UNSUPERVISED PROSODIC PHRASE BOUNDARY LABELING OF MANDARIN SPEECH SYNTHESIS DATABASE USING CONTEXT-DEPENDENT HMM

Chen-Yu Yang, Zhen-Hua Ling, Li-Rong Dai

National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China
yangcy@mail.ustc.edu.cn, zhling@ustc.edu.cn, lrdai@ustc.edu.cn

ABSTRACT

In this paper, an automatic and unsupervised method based on context-dependent hidden Markov model (CD-HMM) is proposed for labeling the phrase boundary positions of a Mandarin speech synthesis database. The initial phrase boundary labels are predicted by clustering the durations of the pauses between every two prosodic words in an unsupervised way. Then, the CD-HMMs for the spectrum, F0 and phone duration are estimated by a means similar to the HMM-based parametric speech synthesis using the initial phrase boundary labels. These labels are further updated by Viterbi decoding under the maximum likelihood criterion given the acoustic feature sequences and the trained CD-HMMs. The model training and Viterbi decoding procedures are conducted iteratively until convergence. Experimental results on a Mandarin speech synthesis database show that this method is able to label the phrase boundary positions much more accurately than the text-analysis-based method without requiring any manually labeled training data. The unit selection speech synthesis system constructed using the phrase boundary labels generated by our proposed method achieves similar performance to that using the manual labels.

Index Terms— speech synthesis, phrase boundary, unsupervised labeling, context-dependent hidden Markov model, Viterbi decoding

1. INTRODUCTION

A speech database with corresponding label information is the precondition for constructing a speech synthesis system. A large-sized and precisely labeled speech database can help improve the naturalness and intelligibility of the constructed speech synthesis system, especially for the unit selection and waveform concatenation synthesis method. Speech database annotation commonly consists of phonetic segmentation and prosodic labeling. In terms of phonetic segmentation, the text-analysis-based phoneme transcription and the HMM-based segmentation techniques have already achieved good performance [1, 2] and are widely adopted in practical systems. In this paper we focus on the task of prosodic labeling. The definition of the prosodic labels varies with language. For Mandarin speech synthesis, the prosodic labels refer to the prosodic boundaries. Among different levels of prosodic boundaries, the prosodic phrase boundary tends to be the most difficult one for automatic and manual labeling. In contrast to the prosodic word boundary which can be precisely predicted from the text, the phrase boundary positions are more context-dependent and speaker-dependent. If they are labeled manually, it is very time-consuming and difficult to guarantee consistency among different annotators. Therefore, we investigate methods of automatic phrase boundary labeling for Mandarin speech synthesis in this paper.

Various methods have been proposed to achieve automatic prosodic boundary labeling for speech synthesis databases. Most of them are supervised classification based approaches [3–7], which implies that a certain amount of manually labeled training data is necessary for annotating each database. Several methods which adopt unsupervised approaches can be found in [8–10]. Ananthakrishnan et al. [8] applied clustering algorithms to partition the acoustic space into two classes and initialize the prosodic boundary labels. These labels were used to train a MAP classifier and were updated iteratively. Huang et al. [9] initialized the labels by some lexical and acoustic cues and used these labels to train a GMM-based prosodic break detector. Chiang [10] proposed a joint prosody labeling and modeling method which determined the prosodic labels and built the prosodic models simultaneously. In this approach, the initial labels were determined by a decision tree which was designed based on prior knowledge of prosodic breaks. After the parameters of the prosodic models were estimated, the prosodic labels were updated iteratively.

In this paper, an unsupervised prosodic phrase boundary labeling method is proposed, which uses the context-dependent hidden Markov model (CD-HMM). A similar structure using a supervised approach has previously been shown to be effective by the authors [11, 12]. We now extend the method to suit an unsupervised condition in this paper. The CD-HMMs of the spectrum, F0, and phone duration are firstly trained without the context information of the prosodic phrase boundaries. Then a state alignment to the acoustic features is performed using the trained models to get the pause durations between every two prosodic words. These pause durations are further normalized according to the context-dependent phone duration distributions. The initial phrase boundary positions are labeled by unsupervised clustering of the normalized pause durations. After the initial labels are given, the CD-HMMs are re-estimated and the phrase boundary labels are updated by Viterbi decoding under the maximum likelihood criterion. The model training and the Viterbi decoding procedures are executed iteratively until the labeling results converge. In contrast to the methods in [8,9] where the prosodic boundary type at each prosodic word boundary was determined independently, our proposed method adopts the Viterbi decoding approach to decide the boundary types of all prosodic word boundaries simultaneously. Different from [10], the influence of the known context information on the acoustic features is considered during initialization and extra prosodic models are not necessary in our proposed method.

