PAPER *Special Section on Advances in Modeling for Real-world Speech Information Processing and its Application*

Unsupervised Prosodic Labeling of Speech Synthesis Databases Using Context-Dependent HMMs*

Chen-Yu YANG[†a)], ***Student Member***, **Zhen-Hua LING**[†b)], ***and*** **Li-Rong DAI**[†c)], ***Nonmembers***

SUMMARY In this paper, an automatic and unsupervised method using context-dependent hidden Markov models (CD-HMMs) is proposed for the prosodic labeling of speech synthesis databases. This method consists of three main steps, i.e., initialization, model training and prosodic labeling. The initial prosodic labels are obtained by unsupervised clustering using the acoustic features designed according to the characteristics of the prosodic descriptor to be labeled. Then, CD-HMMs of the spectral parameters, F0s and phone durations are estimated by a means similar to the HMM-based parametric speech synthesis using the initial prosodic 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 prosodic labeling procedures are conducted iteratively until convergence. The performance of the proposed method is evaluated on Mandarin speech synthesis databases and two prosodic descriptors are investigated, i.e., the prosodic phrase boundary and the emphasis expression. In our implementation, the prosodic phrase boundary labels are initialized by clustering the durations of the pauses between every two consecutive prosodic words, and the emphasis expression labels are initialized by examining the differences between the original and the synthetic F0 trajectories. Experimental results show that the proposed method is able to label the prosodic 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 prosodic phrase boundary labels generated by our proposed method achieves similar performance to that using the manual labels. Furthermore, the unit selection speech synthesis system constructed using the emphasis expression labels generated by our proposed method can convey the emphasis information effectively while maintaining the naturalness of synthetic speech.

key words: *speech synthesis, prosodic labeling, hidden Markov model, prosodic phrase boundary, emphasis expression*

1. Introduction

Nowadays unit-selection-based concatenative synthesis [2] and HMM-based parametric synthesis [3], [4] are the two most popular speech synthesis approaches. For either of them, 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 intelligibility and naturalness of the synthetic speech, especially for the unit-selection-based concatenative synthesis methods. Speech database annotation commonly consists of phonetic segmentation and prosodic labeling. In terms of phonetic segmentation, the text-analysis-based phonetic transcription and the HMM-based segmentation techniques have already achieved satisfactory performance and been widely used in practical systems [5], [6]. On the other hand, manual prosodic labeling is still necessary in most cases in order to construct high quality speech synthesis systems. However, manual labeling is laborious and very time-consuming. Thus, automatic prosodic labeling has attracted the attentions of many researchers.

Various methods have been proposed to label the prosodic descriptors of the speech databases automatically [7]–[21]. Most of these methods are supervised classification based approaches [7]–[15], which means that a certain amount of manually labeled training data of each database is necessary before annotating the remaining sentences automatically. Decision tree is the most popular classifier [7], [8], [10], [11] for the supervised classification. Generally speaking, these supervised approaches have already achieved good performance. However, the manual prosodic labels are not always available. For example, nowadays there are huge amounts of speech data on the Internet, the data can be utilized to construct the speech synthesis systems for different speakers respectively. In this situation, the unsupervised approaches would become more attractive than the supervised ones. Several methods which adopt unsupervised approaches for the prosodic labeling can be found in [16]–[21]. Generally speaking, most of these methods utilize a two-step strategy, i.e. *initialization* and *refinement*. In [16], [17], [19], the iterative processing of model training and prosodic labeling was used in the refinement step. Specifically, Ananthakrishnan et al. [16] initialized the prosodic labels by applying clustering algorithms to partition the acoustic space into two classes. In the refinement step, the initial prosodic labels were used to train a maximum a posteriori (MAP) classifier and were updated iteratively. Ni et al. [17] initialized the accent labels according to the POS of the words. In the refinement step, the accent labels were used to train a series of HMM based classifiers and were updated iteratively. Chiang et al. [19] proposed a joint prosodic labeling and modeling method which determined the prosodic labels and built the prosodic models simultaneously. The initial labels were obtained by a decision tree which was designed based on prior knowledge of the prosodic boundary labels. In the refinement step, the parameters of the prosodic models and the prosodic labels

Manuscript received September 24, 2013.

Manuscript revised January 14, 2014.

[†]The authors are with the National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, China.

[*]Part of this work was presented at ICASSP 2013 [1].

a) E-mail: yangcy@mail.ustc.edu.cn

b) E-mail: zhling@ustc.edu.cn

c) E-mail: lrdai@ustc.edu.cn

DOI: 10.1587/transinf.E97.D.1449

were updated iteratively. Different from the methods using iterative processing of model training and prosodic labeling, Huang et al. treated the prosodic labels as a discrete latent variable of a generative mixture model in [18]. Specifically, some representative samples for one or both classes were identified in the initialization step first. In the refinement step, a generative mixture model was trained from both the identified labeled set and a large pool of unlabeled set using expectation maximization algorithm. In some other work, the refinement step was skipped, such as [20], [21], where the emphasis expression labels were obtained directly by examining the differences between the *log* F0 values of the natural and synthetic speech samples of each phrase.

In this paper, an unsupervised prosodic labeling method using context-dependent hidden Markov models (CD-HMM) is proposed. This method also adopts the "initialization and refinement" framework. In the initialization step, the initial prosodic labels are obtained by unsupervised clustering using the acoustic features which can represent the specific characteristics of the prosodic descriptor to be labeled. In the refinement step, the iterative processing of model training and prosodic labeling was conducted to refine the initial prosodic labels. Specifically, the CD-HMMs of the spectral parameters, F0s and phone durations are estimated using the initial prosodic labels. Then, the prosodic labels are updated by Viterbi decoding under the maximum likelihood criterion given the sequences of acoustic features and the trained CD-HMMs. The model training and prosodic labeling procedures are executed iteratively until the labeling results converge.

This proposed method has several advantages. Compared with the initialization steps in [17]–[19], the influence of the known context features on the acoustic features is considered in the unsupervised clustering. In contrast to the refinement steps in [16], [18] where the type of the prosodic boundary following each prosodic word was determined independently, the proposed method adopts the Viterbi decoding approach to decide the types of all prosodic boundaries within a sentence simultaneously. Maeno et al. [20] also utilized the CD-HMMs in the emphasis expression labeling by examining the differences between the *log* F0 values of the natural and synthetic speech samples. However, no further refinement was conducted in their method. The iterative processing of model training and prosodic labeling in our proposed method is intended to improve the accuracy of prosodic labeling further.

