; they can be zero-crossing rate, energy, pitch, and other popular acoustic features.

For example, in Fig. 1, there are three distinct boundaries labeled by human, DTW, and HMM, respectively. The manually labeled boundary is closer to the HMM labeled boundary even though its corresponding acoustic features (ex. zero-crossing rate and pitch) are similar to those of the DTW labeled boundary rather than the HMM labeled boundary.

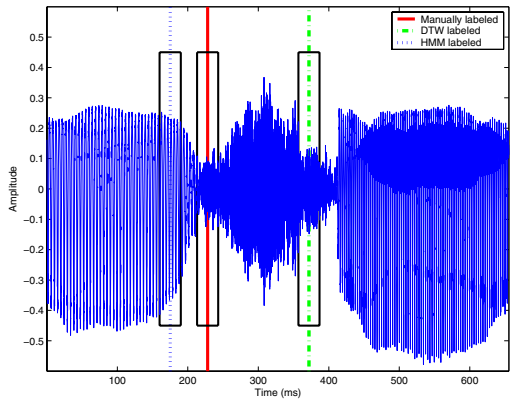

Fig. 1 Three boundaries with distinct locations.

On the basis we adopted a score function which sets higher scores for those boundaries with smaller ranges around a true (manually labeled) boundary. We used this setup simply because it can easily distinguish between two kinds of boundaries, one close to a true boundary and the other not. In this paper, the score function is defined as equation (1) which is modified from the Gaussian equation.

$$Score(D) = K \times \left((2\pi)^{-1/2} \times \sigma^{-1} \times \exp\left(\left[-\frac{D^2}{2 \times \sigma^2} \right] \right) \right), \quad (1)$$

where D denotes the distance between a true boundary and another candidate boundary. The unit of D is ms. In addition, K is a constant used to adjust the scale of the score function. In this paper, K is set at 7018.559 and σ is set at 28 respectively. Such setups generates a curve with scores ranging from 0~100. Fig. 2 shows its corresponding curve of the score function when input D ranges from -200 ms to 200 ms.

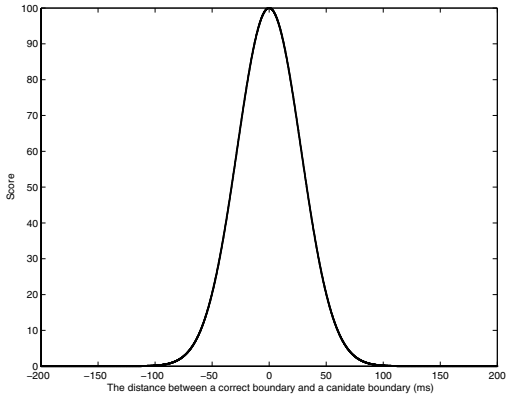

Fig. 2 The proposed score function.

2.3 The construction of a score-predictive model

Once the score function has been defined, a score predictive model (SPM) is subsequently constructed for each phonetic transition category. This construction involves three essential phases, including the collection of training data, a set of useful acoustic features, and a Neural Network (NN) based regression model. The details of these phases are as follows.

2.3.1 Defining candidate boundaries for training data

As noted in Section 2.1, there are 54 possible phonetic transition categories for a phoneme boundary. Next, we collect the boundaries according to the corresponding phonetic transition category. In addition, given a true (manually labeled) boundary, the candidate boundaries located in the nearby area of this boundary are collected. Based on our observation, the segmentation accuracy (range error < 200 ms) is about 97% regardless of using DTW or HMM for the singing voice corpus. Therefore, these candidate boundaries located within ± 200 ms of a true boundary are collected. In this study, these candidate boundaries are collected via the following rules:

1). Add a set of candidate boundaries, 5 ms apart, located within ± 50 ms of a true boundary.
2). Add a set of candidate boundaries, 10 ms apart, located within two intervals (50 ~ 100 ms and -50 ~ -100 ms) around a true boundary.
3). Add a set of candidate boundaries, 20 ms apart, located within two intervals (100 ~ 200 ms and -100 ~ -200 ms) around a true boundary.