This paper is organized as follows. In Section 2, the proposed unsupervised prosodic phrase labeling method is introduced. Sec-

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

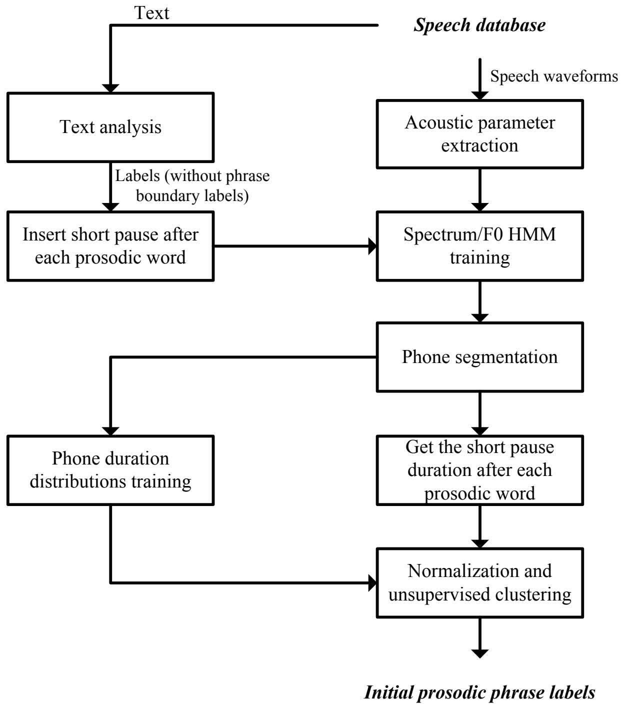
Text
Speech database
Speech waveforms
Text analysis
Acoustic parameter extraction
Labels (without phrase boundary labels)
Insert short pause after each prosodic word
Spectrum/F0 HMM training
Phone segmentation
Phone duration distributions training
Get the short pause duration after each prosodic word
Normalization and unsupervised clustering
Initial prosodic phrase labels

Fig. 1. *Flowchart of the initialization step in our proposed method.*

tion 3 reports the objective and subjective experimental results and Section 4 gives the conclusions.

2. METHODS

2.1. Prosodic labeling of Mandarin speech synthesis database

In Mandarin speech synthesis systems, a three-level structure for describing the prosodic characteristics of an utterance is commonly adopted, which consists of prosodic word, prosodic phrase and sentence levels [13]. For each utterance in the speech database, the boundary positions of these three levels need to be labeled manually or automatically. Among them, the prosodic phrase boundary is the most difficult one for labeling. It is strongly context-dependent and speaker-dependent and cannot be simply predicted from the text like the prosodic word boundary. For manual labeling, it is very time-consuming when a large speech database is used and it is difficult to guarantee the consistency among different annotators. Thus, an automatic and unsupervised prosodic phrase boundary labeling method using CD-HMM is presented in this paper.

2.2. Unsupervised prosodic phrase boundary labeling

Our proposed method contains three main steps, which are initialization, model training, and prosodic labeling. Fig. 1 and Fig. 2 show the flowcharts of these steps, in which the model training and prosodic labeling steps are conducted iteratively.

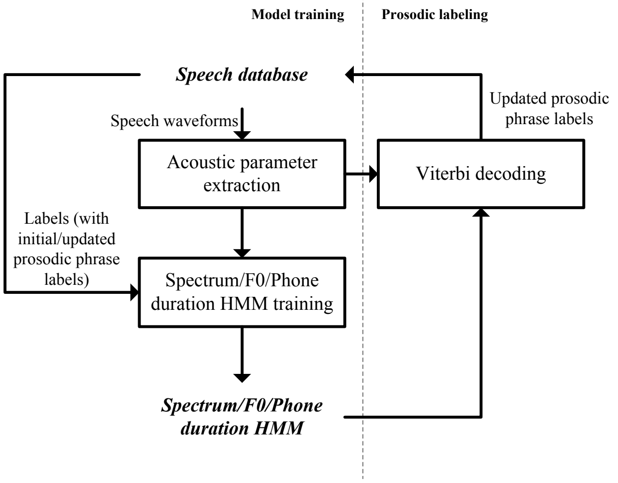
Model training
Prosodic labeling
Speech database
Updated prosodic phrase labels
Speech waveforms
Acoustic parameter extraction
Viterbi decoding
Labels (with initial/updated prosodic phrase labels)
Spectrum/F0/Phone duration HMM training
Spectrum/F0/Phone duration HMM

Fig. 2. *Flowchart of the model training and prosodic labeling steps in our proposed method.*

2.2.1. Initialization

In the initialization step, the initial labels of the prosodic phrase boundaries are obtained without human intervention in preparation for the following iterative model training and prosodic labeling. Once the prosodic word and sentence boundaries are given, this becomes an unsupervised two-class classification problem. It is judging whether each prosodic word boundary should be a prosodic phrase boundary or not. The pause durations at the prosodic word boundaries are extracted as the features for classification in our method because of the phonetic knowledge that the prosodic phrase boundary tends to span much longer pauses than the prosodic word boundary.