The contents of the prosodic labels vary with the languages and the styles of the databases. In this paper, Mandarin speech synthesis databases are adopted in our experiments, and two prosodic descriptors, i.e., the prosodic phrase boundary and the emphasis expression, are used to evaluate the performance of our proposed method. The prosodic labels of the reading-style Mandarin speech synthesis databases commonly refer to prosodic boundaries [8], [10], [11], [14], [15], [18], [19]. Among different levels of prosodic boundaries, the prosodic phrase boundaries tend to be the most difficult ones for either manual or automatic labeling. In contrast to the prosodic word boundaries which can be precisely predicted by the text-analysis module, the prosodic phrase boundaries are much more context-dependent and speaker-dependent. If they are labeled manually, it is not only time-consuming but also difficult to guarantee the consistency among different human annotators. In addition to the prosodic phrase boundaries, the labeling of emphasis expressions for an audiobook database is also studied in this paper. The audiobook databases have been adopted to construct speech synthesis systems [22]–[24] in order to improve the expressiveness of the systems constructed using the traditional reading-style databases. The emphasis expression is one of the most representative prosodic descriptors for the Mandarin audiobook databases. Here, the emphasis expression labels are defined at the prosodic word level, i.e., to indicate each prosodic word to be emphasized or not. Similarly, it is also time-consuming and difficult to label the emphasis expressions manually for a large-sized audiobook database. In our proposed method, the initialization step need to be designed for different prosodic descriptors respectively. In our implementation, the prosodic phrase boundary labels are initialized according to the duration of pauses between every two consecutive prosodic words [1]. The initial labels of the emphasis expression are obtained by examining the differences between the *log* F0 values of the natural and synthetic speech samples for each prosodic word.

This paper is organized as follows. Section 2 introduces the framework of our proposed unsupervised prosodic labeling method. Section 3 and Sect. 4 present some detailed introduction to the prosodic phrase boundary labeling and emphasis expression labeling respectively. Section 5 reports the objective and subjective experimental results and the conclusions are presented in Sect. 6.

2. Unsupervised Prosodic Labeling Using CD-HMM

2.1 Overview of the Proposed Method

Figure 1 shows the flowchart of the proposed method for the unsupervised prosodic labeling. The whole method consists of three main steps: initialization, model training and prosodic labeling. The initial prosodic labels are firstly obtained in the initialization step by unsupervised clustering. The acoustic features used for clustering vary with the prosodic descriptor to be labeled. After that, the CD-HMMs of the spectral parameters, F0s and phone durations are estimated using the initial prosodic labels. In the prosodic labeling step, the prosodic labels of each utterance are then updated by Viterbi decoding under the maximum likelihood criterion given the acoustic feature sequences and the trained CD-HMMs. Once all the utterances in the speech database are processed, a new model training procedure is conducted using the updated prosodic labels. The model training and prosodic labeling procedures are conducted iteratively until the labeling results converge. More detailed descriptions to the prosodic labeling step and model training

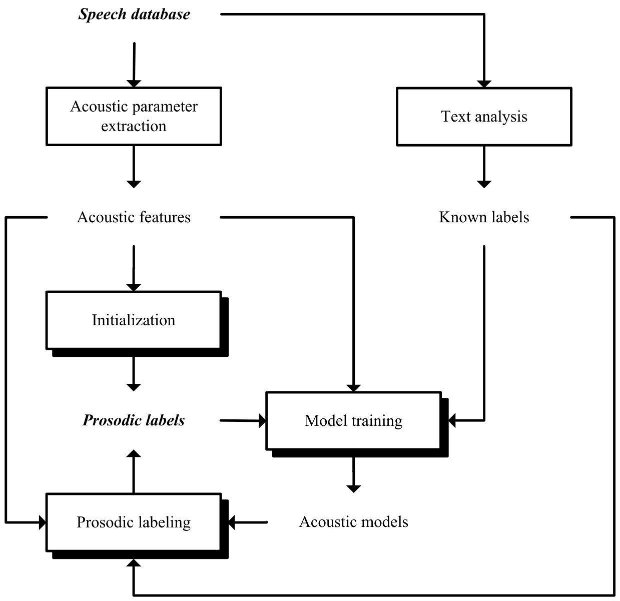

Fig. 1 Flowchart of the unsupervised prosodic labeling method. "Known labels" stand for the known phonetic and prosodic labels generated by the text-analysis module. "Prosodic labels", which are initialized first and then updated iteratively, stand for the prosodic labels expected to be labeled.

step are given in the following sub-sections.

2.2 Prosodic Labeling

The basic idea of prosodic labeling in our proposed method is similar to the statistical approach of automatic speech recognition (ASR), which can be expressed as

$$C^* = \arg\max_{C} P(\boldsymbol{O}|\boldsymbol{\lambda}, C_g, C)P(C), \tag{1}$$

where $\boldsymbol{O}$ stands for the acoustic features extracted from the speech waveforms of an utterance to be labeled; $\boldsymbol{\lambda}$ denotes the trained acoustic models; C_g represents the known phonetic and prosodic labels and C stands for the prosodic labels that are expected to be predicted. Once C is determined, it can be combined with C_g to calculate the output probability $P(\boldsymbol{O}|\boldsymbol{\lambda}, C_g, C)$ of the acoustic features for the corresponding acoustic models. $P(C)$ denotes a prior distribution of the unknown labels without any information of the acoustic features. In order to estimate prior distribution of the unknown labels, a large number of manual labels are usually necessary. In this paper, we ignore this prior distribution for simplicity and thus Eq. (1) can be simplified as

$$C^* = \arg\max_{C} P(\boldsymbol{O}|\boldsymbol{\lambda}, C_g, C). \tag{2}$$

The Viterbi decoding algorithm in ASR [25] is used here to solve Eq. (2). A "word graph" representing all of the possible prosodic labeling results is firstly constructed for each utterance based on the known phonetic and prosodic labels together with the possible values of the prosodic labels to be predicted. Then Viterbi decoding algorithm is applied to search the best path within the "word graph" and to derive the labeling results.

2.3 Model Training

The aim of model training is to estimate the acoustic models which describe the distributions of the acoustic features given the phonetic and prosodic labels. Here, the CD-HMM is adopted as the acoustic models and the model training procedure is similar to the one used in the HMM-based parametric speech synthesis [26]. 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 features 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) [27]. A decision tree based model clustering method using the minimum description length (MDL) criterion is applied to the model training in order to avoid the data-sparsity problem. Then each utterance in the training database is segmented into states by Viterbi alignment using the trained CD-HMMs. Based on the results of state segmentation, CD-HMMs of the phone durations can be estimated using the same decision-tree-based model clustering technique.

3. Unsupervised Prosodic Phrase Boundary Labeling

In our Mandarin speech synthesis systems, a three-level structure is commonly adopted to describe the prosodic characteristics of an utterance, which consists of prosodic word, prosodic phrase and sentence levels. Among them, the prosodic phrase boundary is the most difficult one for manual and automatic labeling.