Finally, a total of 41 candidate boundaries (including a true boundary) can be collected for each true boundary as shown in Fig. 3 according to the three rules mentioned above.

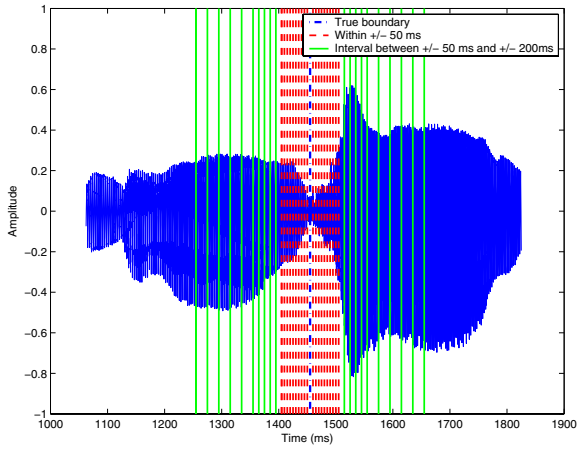

Fig. 3 A typical example of all candidate boundaries for a true boundary.

2.3.2 *Defining the SPM-feature for each candidate boundary*

Once the locations of these candidate boundaries are defined, we take seven popular acoustic features, including zero-crossing rate, log energy, entropy [10], bisector frequency [2], pitch and line spectrum pairs (LSFs), Mel-scale frequency cepstral coefficients (MFCCs), as a basic feature set. All except 2 features are scalars, LSFs and MFCCs, which are both a 12-dimensional vector. However, these features are not taken directly as training features used for generating a SPM. In practice, for each candidate boundary, the differences between all acoustic features of its left and right frames are evaluated. (The size of a frame is set to 20 ms). In addition, these features are then normalized to the range [0, 1]. Consequently, there is a seven dimensional feature vector representing each candidate boundary. In addition, it is better if the affect of the neighboring frames can be considered simultaneously. Thus, another potential feature vector is computed via a delta function which estimates the rate of change of the original feature. Equation (2) shows the delta function:

$$\Delta F_{\dim}(t) = \frac{\sum_{\tau=-M}^{M} F_{\dim}(t+\tau)\tau}{\sum_{\tau=-M}^{M} \tau^2}, \dim = 1,2,...,7 \qquad (2)$$

where F indicates the original seven dimensional feature vector, M is set as 2, and t denotes a candidate boundary index. Thus, for each boundary t, it still needs to extract other features of the four neighboring boundaries in addition to its original feature F. In this paper, the four neighboring boundaries, 10 ms apart, located within ± 20 ms of this boundary t, and the feature extraction of the four boundaries are the same with that of the original feature F. As a result, delta F can be derived via equation (2) according to F and the features of these neighboring boundaries. Finally, the SPM-feature is a 14 dimensional feature vector when we combine F and its delta F.

2.3.3 *Constructing SPMs by using a NN-based regression model*

In this section, we employ a NN (Neural Network) based regression model to construct a SPM for each kind of phonetic category transition. The input of the NN-based regression model is the SPM-feature while its corresponding output is score obtained by the proposed score function. In this study, we applied Levenberg-Marquardt back propagation [5] as the primary training function. Two hidden layers with 30 neurons each were adopted and the corresponding transfer function of the hidden layer was the hyperbolic tangent function, but the linear function was used in the output layer.

Since we employed a NN-based method to construct SPMs, we are faced with the possibility of a local minimum point. That is to say, it cannot guarantee that the obtained SPMs are always reliable enough. To cope with this problem, for each phonetic transition category, it costs ten iterative training cycles to obtain the optimum SPM which has smallest root mean squared error (RMSE). Fig. 4 shows the construction of SPMs for the 54 phonetic category transitions.