As shown in Fig. 1, a text analysis module is adopted to determine the phonetic and prosodic labels excluding the phrase boundary positions for each utterance in the speech database. In order to extract the pause duration at each prosodic word boundary, a symbol “sp” is inserted at the end of the phoneme transcriptions of each prosodic word. The CD-HMMs of the spectrum, F0 and phone durations are trained without using the context information about the prosodic phrase boundaries. The duration of “sp” at the end of each prosodic word is obtained by performing a state alignment to the acoustic features using the training models. Considering that other context information besides the prosodic boundary type may also affect the duration of these short pauses, a normalization is applied, according to the trained context-dependent phone duration distributions.

$$\hat{d}_{sp} = \frac{d_{sp} - \mu}{\sigma}, \tag{1}$$

where d_{sp} and $\hat{d}_{sp}$ are the pause durations before and after normalization respectively; μ and σ stand for the mean and standard deviation of the corresponding duration distribution given the context information of the pause. Then a k-medians clustering algorithm is applied to partition the prosodic word boundaries into two classes according to the pause durations. The boundaries belonging to the class with longer pauses are initialized as the prosodic phase boundaries, while the other ones are kept as the prosodic word boundaries.

2.2.2. *Model training*

The model training procedure here is similar to that for HMM-based parametric speech synthesis [14]. Firstly, acoustic features are extracted from the speech waveforms. The feature vector for each frame consists of static, delta and delta-delta components of spectral parameters and F0. The context-dependent HMMs are estimated under the maximum likelihood criterion according to the extracted acoustic features and the context information derived from the database labels. The spectrum part is modeled by a continuous probability distribution and the F0 part is modeled by a multi-space probability distribution (MSD) [15]. A decision tree based model clustering method using the minimum description length (MDL) criterion is applied in the context-dependent HMM training to avoid the data-sparsity problem. Then each utterance in the training database is segmented into states by Viterbi alignment using the trained acoustic HMMs. Based on the results of state segmentation, context-dependent phone duration models are estimated using the same decision-tree-based model clustering technique. Table 1 lists the context features used in the model training. Compared to HMM-based parametric speech synthesis, the number of the context features is reduced here in order to control the complexity of the following Viterbi decoding step [12].

Category	Context features
Phone Groups	{current, next} phone
Tone Groups	the tone of {previous, current, next} syllable
Boundary Groups	the prosodic boundary type at current syllable

Table 1. The context features used in the CD-HMM training.

2.2.3. *Prosodic labeling*

The basic idea of the prosodic labeling step is similar to the automatic speech recognition (ASR) problem. It can be expressed as

$$C^* = \arg\max_C P(\boldsymbol{O}|\boldsymbol{\lambda}, C_g, C)P(C), \qquad (2)$$

where $\boldsymbol{O}$ stands for the acoustic features extracted from the speech waveforms of an utterance; $\boldsymbol{\lambda}$ denotes the trained CD-HMMs; C_g represents the known phonetic and prosodic labels and C stands for the prosodic labels that are expected to be labeled, i.e., the prosodic boundaries in this paper. Once C is determined, it can be combined with C_g to generate the context features listed in Table 1 and calculate the output probability $P(\boldsymbol{O}|\boldsymbol{\lambda}, C_g, C)$ of the acoustic features for the CD-HMMs. $P(C)$ denotes a prior distribution of the unknown labels without observing any acoustic features. In this paper, we ignore this prior distribution and (2) can be simplified as

$$C^* = \arg\max_C P(\boldsymbol{O}|\boldsymbol{\lambda}, C_g, C). \qquad (3)$$

The Viterbi decoding algorithm in ASR [16] is adopted here to solve (3). A “word graph” representing all possible prosodic labeling results is firstly constructed for each utterance based on the known phonetic and prosodic labels and the possible values of the unknown labels. Then a two-pass Viterbi decoding strategy is applied. The N-best paths of each utterance are firstly obtained by Viterbi decoding using the CD-HMMs of the spectrum and F0 features. Then, these N hypotheses are rescored using the context-dependent models of the phone duration. After that, the phrase boundary labels of the utterance can be derived from the best path in the “word graph”.