The definition of prosodic word and prosodic phrase boundaries in Mandarin can be given from two points of view, syntax [28] and phonetics [29], respectively. The proposed method is designed according to the acoustic characteristics described in the phonetic definition. From the perspective of phonetics, the labels of prosodic word and prosodic phrase boundaries are given through the perception experiments conducted by human annotators. A prosodic word is a tone sandhi group bearing one word stress; A prosodic phrase contains one or more prosodic words bearing one phrasal stress. The perceived pause between prosodic phrases is longer than that of prosodic words. Usually, there is a pitch reset between prosodic phrases [29].

3.1 Initialization

In the initialization step, the prosodic phrase boundary labels are generated without human intervention. These initial labels are used in the following iterative processing of model training and prosodic labeling. Once the prosodic word boundaries and the sentence boundaries are given, initializing the labels of the prosodic phrase boundaries becomes an unsupervised binary classification problem, to determine whether each prosodic word boundary should be a prosodic phrase boundary or not. From the phonetic definition described above, it can be found that the "longer pause", which is frequently accompanied by a significant pitch reset, is a typical characteristic to differentiate prosodic phrase boundary from the prosodic word boundary. Therefore, the durations of the pauses at the prosodic word boundaries are extracted as the features for classification in the initialization step.

As shown in Fig. 2, a text analysis module is adopted to give the phonetic and prosodic labels excluding the prosodic phrase boundary positions of each utterance in the speech database. In order to extract the pause duration at each prosodic word boundary, a phonetic symbol "sp", which stands for the short pause, is inserted at the end of the phonetic transcriptions of each prosodic word. The CD-HMMs of the spectral parameters, F0s and phone durations are trained without using the context features related to 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 trained models. Considering that other context features besides the prosodic boundaries may also affect the duration of these short pauses, a normalization is applied using the trained context-dependent phone duration distributions as

$$\hat{d}_{sp} = \frac{d_{sp} - \mu}{\sigma}, \tag{3}$$

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 features of the pause. Then a K-medians clustering algorithm is used to divide the prosodic word boundaries into two classes according to the normalized pause-durations. The midpoint of cluster centers is used here as the threshold value. The boundaries belonging to the class with longer pauses, i.e. normalized pause-durations spanning longer than the threshold value, are initialized as the prosodic phase boundaries, while the other ones are kept as the prosodic word boundaries.

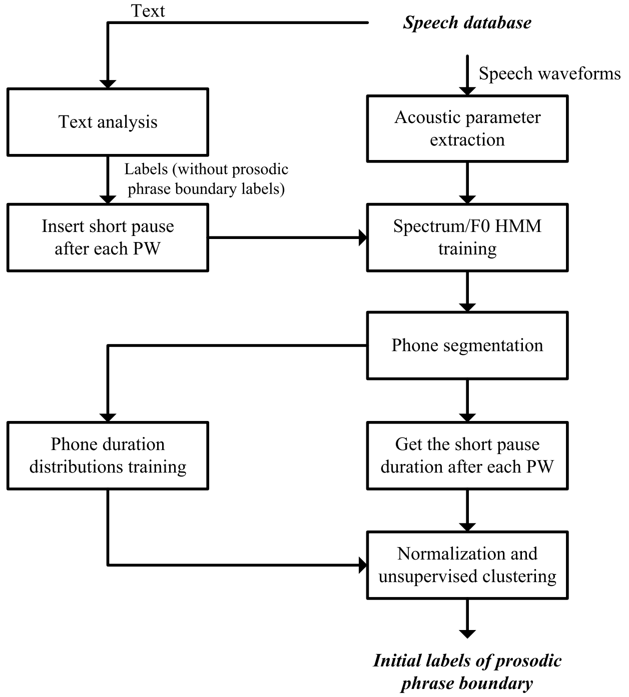
Text
Speech database
Speech waveforms
Text analysis
Acoustic parameter extraction
Labels (without prosodic phrase boundary labels)
Insert short pause after each PW
Spectrum/F0 HMM training
Phone segmentation
Phone duration distributions training
Get the short pause duration after each PW
Normalization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

Fig. 2 Flowchart of the initialization step for prosodic phrase boundary labeling. "PW" stands for prosodic word.

3.2 Context Features for Model Training

Table 1 lists the context features used in the model training of prosodic phrase boundary labeling. Compared to HMM-based parametric speech synthesis, the number of the context features is reduced here. In our previous work [30], [31], it was found that the context features listed in Table 1 were more important for the prosodic phrase boundary labeling than other context features used in the HMM-based parametric speech synthesis system. The reduced features can help to simplify the construction of "word graph" and control the complexity of the Viterbi decoding procedure in the prosodic labeling [31].

3.3 Two-Pass Viterbi Decoding

According to the context features listed in Table 1, the "word graph" for prosodic phrase boundary labeling can be derived from the outputs of text analysis module. Taking the short sentence "语音合成 (speech synthesis)" as an example, the text analysis results of this sentence are shown in Fig. 3

Table 1 Context features used in the labeling of prosodic phrase boundary.

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

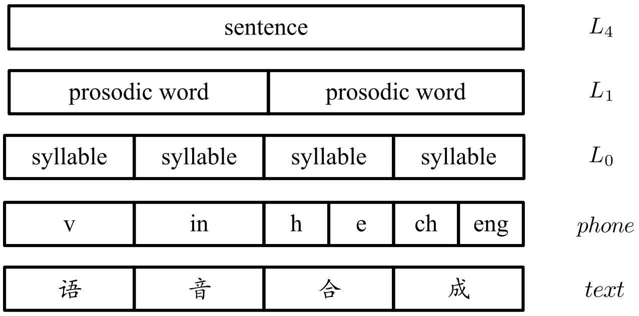

Fig. 3 Phonetic transcriptions and prosodic structure of the short sentence "语音合成 (speech synthesis)". The L_3, i.e. the prosodic phrase boundary, is missing, because it is the prosodic labels to be predicted in this task.

which contains the phonetic transcriptions and the known prosodic labels. From this figure, it can be found that there are four syllables in this sentence. A prosodic word boundary position is located between the second and third syllable. “v”, “in”, “h”, “e”, “ch”, “eng” stand for the phones of these syllables. L_0, L_1, L_3, L_4 represent for the syllable, prosodic word, prosodic phrase, and sentence boundary respectively. The “word graph” of this sentence is shown in Fig. 4. The context features of Tone Groups is omitted in this figure for simplification. Here, “sil” represents for the silence at the beginning and end of the sentence to be labeled. “null” means that the next phone does not exist. The best path in the “word graph” can be found by Viterbi decoding under the maximum likelihood criterion given the acoustic feature sequences and the trained CD-HMMs. The labeling results can be derived from the best path then.

