



# Automatic Phrase Boundary Labeling of Speech Synthesis Database Using Context-Dependent HMMs and N-Gram Prior Distributions

Qian Chen<sup>1</sup>, Zhen-Hua Ling<sup>1</sup>, Chen-Yu Yang<sup>2</sup>, Li-Rong Dai<sup>1</sup>

<sup>1</sup>National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China

<sup>2</sup>Human Language Technology Department, Institute for Infocomm Research, A\*STAR, Singapore

cq1231@mail.ustc.edu.cn      zhling@ustc.edu.cn  
 yangc@i2r.a-star.edu.sg      lrdai@ustc.edu.cn

## Abstract

This paper presents an automatic phrase boundary labeling method for speech synthesis database annotation using context-dependent hidden Markov models (CD-HMMs) and n-gram prior distributions. At training stage, CD-HMMs are built to describe the conditional distribution of acoustic features given phonetic label and phrase boundary. In addition, n-gram models are estimated to represent the prior distributions of the phrase boundaries to be predicted. At decoding stage, the CD-HMMs and n-gram models are combined to predict the phrase boundaries by Viterbi decoding under maximum a posteriori (MAP) criterion. In our experiments, the proposed method utilizing context-dependent bigram prior distributions improved the F-score of phrase boundary labeling from 72.2% to 79.6% on the Boston University Radio News Corpus (BURNC), and from 69.6% to 81.0% on the Blizzard Challenge 2007 database respectively, comparing with the method using only acoustic models.

**Index Terms:** speech synthesis, phrase boundary, hidden Markov model, n-gram, maximum a posteriori

## 1. Introduction

A speech corpus with rich and precise annotation is important for building a speech synthesis system with highly intelligible and natural synthetic output. Annotating speech synthesis databases mainly consists of two sub-tasks, phonetic segmentation and prosodic labeling. This paper focuses on the issue of prosodic labeling. Prosodic structures of the utterances in speech synthesis databases provide suprasegmental descriptions of phonetic units [1], which play important roles in either context-dependent acoustic modeling or cost-function-based unit selection. The contents of prosody labeling vary among languages. In this paper, we investigate the methods for automatic phrase boundary labeling, which is a common task for many languages, such as English and Mandarin Chinese. It is laborious and time-consuming to annotate all the phrase boundary positions manually, especially for large-scale speech synthesis databases. In addition, it is difficult to guarantee the consistency among different annotators during manual phrase boundary labeling. Therefore, an automatic phrase boundary labeling method becomes necessary.

Various kinds of techniques to identify phrase boundary positions from speech signals have been proposed. In the early, Wightman and Ostendorf [2] used decision trees and a Markov sequence model to predict phrase boundaries. Sridhar et al. [3]

developed a maximum entropy-based automatic phrase boundary labeling framework that exploited both language and speech information. Rosenberg [4] used AdaBoost to train a classifier based on acoustic features and syntactic features, resulting in the state-of-the-art performance with an F-score of 76.1% on the BURNC corpus. Recently, Yang [5] used CD-HMMs derived from acoustic features and context information to label phrase boundaries automatically for Mandarin corpus. This method had two advantages. First, rich context features were adopted to model the context-dependent distributions of acoustic features using CD-HMMs. Therefore, the influence of known context features were also taken into account when determining the phrase boundary labels according to acoustic observations. Second, the Viterbi decoding [6] approach was adopted to determine the prosodic phrase boundaries within the entire utterance simultaneously. While in other methods [2, 3, 4], phrase boundaries were commonly predicted independently. On the other hand, an obvious disadvantage of this CD-HMM-based method [5] is that only acoustic observations were utilized and the textual prior knowledge of phrase boundary positions was ignored.

The method proposed in this paper is an extension of the work introduced in [5]. The proposed method is composed of a training part and a decoding part. In the training part, CD-HMMs are trained using acoustic features including spectra and fundamental frequencies (F0) together with rich context information. In addition, the prior distribution of phrase boundary positions without seeing the acoustic features is also estimated using an n-gram model. In the decoding part, a custom-designed network which is made up of paths that represent all possible phrase boundary labeling results is built automatically at first. Then, the maximum a posteriori (MAP) criterion which combines acoustic likelihood with textual prior probability is followed to decode the network using Viterbi algorithm.

This paper is organized as follows. In Section 2, the details of the proposed method are presented. In Section 3, the results of several experiments are shown and discussed. The conclusions are drawn in Section 4.