Once all the utterances in the speech database are processed, a new model training procedure is conducted using the updated phrase boundary labels. The model training and Viterbi decoding procedures are conducted iteratively until the phrase boundary labeling results converge.

3. EXPERIMENTS

3.1. Experimental conditions

A Mandarin speech synthesis database containing 13,000 utterances was used in our experiments. The prosodic boundaries of all these utterances were labeled by experienced annotators. Besides the manual labeling results, three sets of phrase boundary labels were generated and compared in our experiments.

- **Text-based labeling**. A C4.5 decision tree based classifier was constructed using Weka tools [17] to determine whether each prosodic word boundary should be a phrase boundary. All of the features used for classification came from the output of text analysis, such as the POS and the number of syllables in a prosodic word. The training set was composed of 20,000 utterance with manual phrase boundary labels.
- **CD-HMM-based supervised labeling**. 1,000 utterances were picked up from the speech database. 900 of them with manual phrase boundary labels were used for the model training as introduced in Section 2.2.2. The remaining 100 utterances were used as a test set for the objective evaluation. Finally, the phrase boundary positions of all utterances in the database were labeled by Viterbi decoding using the trained model.
- **CD-HMM-based unsupervised labeling**. The same 1,000 sentences as the CD-HMM-based supervised labeling were used for the unsupervised model training. After the initialization of phrase boundary labels, the iterative model training with Viterbi decoding was started. Here a two-pass Viterbi decoding strategy was applied. The 40-best paths of each sentence were firstly obtained by Viterbi decoding using the CD-HMMs of the spectrum and F0 features. These 40 hypotheses were rescored using the context-dependent models of the phone duration. Then, the updated phrase boundary labels of the utterance could be derived. After six iterations, the phrase boundary labels and the CD-HMMs converged. The converged CD-HMMs were applied to label all utterances in the database.

In both the supervised and unsupervised labeling, the speech waveforms were sampled at 16kHz. The acoustic parameters were extracted by STRAIGHT [18], including 40-order line spectral pairs (LSP) and F0. A 5-state left-to-right HMM structure was adopted to train the context-dependent models, where a single Gaussian distribution was used for each HMM state.

3.2. Objective evaluation

We performed an objective evaluation among these methods by comparing their labeling results with the manual labels on a test set. F-score was chosen here as the measurement. The test set consisted of 100 sentences, which were not included in the training set of the supervised labeling but were used during the iterative model training of the unsupervised labeling. We consider this to be reasonable because the manual labels are not required for the unsupervised labeling and all the utterances in the database can be used for the iterative model training. The manual labels were determined by the voting results among three annotators. Table 2 lists the F-scores of different methods. From this table, we can see that the initial labels of the CD-HMM-based unsupervised approach is much higher than that of the text-based approach. This indicates the importance of acoustic cues in determining the phrase boundary positions for a speech synthesis database. After iterative model training, the F-score of the unsupervised labeling increases from 62.73% to 72.51%, which is close to the results of the CD-HMM-based supervised labeling. This is a satisfactory result when compared with the consistency among different human annotators. In our experiments, the average F-score among the three annotators is 75.83% on the test set.

Method	F-score(%)
Text-based	49.22
Supervised CD-HMM	76.09
Unsupervised CD-HMM(initial)	62.73
Unsupervised CD-HMM(converged)	72.51

Table 2. The F-scores of phrase boundary labeling on the test set for different methods.

3.3. Subjective evaluation

Four speech synthesis systems were constructed using the manual phrase boundary labels, and the results of the three automatic labeling methods listed in Section 3.1. The HMM-based unit selection speech synthesis approach [19] was followed and all the 13,000 sentences were used for these systems. Twenty sentences, which were not included in the database, were synthesized by the four systems respectively. These sentences were evaluated by 8 listeners. Each listener was required to give a score from 1 (bad) to 5 (good) on the naturalness of each synthetic sentence. The average mean opinion scores (MOS) for these systems were shown in Fig. 3.[1] From Fig. 3, we can see that the quality of the phrase boundary labeling plays an important role in the performance of a Mandarin speech synthesis system using the unit selection approach. The difference of naturalness between the systems using the text-based labeling method and the other three methods is significant, but no significant difference is observed among the systems using those three methods(Tukey's HSD test at $\alpha \leq 0.01$). That is to say, the systems using the prosodic labels given by the CD-HMM-based supervised and unsupervised labeling methods are both comparable to the one constructed using the manual labels.