Three kinds of acoustic features are used in our methods. They are spectral parameters, F0s and phone durations. In our previous work [30], [31], it was found that all of these features, especially phone durations, played an important role in the prosodic phrase boundary labeling. However, these features can not be used in the Viterbi decoding simultaneously, because spectral parameters and F0s are frame-level features but phone durations are phone-level features. Therefore, a two-pass Viterbi decoding strategy is applied here. 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 the N-best paths of each utterance are firstly obtained by Viterbi decoding using the CD-HMMs of the spectrum and F0 features. After that, these N hypotheses are rescored using the context-dependent models of the phone durations. Finally, the prosodic phrase boundary labels of the utterance are derived from the best path.

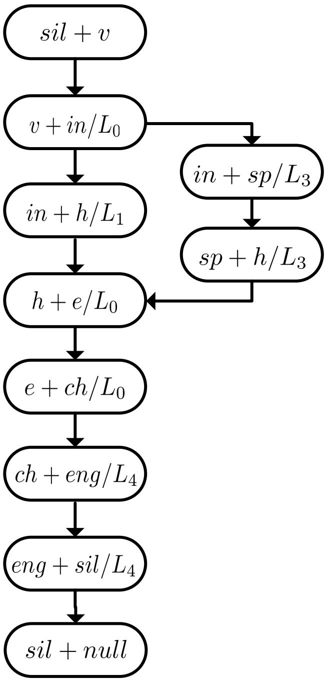

Fig. 4 “Word graph” of the short sentence “语音合成 (speech synthesis)” for prosodic phrase boundary labeling. It is constructed according to the phonetic transcriptions and prosodic structure shown in Fig. 3. The definitions of the symbols in the “word graph” are given in Sect. 3.3.

4. Unsupervised Emphasis Expression Labeling of Mandarin Audiobook Database

Audiobook databases consist of a large amount of expressive speech. Thus, it is promising to use this kind of databases to produce expressive synthetic speech. Emphasis expression is one of the most important characteristics of expressive speech. In this section, the unsupervised prosodic labeling framework proposed in Sect. 2 is applied to the emphasis expression labeling of a Mandarin audiobook database.

4.1 Initialization

In the initialization step, the method proposed in [20] was adopted to obtain the initial labels of emphasis expression. Though Mandarin is not a pitch accent language, the raise of F0 contour is still an important characteristic of emphasis expression [32]. The labeling of emphasis expression is treated as a binary classification problem for simplicity. As shown in Fig. 5, the acoustic features are extracted from the speech waveforms at first. Then the CD-HMMs of spectral parameters, F0s and phone durations are trained without using the context features of emphasis expressions. After that, the *log* F0 sequence can be generated from these models. Finally the initial labels of emphasis expression can be obtained by examining the differences of mean *log* F0 values between natural and generated parameters of each prosodic

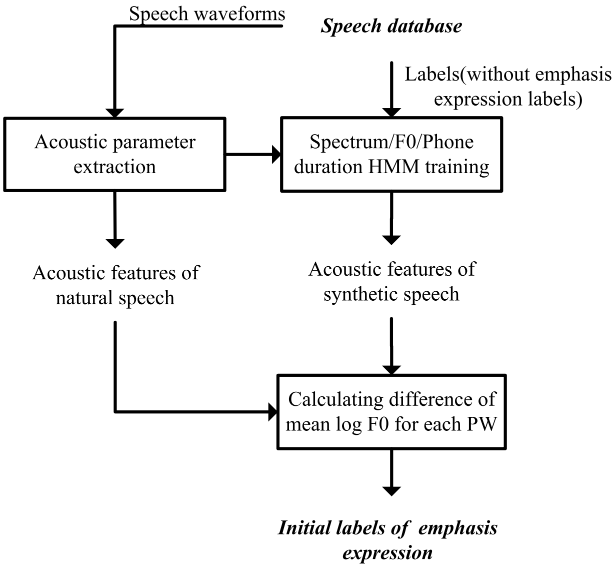
Speech waveforms
Speech database
Labels(without emphasis expression labels)
Acoustic parameter extraction
Spectrum/F0/Phone duration HMM training
Acoustic features of natural speech
Acoustic features of synthetic speech
Calculating difference of mean log F0 for each PW
Initial labels of emphasis expression

Fig. 5 Flowchart of the initialization step for emphasis expression labeling. “PW” stands for prosodic word.

Table 2 Context features used in the labeling of emphasis expression ("PW" stands for "prosodic word", "PP" stands for prosodic phrase).

Category	Context features
Phone Groups	{previous, current, next} phone
Tone Groups	the tone of {previous, current, next} syllable
Position Groups	$\left\{\begin{array}{c}\text{relative}\\\text{absolute}\end{array}\right\}$ positions of current $\left\{\begin{array}{c}\text{syllable}\\\text{PW}\\\text{PP}\\\text{sentence}\end{array}\right\}$
Boundary Groups	the prosodic boundary type at current syllable
Emphasis Groups	the emphasis expression labels of current PW

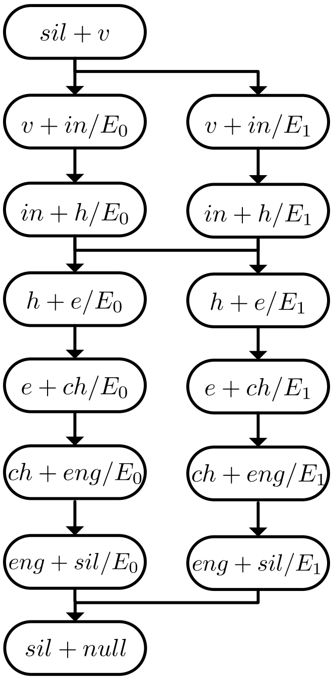

Fig. 6 "Word graph" of the short sentence "语音合成" (speech synthesis) for emphasis expression labeling. It is constructed according to the phonetic transcriptions and prosodic structure shown in Fig. 3. The definitions of the symbols in the "word graph" are given in Sects. 3.3 and 4.3.

word. A prosodic word is labeled as an emphatic prosodic word if the difference is larger than zero, otherwise it is labeled as a neutral prosodic word.

4.2 Context Features for Model Training

Table 2 lists the context features used in the model training of emphasis expression labeling. All of the context features used in the HMM-based parameter speech synthesis are adopted here. Among all of these context features, only the Emphasis Groups depend on the labels of emphasis expression. The context features of other groups are known at the labeling time. Therefore, using such detailed context features can improve the accuracy of the acoustic models without increasing the complexity of decoding.