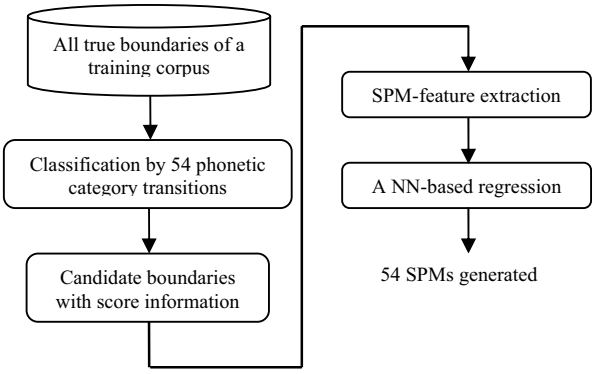

Fig. 4 The construction of SPMs.

2.4 Using the SPM-based approach to integrate the results obtained by DTW and HMM

In this study, two feasible phonetic segmentation methods are performed by using DTW and HMM, respectively. The descriptions concerning automatic segmentation by means of DTW can be found in [1]. For HMM, the TCC-300 corpus [6] was used to train context-dependent triphone models at the beginning, then an embedded-reestimation was employed to implement the HMM-based alignment. The overall segmentation procedure is described as follows.

1. Since two recognizers were performed initially, there are two initial estimates for each phoneme boundary between two syllables.
2. Each phoneme boundary is classified according to its phonetic transition category. Next, we use its corresponding SPM to predict the scores of two initial boundaries. Only the boundary with the higher score is preserved and the lower score is discarded.
3. A dynamic search area for refinement is determined according to the score of the preserved boundary. The size of the search area is set empirically according to the following rules:

$$Search\ \ area = \begin{cases} 0\ \ ms, & when\ \ score \geq 90 \\ 20\ \ ms, & when\ \ 60 \leq score < 90 \\ 30\ \ ms, & when\ \ 20 \leq score < 60 \\ 40\ \ ms, & when\ \ score < 20 \end{cases}$$

4. Finally, for the preserved boundary, we select candidate boundaries as the test set, which are 2 ms apart, and within a search area at both sides of this boundary. Then we use the corresponding SPM to obtain scores of these boundaries, and the highest score boundary is the final boundary.

3. Experimental Results and Analysis

The singing voice corpus used in our experiments was recorded from the voice of a professional female recording artist. The corpus is composed of a total of 1384 non-uniform length utterances which correspond to 9561 syllables in total. The related design and collection of the singing voice corpus was elaborated in [3]. The entire corpus was divided into 800 utterances for training and 584 utterances for the test. The boundaries of these utterances were labeled in advance by an expert. The training utterances were used to construct SPMs for 54 phonetic transition categories. The test utterances were then used to verify the feasibility of the proposed method.

As noted in Section 1, most of the boundary refinement methods are not very efficient in refining the boundaries if initially there are larger segmental errors caused by HMM. This is due to the fact that these large segmental errors caused by HMM can result in serious outlier problems [7]. In order to confirm this phenomenon further, a boundary refinement based on a hybrid approach proposed in our previous study [2] was compared in the following experiments. This approach was verified to be able to perform well on the phonetic segmentation of speech data. In addition, this approach refines the HMM-based segmental results within a fixed search area (± 40 ms) instead of the dynamic one used in the proposed SPM-based approach.

As mentioned above, a total of four kinds of schemes were used in the experiments including an inside test and an outside test, they are HMM, DTW, a hybrid approach, and the proposed SPM approach. Fig. 5 demonstrates the performance of the inside test (800 training utterances) and the outside test (584 test utterances) for these schemes.