4. CONCLUSIONS

In this paper, an unsupervised prosodic phrase labeling method has been proposed for Mandarin speech synthesis database. The phrase boundary labels are initialized by unsupervised clustering of the pause duration between every two prosodic words. Then, CD-HMM training and Viterbi decoding are conducted iteratively to refine the acoustic models and the labeling results. The objective evaluation results have shown that this method can achieve satisfactory phrase boundary labeling performance without requiring any manual labels. Also, the unit selection speech synthesis system constructed using the labels given by our proposed method is comparable to the one constructed using manual annotations. To extend this method to other languages and prosodic labels will be the tasks of our future work.

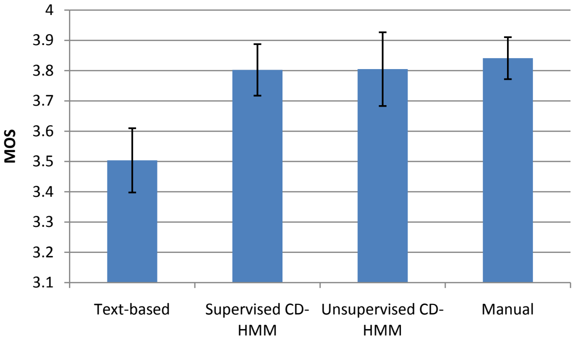

Fig. 3. *Mean opinion scores (MOS) with 95% confidence intervals.*

5. ACKNOWLEDGEMENTS

This work is partially funded by the National Nature Science Foundation of China (Grant No.61273032,61273264) and National 973 program of China (Grant No. 2012CB326405). We would like to thank Prof. Ian McLoughlin for proofreading the paper and making helpful suggestions, we also gratefully acknowledge the assistance of the research division of iFLYTEK Co. Ltd., Hefei, China, for providing the speech database with manual annotations.

6. REFERENCES

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

[2] Y.-J. Wu, H Kawai, J.-F. Ni, and R.-H. Wang, "Minimum segmentation error based discriminative training for speech synthesis application," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, pp. 629–632.

[3] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 469 –481, 1994.

[4] F.-C. Chou, C.-Y. Tseng, and L.-S. Lee, "Automatic segmental and prosodic labeling of Mandarin speech database," in *International Conference on Spoken Language Processing (ICSLP)*, 1998.

[5] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, pp. 509–512.

[6] S. Ananthakrishnan and S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, 2005, pp. 269 –272.

[1] Some examples of the synthetic speech are available at `http://home.ustc.edu.cn/~yangcy/USProsodyLabeling/demo.html`.

[7] F.-Z. Liu, H.-B. Jia, and J.-H. Tao, "A maximum entropy based hierarchical model for automatic prosodic boundary labeling in Mandarin," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2008, pp. 1–4.

[8] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling," in *Proc. Interspeech*, 2006, pp. 829 – 832.

[9] J.-T. Huang, M. Hasegawa-Johnson, and C. Shih, "Unsupervised prosodic break detection in Mandarin speech," in *Speech Prosody*, 2008, pp. 165–168.

[10] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, "Unsupervised joint prosody labeling and modeling for Mandarin speech," *Acoustical Society of America Journal*, vol. 125, pp. 1164–1183, 2009.

[11] C.-Y. Yang, Z.-H. Ling, H. Lu, W. Guo, and L.-R. Dai, "Automatic phrase boundary labeling for Mandarin TTS corpus using context-dependent HMM," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2010, pp. 374 –377.

[12] C.-Y. Yang, L.-X. Zhu, Z.-H. Ling, and L.-R. Dai, "Automatic phrase boundary labeling for a Mandarin TTS corpus using the Viterbi decoding algorithm," *Journal of Tsinghua University*, vol. 51, no. 9, pp. 1276 – 1281, 2011.

[13] A.-J. Li, "Chinese prosody and prosodic labeling of spontaneous speech," in *Speech Prosody*, 2002, pp. 39 – 46.

[14] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC System for Blizzard Challenge 2006," in *Blizzard Challenge Workshop*, 2006.

[15] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, 1999, pp. 229–232.

[16] H. Ney and S. Ortmanns, "Dynamic programming search for continuous speech recognition," *Signal Processing Magazine, IEEE*, vol. 16, no. 5, pp. 64 –83, 1999.

[17] "Weka 3: Data mining software in java," `http://www.cs.waikato.ac.nz/ml/weka/`.

[18] H. Kawahara, I. Masuda-katsuse, and A. D. Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[19] Z.-H. Ling, X.-J. Xia, Y. Song, C.-Y. Yang, L.-H. Chen, and L.-R. Dai, "The USTC System for Blizzard Challenge 2012," in *Blizzard Challenge Workshop*, 2012.