4.3 An Example of "Word Graph" for Emphasis Expression Labeling

The "word graph" for emphasis expression labeling can be constructed according to the context features listed in Table 2. Take the short sentence "语音合成 (speech synthesis)" for example and assume that the result of text-analysis module can also be expressed as Fig. 3. Then, the "word graph" for emphasis expression labeling can be constructed as shown in Fig. 6. In this figure, only the context features representing current phone, next phone, and emphasis expression are used in order to simplify the illustration. As for the emphasis expression labels, E_0 stands for the neutral prosodic word and E_1 stands for the emphatic one.

5. Experiments

5.1 Experiments on Prosodic Phrase Boundary Labeling

5.1.1 Experimental Conditions

For the experiments of prosodic phrase boundary labeling, a standard reading-style Mandarin speech synthesis corpus was used. This corpus was uttered by a professional female speaker. It contains 13,000 utterances and lasts about 20 hours. The prosodic boundaries of all these utterances were labeled by professional human annotators. Besides the manual labeling results, the prosodic phrase boundary labels generated by three automatic labeling methods were compared in our experiments. These three labeling methods are described as follows:

- **Text-based labeling**. A C4.5 decision tree based classifier was constructed using the Weka tools [33] to determine whether each prosodic word boundary should be a prosodic phrase boundary or not. All the features used for classification were generated by the text-analysis module, such as the POS and the number of syllables in a prosodic word. The text set for training the decision tree based classifier consisted of 20,000 sentences with manual labels of prosodic phrase boundaries.
- **CD-HMM-based supervised labeling**. 1,000 utterances were picked up from the speech database. 900 of them with manual prosodic phrase boundary labels, which were phonetically and prosodically balanced, were used for the model training introduced in Sect. 2.3. The remaining 100 utterances were used as the test set for the objective evaluation. Finally, the prosodic phrase boundary positions of all utterances in the database were labeled by Viterbi decoding using the trained models.
- **CD-HMM-based unsupervised labeling**. The same 1,000 utterances as the CD-HMM-based supervised labeling were used for the unsupervised model training. The threshold of $\hat{d}_{sp}$, which is defined as (3), was obtained by unsupervised clustering in the initialization step. The value of the threshold is 0.0407. After the initialization of prosodic phrase boundary labels, the iterative processing of model training and prosodic labeling was implemented. The two-pass Viterbi decoding strategy was adopted here. The 40-best paths of

each utterance were firstly generated by the Viterbi decoding using the CD-HMMs of the spectrum and F0 features. These 40 hypotheses were then rescored using the CD-HMMs of the phone durations. After that, the updated prosodic phrase boundary labels of the utterance were derived. The model training and prosodic labeling were conducted iteratively, until the prosodic boundary labels and the CD-HMMs converged. The converged CD-HMMs were then applied to label all utterances in the database.

In both the supervised and unsupervised prosodic phrase boundary labeling, the speech waveforms were sampled at 16kHz. The acoustic parameters were extracted by STRAIGHT [34], 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. In order to evaluate the labeling result, three kinds of measurement were used in the experiments. They were precision, recall and F-score. The definitions of these three measurements are as follows:

1) Precision: the percentage that an automatic labeled prosodic phrase boundary agrees with the label of the reference result.

2) Recall: the percentage that a prosodic phrase boundary of the reference result is detected by the automatic labeling method.

3) F-score (F-measure): the harmonic mean of the precision and recall. The definition can be expressed as the following equation.

$$\text{F-score} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

In the unsupervised prosodic phrase boundary labeling, the F-score between the labeling results of adjacent iterations is used as a measurement of convergence. Here, the reference labels is defined as the labeling result of the former iteration. The condition of convergence is that F-score between last two iterations is above 99%. The F-scores of adjacent iterations calculated in this experiment are shown in Table 3. It can be found that the prosodic phrase boundary labels converged after six iterations. Figure 7 gives the log likelihood per frame of the training data in the iterative processing of model training and prosodic labeling, which shows that the converged CD-HMMs can model the training data better than the initial models.

Table 3 The F-score (%) of adjacent iterations in the iterative processing of model training and prosodic labeling for prosodic phrase boundary labeling.

Iteration	0	1	2	3	4	5	6
F-score	N/A	73.7	92.8	96.7	98.1	98.8	99.2

5.1.2 Objective Evaluation

We performed an objective evaluation among the three automatic labeling methods by comparing their labeling results with the manual labels on a test set. Precision, recall and F-score were chosen here as the measurements. These measurements have just been defined in Sect. 5.1.1. The manual labels are used as the reference result.

The test set for all these automatic labeling methods consisted of 100 utterances, which were not included in the training set of the supervised labeling but were used during the iterative processing of model training and prosodic labeling for the unsupervised labeling. This is considered to be reasonable because the manual labels are not required for the unsupervised labeling and all the utterances in the database can be used in the iterative processing. The manual labels were obtained by the voting results among three human annotators. There are 264 prosodic phrase boundary labels and 788 prosodic word boundary labels in the test set. Table 4 shows the labeling consistency between every two human annotators.

Table 5 lists the precisions, recalls and F-scores of different methods. Because F-score considers both precision and recall, it is also used here as an overall measurement of the labeling performance. From this table, it can be found that the initial labels of the CD-HMM-based unsupervised approach is much better than that of the text-based approach. This indicates the importance of acoustic cues in determining the prosodic phrase boundary positions for a speech syn-

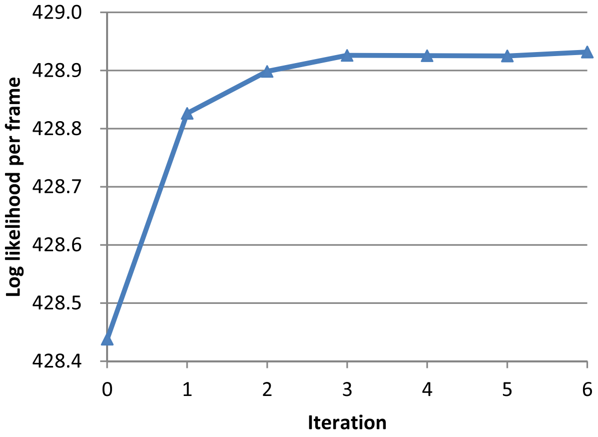

Fig. 7 Log likelihood per frame of the training data in the iterative processing of model training and prosodic labeling for prosodic phrase boundary labeling. The x-axis refers to the number of iterations in the iterative processing.

Table 4 The F-score (%) between every two human annotators on the test set for the prosodic phrase boundary labeling.

	Annotator 1	Annotator 2	Annotator 3
Annotator 1	N/A	79.7	70.9
Annotator 2	79.7	N/A	76.9
Annotator 3	70.9	76.9	N/A

Table 5 The precision (P), recall (R) and F-score (F) of prosodic phrase boundary labeling on the test set for different methods.