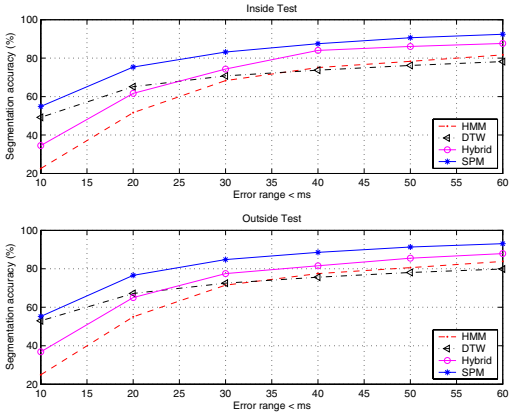

Fig. 5 The comparison of four schemes for automatic segmentation. Top: inside test. Bottom: outside test.

From Fig. 5, it is evident that both the performance of DTW and HMM are not accurate enough regardless if it is an inside test or an outside test. Although the previous hybrid approach certainly improves the performance as compared with that of DTW or HMM, its segmentation accuracy is lower than that of the proposed SPM-based method. In other words, the proposed SPM-based method achieves the desirable performance, as anticipated. This also indicates that if we were able to integrate two results (DTW and HMM) more perfectly, it would be quite useful for the automatic segmentation task of a singing voice corpus.

4. Conclusions

This paper proposed a SPM-based approach to be used to refine the segmentation results obtained by DTW and HMM. In order to verify its feasibility, a boundary refinement based on a hybrid approach proposed in our previous study was used for comparison. The experimental results indicated that our proposed method has a better performance than that of other approaches.

In practice, it will be quite possible to improve the performance of the SPM-based method. This is due to the fact that some SPMs did not work well due to the co-articulation problems. For example, for "vowel + nasal" phonetic transition, there is always a stronger co-articulation effect between two syllables. In this paper, we did not address any special handling of these cases. In future work we will try other rule-based or statistics-based methods to cope with the problem of stronger co-articulation. In addition, more influential acoustic features for score prediction will be attempted as well.

5. References

[1] Cheng-Yuan Lin, Jyh-Shing Roger Jang and Mao-Yuan Hsu, "An Automatic Singing Voice Rectifier," *ACM Multimedia Conference*, 2003.

[2] Cheng-Yuan Lin, Kuan-Ting Chen, J.-S. Roger Jang, "A Hybrid Approach to Automatic Segmentation and Labeling for Mandarin Chinese Speech Corpus," *Eurospeech*, 2005.

[3] Cheng-Yuan Lin, Tzu-Ying Lin and J.-S. Roger Jang, "A Corpus-based Singing Voice Synthesis System for Mandarin Chinese," *ACM Multimedia Conference*, 2005.

[4] E.-Y. Park, S.-H. Kim, and J.-H. Chung, "Automatic speech synthesis unit generation with MLP based postprocessor against auto-segmented phoneme errors," *in Proc. Int. Joint Conf. Neural Networks*, pp. 2985–2990, 1999.

[5] Gill, P. R, Murray, W.; and Wright, M. H. "The Levenberg-Marquardt Method," §4.7.3 in *Practical Optimization. London: Academic Press*, pp. 136-137, 1981.

[6] http://rocling.iis.sinica.edu.tw/ROCLING/MAT/Tcc_300brief.htm

[7] Ki-Seung Lee, "MLP-Based Phone Boundary Refining for a TTS Database," *IEEE Trans. on speech and audio processing*, 2005.

[8] LiJuan Wang, Yong Zhao, Min Chu, Jianlai Zhou and Zhigang Cao, "Refining Segmental Boundaries for TTS database Using Fine Contextual-Dependent Boundary Models," *ICASSP*, 2004.

[9] Meron Y., "High quality singing synthesis using the selection-base synthesis scheme," *PhD dissertation*, 1999.

[10] Shen, J.-L., et al., "Robust entropy-based endpoint detection for speech recognition in noisy environments," *ICSLP*, 1998.