Method	P(%)	R(%)	F(%)
Text-based	36.1	77.3	49.2
Supervised CD-HMM	68.5	85.6	76.1
Unsupervised CD-HMM (initial)	55.4	72.4	62.7
Unsupervised CD-HMM (converged)	59.1	94.0	72.5

thesis database. After iterative processing of model training and prosodic labeling, the F-score of the unsupervised labeling increases from 62.7% to 72.5%, which is close to the F-score of the CD-HMM-based supervised labeling. From Table 4, it can be found that the average F-score among the three human annotators is 75.8% on the test set. Hence, the converged labels of unsupervised prosodic phrase boundary labeling is satisfactory when compared with the consistency among different human annotators. In addition, comparing the labeling results of CD-HMM-based supervised labeling and CD-HMM-based unsupervised labeling after iterative processing of model training and prosodic labeling, it can be found that the precision decreases from 68.5% to 59.1% and the recall increases from 85.6% to 94.0%. This indicates that the CD-HMM-based unsupervised labeling method tends to assign more prosodic phrase boundary labels than the supervised method.

5.1.3 Subjective Evaluation

Four speech synthesis systems were constructed using the manual prosodic phrase boundary labels, and the results of the three automatic labeling methods listed in Sect. 5.1.1. The HMM-based unit selection speech synthesis approach [35] was adopted and all of the 13,000 utterances were used for constructing these systems. Twenty utterances, which were not included in the database, were synthesized by the four systems respectively and were evaluated by eight listeners. Each listener was required to give a score from 1 (bad) to 5 (good) on the naturalness of each synthetic utterance. The average mean opinion scores (MOS) for these systems are shown in Fig. 8.[†] From Fig. 8 and Table 5, it can be found that the quality of the prosodic phrase boundary labeling plays an important role in the performance of the unit selection system. This is expectable by the HMM-based unit selection approach [35]. In this approach, calculations of the target and concatenation cost are depending on the context features of the candidate units. Many important context features are related to the labels of prosodic phrase boundary. Therefore, the inaccurate labels of prosodic phrase boundary can result in the selections of inappropriate units. Here, 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 these three methods (Tukey's HSD test at $\alpha \leq 0.01$). This indicates that the systems using the prosodic phrase boundary labels given by the CD-HMM-based supervised and unsupervised labeling methods are both comparable to the one constructed using the manual labels.

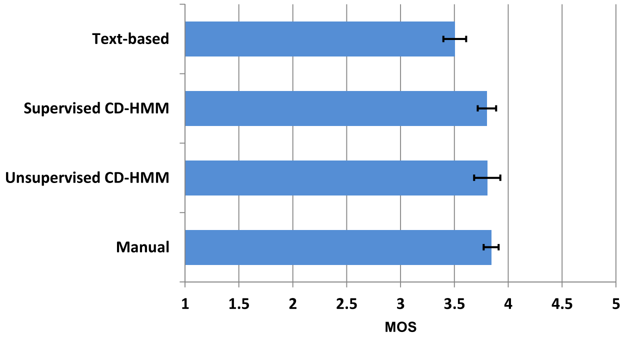

Fig. 8 Mean opinion scores (MOS) with 95% confidence intervals of the systems constructed using the manual prosodic phrase boundary labels and the results of three automatic prosodic phrase boundary labeling methods.

5.2 Experiments on Emphasis Expression Labeling

5.2.1 Experimental Conditions

For the experiments of emphasis expression labeling, a Mandarin audiobook database was used. The database is composed of the recordings of essays read by a male narrator. It contains 3,000 utterances, which lasts about 3 hours. The prosodic phrase boundaries of all these utterances were labeled by professional human annotators and the unsupervised labeling methods described in Sect. 3 respectively. The F-score of the prosodic phrase boundary labeling on the whole audiobook database is 84.43%, which is a satisfactory result. So the following experiments were conducted based on the prosodic phrase boundaries obtained by the unsupervised labeling methods.

In the experiments of emphasis expression labeling, all of the 3,000 utterances were used in the unsupervised model training. The speech waveforms were sampled at 16kHz. The acoustic parameters were extracted by STRAIGHT [34], 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.

After the initialization of emphasis expression labels, the iterative processing of model training and prosodic labeling was applied. The model training and prosodic labeling were conducted iteratively, until the emphasis expression labeling results and the CD-HMMs converged. Similar to the prosodic phrase boundary labeling in Sect. 5.1.1, the F-score between the labeling results of two adjacent iterations was used here as a measurement of convergence. The F-score values calculated in this experiments are shown in Table 6, which indicate the emphasis expression labels converged after five iterations. Figure 9 shows the log likelihood per frame of the training data in the iterative process-

[†] Some examples of the synthetic speech generated by the systems with different prosodic phrase labels are available at http://home.ustc.edu.cn/˜yangcy/USProsodyLabelingFull/demo.html.

Table 7 The F-scores (%) among human annotators and the CD-HMM-based unsupervised labeling method on the test set for emphasis expression labeling.

	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5	Unsupervised
Annotator 1	N/A	55.8	56.1	58.4	54.5	33.5
Annotator 2	55.8	N/A	62.3	58.9	46.4	45.5
Annotator 3	56.1	62.3	N/A	52.9	42.0	41.5
Annotator 4	58.4	58.9	52.9	N/A	46.9	43.1
Annotator 5	54.5	46.4	42.0	46.9	N/A	39.3
Unsupervised	33.5	45.5	41.5	43.1	39.3	N/A

Table 6 The F-score (%) of adjacent iterations in the iterative processing of model training and prosodic labeling for emphasis expression labeling.

Iteration	0	1	2	3	4	5
F-score	N/A	94.3	97.5	98.4	98.8	99.1

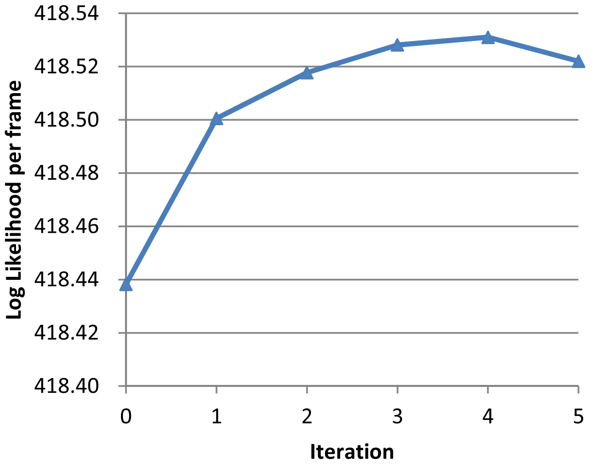

Fig. 9 Log likelihood per frame of the training data in the iterative processing of model training and prosodic labeling for emphasis expression labeling. The *x*-axis refers to the number of iterations in the iterative processing.

ing of model training and prosodic labeling. It can also be found that the converged CD-HMMs can model the training data better than the initial models. The log likelihood decreases a little in the fifth iteration, this is considered to be reasonable because the proposed method doesn't ensure the log likelihood to increase strictly.

In order to evaluate the performance of emphasis expression labeling, five professional human annotators were asked to give the emphasis expression labels on 100 utterances which were randomly selected from the audiobook database. The average percentage of prosodic words with the emphasis expression label in the test set given by human annotators is 18.7%. Table 7 shows the labeling consistency among the human annotators and the CD-HMM-based unsupervised labeling method on the test set. From this table, it can be found that the average F-score among these human annotators is 54.4%. Comparing Table 7 and Table 4, it can be found that the emphasis expression labeling consistencies between human annotators are much worse than those of prosodic phrase boundary labeling. This implies that it is more difficult to obtain consistent labeling results of emphasis expressions than prosodic phrase boundaries. Similarly, it can be found that the emphasis expression labeling consistency between CD-HMM-based unsupervised labeling and manual labeling is not as good as the one of the prosodic phrase boundary labeling. Considering that the purpose of emphasis expression labeling here is to synthesize the speech with appropriate emphasis expressions, a subjective listening test was conducted to measure the ability of conveying emphasis expressions for the synthetic speech.

5.2.2 Perception of Emphasis Expressions

The HMM-based unit selection speech synthesis system was constructed using the emphasis expression labels. The utterances of the same text but with different emphasis expression labels were synthesized and compared here. Twenty utterances, which were not included in the audiobook database, were synthesized first without any prosodic word emphasized. Another twenty utterances of the same text were synthesized but with an emphasized prosodic word in each utterance. The prosodic words to be emphasized were randomly selected from the utterances. Then, these utterances were used to compose twenty contrasting pairs which were evaluated by a group of native listeners subjectively. The first utterance of each pair, which was synthesized without any prosodic word emphasized, was used as the reference in the evaluation. If a difference of emphasis expression between a certain pair of the synthetic utterances was perceived, the listener was asked to mark the prosodic word that carried the emphasis. If no difference of emphasis expression was perceived, the listener didn't need to mark anything. Altogether eight native Mandarin listeners took part in this test. Precision, Recall and F-score were calculated here as the measurement. The definitions of these three measurements are:

1) Precision: the percentage that an emphatic prosodic word perceived by listeners agrees with the one intended to convey emphasis by the speech synthesis system.

2) Recall: the percentage that a prosodic word intended to convey emphasis by the speech synthesis is detected by the listeners.

3) F-score (F-measure): the harmonic mean of the precision and recall. The definition can be expressed as Eq. (4).

Table 8 Precision (P), recall (R) and F-score (F) of emphasis perception. The "Total" results are obtained by analyzing the results of all listeners altogether.

	P(%)	R(%)	F(%)
Listener 1	100.0	95.0	97.4
Listener 2	100.0	100.0	100.0
Listener 3	100.0	90.0	94.7
Listener 4	95.0	95.0	95.0
Listener 5	100.0	95.0	97.4
Listener 6	100.0	65.0	78.8
Listener 7	95.0	95.0	95.0
Listener 8	80.0	80.0	80.0
Total	96.0	89.4	92.6

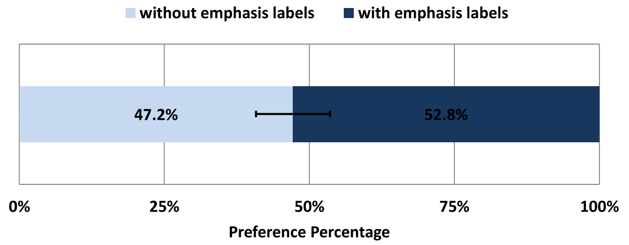

Fig. 10 Preference score of the systems constructed with and without emphasis expression labels with 95% confidence intervals.

The results of this experiment were listed in Table 8[†]. We can see that the unit selection speech synthesis system constructed using the emphasis expression labels can convey the emphasis expressions very well.

5.2.3 Naturalness Evaluation

In order to evaluate the naturalness of the speech synthesized by the systems using the unsupervised emphasis expression labeling method, two speech synthesis systems with and without emphasis expression labels were constructed and compared in our experiments. The HMM-based unit selection speech synthesis approach was adopted and all the 3,000 utterances were used for constructing the systems. Twenty utterances, which were not included in the training set, were synthesized by the two systems respectively. A paired-comparison preference listening test was conducted by 8 listeners. This test compared the naturalness of the synthesized speech generated from these two systems. For each pair, listeners were asked to tell which sentence is more natural. Then the preference score of these two systems could be calculated. Because the text analysis module used in this experiment can not predict the emphasis labels from the texts, the labels of these twenty sentences were decided manually here. The evaluation results are shown in Fig. 10[††]. It can be seen that the unsupervised emphasis expression labeling can maintain the naturalness of synthetic speech while the corresponding labels are given at synthesis time.

6. Conclusions

In this paper, an unsupervised method using CD-HMMs has been proposed for the prosodic labeling of speech synthesis databases. The method consists of three steps which are initialization, model training and prosodic labeling. The initial prosodic labels are firstly obtained by unsupervised clustering using the task-specific acoustic features. Then, the model training step and the prosodic labeling step are conducted iteratively to update the acoustic models and the labeling results. The unsupervised labeling of the prosodic phrase boundaries and the emphasis expressions of Mandarin speech synthesis databases has been investigated in this paper. In the experiments of prosodic phrase boundary labeling, the objective evaluation results have shown that this proposed method can achieve satisfactory prosodic phrase boundary labeling accuracy without requiring any manual labels. Also, the unit selection speech synthesis system constructed using the prosodic phrase boundary labels given by our proposed method is comparable to the one constructed using manual labels. In the experiments of emphasis expression labeling, the unit selection speech synthesis system constructed using the emphasis expression labels given by our proposed method can convey the emphasis information well while maintaining the naturalness of synthetic speech.

Although the experimental results demonstrate the effectiveness of the proposed method, several aspects of the current implementation can be improved in the future work. First, the prior distribution of the unknown prosodic labels is omitted for simplicity, while integrating the prior distribution may help improve the labeling performance further. Second, only the pause duration is used as the feature for classification in the initialization step of prosodic phrase boundary labeling. Some other features, e.g. pitch resets, can also be used here as supplements. Third, CD-HMMs with a single Gaussian distribution for each state are adopted as the acoustic models in this paper. They belong to the generative models with shallow architectures. Recently, the deep learning techniques have already achieved great success in the field of ASR [36]. Therefore, to improve the acoustic modeling by deep learning techniques will also be the task of our future work.

Acknowledgments

This work is partially funded by the National Nature Science Foundation of China (Grant No.61273032, No. 61273264) and National 973 program of China (Grant No. 2012CB326405). We also gratefully acknowledge the assistance of the research division of iFLYTEK Co. Ltd., Hefei, China, for providing the speech database with manual annotations.

[†] Some examples of the synthetic speech used for emphasis perception are available at http://home.ustc.edu.cn/˜yangcy/USProsodyLabelingFull/demo.html.

[††] Again, some examples of the synthetic speech generated by the systems with and without emphasis expression labels are available at http://home.ustc.edu.cn/˜yangcy/USProsodyLabelingFull/demo.html.

References

[1] C.-Y. Yang, Z.-H. Ling, and L.-R. Dai, "Unsupervised prosodic phrase boundary labeling of Mandarin speech synthesis database using context-dependent HMM," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.6875–6879, 2013.

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

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. EUROSPEECH, pp.2347–2350, 1999.

[4] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," Speech Commun., vol.51, no.11, pp.1039–1064, 2009.

[5] D.T. Toledano, L.A.H. Gomez, and L.V. Grande, "Automatic phonetic segmentation," IEEE Trans. Speech Audio Process., vol.11, no.6, pp.617–625, 2003.

[6] Y.-J. Wu, H. Kawai, J.-F. Ni, and R.-H. Wang, "Discriminative training and explicit duration modeling for HMM-based automatic segmentation," Speech Commun., vol.47, no.4, pp.397–410, 2005.

[7] C.W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," IEEE Trans. Speech Audio Process., vol.2, no.4, pp.469–481, 1994.

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

[9] A. Conkie, G. Riccardi, and R.C. Rose, "Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events," Proc. EUROSPEECH, pp.523–526, 1999.

[10] W.-X. Hu, T.-Y. Huang, and B. Xu, "Study on prosodic boundary location in Chinese Mandarin," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.I–501, 2002.

[11] X.-J. Ma, W. Zhang, Q. Shi, W.-B. Zhu, and L.-Q. Shen, "Automatic prosody labeling using both text and acoustic information," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.I–516, 2003.

[12] 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," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.509–512, 2004.

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

[14] C.-J. Ni, W.-J. Liu, and B. Xu, "Automatic prosody boundary labeling of Mandarin using both text and acoustic information," International Symposium on Chinese Spoken Language Processing (ISCSLP), pp.354–357, 2008.

[15] F.-Z. Liu, H.-B. Jia, and J.-H. Tao, "A maximum entropy based hierarchical model for automatic prosodic boundary labeling in Mandarin," International Symposium on Chinese Spoken Language Processing (ISCSLP), pp.257–260, 2008.

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

[17] X.-Q. Ni, Y.-N. Chen, F.K. Soong, M. Chu, and P. Zhang, "An unsupervised approach to automatic prosodic annotation," Proc. INTERSPEECH, pp.486–489, 2007.

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

[19] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, "Unsupervised joint prosody labeling and modeling for Mandarin speech," J. Acoust. Soc. Am., vol.125, pp.1164–1183, 2009.

[20] Y. Maeno, T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, "HMM-based emphatic speech synthesis using unsupervised context labeling," Proc. INTERSPEECH, pp.1849–1852, 2011.

[21] Y. Maeno, T. Nose, T. Kobayashi, T. Koriyama, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, "HMM-based expressive speech synthesis based on phrase-level F0 context labeling," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.7859–7863, 2013.

[22] Y. Zhao, D. Peng, L.-J. Wang, M. Chu, Y.-N. Chen, P. Yu, and J. Guo, "Constructing stylistic synthesis databases from audio books," Proc. INTERSPEECH, pp.1750–1753, 2006.

[23] K. Prahallad and A.W. Black, "Segmentation of monologues in audio books for building synthetic voices," IEEE Trans. Audio Speech Language Process., vol.19, no.5, pp.1444–1449, 2011.

[24] F. Eyben, S. Buchholz, and N. Braunschweiler, "Unsupervised clustering of emotion and voice styles for expressive TTS," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.4009–4012, 2012.

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

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

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

[28] H.-J. Wang, "The prosodic word and prosodic phrase of Chinese," Studies of The Chinese Language, vol.6, pp.525–536, 2000.

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

[30] 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," International Symposium on Chinese Spoken Language Processing (ISCSLP), pp.374–377, 2010.

[31] 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," J. Tsinghua University, vol.51, no.9, pp.1276–1281, 2011.

[32] A.-J. Li, "Prosodic analysis on conversations in Standard Chinese," Studies of The Chinese Language, vol.6, pp.525–535, 2002.

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

[34] 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 Commun., vol.27, pp.187–207, 1999.

[35] 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," Blizzard Challenge Workshop, 2012.

[36] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Process. Mag., vol.29, no.6, pp.82–97, 2012.

Chen-Yu Yang received the B.E. degree in electronic information engineering from University of Science and Technology of China, Hefei, China, in 2009. He is currently a Ph.D student of the National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China. His research interests include prosodic labeling and speech synthesis.

Zhen-Hua Ling received the B.E. degree in electronic information engineering, M.S. and Ph.D. degree in signal and information processing from University of Science and Technology of China, Hefei, China, in 2002, 2005, and 2008 respectively. From October 2007 to March 2008, he was a Marie Curie Fellow at the Centre for Speech Technology Research (CSTR), University of Edinburgh, UK. From July 2008 to February 2011, he was a joint postdoctoral researcher at University of Science and Technology of China and iFLYTEK Co., Ltd., China. He is currently an associate professor at University of Science and Technology of China. He also worked at University of Washington, USA as a visiting scholar from August 2012 to August 2013. His research interests include speech processing, speech synthesis, voice conversion, speech analysis, and speech coding. He was awarded IEEE Signal Processing Society Young Author Best Paper Award in 2010.

Li-Rong Dai was born in China in 1962. He received the B.S. degree in electrical engineering from Xidian University, Xi'an, China, in 1983 and the M.S. degree from Hefei University of Technology, Hefei, China, in 1986, and the Ph.D. degree in signal and information processing from the University of Science and Technology of China, Hefei, in 1997. He joined University of Science and Technology of China in 1993. He is currently a Professor of the School of Information Science and Technology, University of Science and Technology of China. His current research interests include speech synthesis, speaker and language recognition, speech recognition, digital signal processing, voice search technology, machine learning, and pattern recognition. He has published more than 50 papers in these areas.