## 2. Method

### 2.1. Framework

Under MAP criterion, the phrase boundary labeling problem can be regarded as solving

$$C^* = \arg \max_C P(C|O, C_g), \quad (1)$$

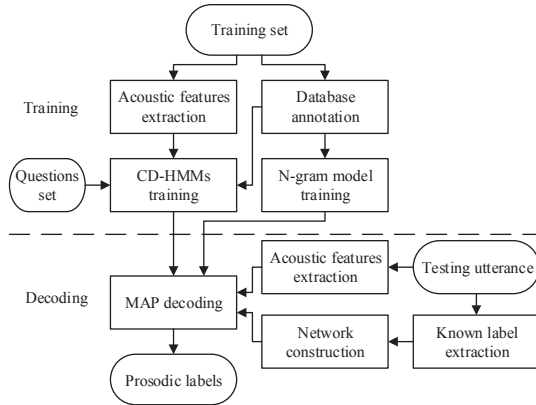


Figure 1: Flowchart of our proposed method.

where  $\mathcal{C}$  denotes the phrase boundary labels that are expected to be predicted and  $\mathcal{C}^*$  denotes the labeling results,  $\mathcal{O}$  represents the acoustic features derived from the speech waveforms of an utterance and  $\mathcal{C}_g$  denotes the known phonetic and context information. It is difficult to calculate  $P(\mathcal{C}|\mathcal{O}, \mathcal{C}_g)$  directly. By applying Bayes' Rule, we can get

$$\begin{aligned} P(\mathcal{C}|\mathcal{O}, \mathcal{C}_g) &= \frac{P(\mathcal{O}|\mathcal{C}, \mathcal{C}_g)P(\mathcal{C}, \mathcal{C}_g)}{P(\mathcal{O}, \mathcal{C}_g)} \\ &= \frac{P(\mathcal{O}|\mathcal{C}, \mathcal{C}_g)P(\mathcal{C}|\mathcal{C}_g)P(\mathcal{C}_g)}{P(\mathcal{O}, \mathcal{C}_g)}, \end{aligned} \quad (2)$$

and (1) can be simplified as

$$\mathcal{C}^* = \arg \max_{\mathcal{C}} P(\mathcal{O}|\mathcal{C}, \mathcal{C}_g)P(\mathcal{C}|\mathcal{C}_g). \quad (3)$$

Thus, the labeling result  $\mathcal{C}^*$  depends on the likelihood  $P(\mathcal{O}|\mathcal{C}, \mathcal{C}_g)$  which can be calculated by context-dependent acoustic models and the conditional probability  $P(\mathcal{C}|\mathcal{C}_g)$  which is regarded as a prior distribution of  $\mathcal{C}$ .

Figure 1 shows the flowchart of our proposed method for automatic phrase boundary labeling, which consists of a training part and a decoding part. In the training part, the CD-HMM based acoustic model and the n-gram based prior distribution are trained separately. In the decoding part, a custom-designed network which contains all possible phrase boundary labeling results is constructed automatically based on the known phonetic labels. Then the network is decoded under MAP criterion using the estimated CD-HMMs and the prior model to obtain the phrase boundary labels that are expected to be predicted.

## 2.2. Training

### 2.2.1. Acoustic model training

The acoustic model training process of our proposed method is similar to the HMM-based parametric speech synthesis (HTS) [7, 8]. Firstly, the spectral and F0 features are extracted from speech waveforms, and phonetic and contextual features are derived based on text analysis and manual annotation. The vector of spectral and F0 features for each speech frame consists of static, delta and delta-delta components. Then, the CD-HMMs are estimated under maximum likelihood criterion. The spectral features are modeled by a continuous probability distribution at each HMM state, while the F0 features are modeled by

a multi-space probability distribution (MSD) [9] at each HMM state because of the existing of unvoiced frames. A model clustering method using decision trees and minimum description length (MDL) criterion is utilized during the CD-HMM training to avoid the data-sparsity problem and to improve the robustness of the estimated model parameters.

### 2.2.2. Prior model training

In (2), the conditional probability  $P(\mathcal{C}|\mathcal{C}_g)$  is regarded as a prior probability model of  $\mathcal{C}$  because no acoustic observations are considered in this model. In our implementation, some assumptions to its model structure are made so as to simply the problem of training  $P(\mathcal{C}|\mathcal{C}_g)$ . Two model structures are considered as follows.

- *Context-independent modeling.* Here,  $\mathcal{C}$  and  $\mathcal{C}_g$  are assumed to be independent. Therefore,

$$P(\mathcal{C}|\mathcal{C}_g) = P(\mathcal{C}). \quad (4)$$

N-gram model, which has been popularly used in the language models of automatic speech recognition (ASR), is applied here to describe  $P(\mathcal{C})$ . In this task, the boundaries of prosodic words are known. For each word, it is required to determine whether it corresponds to a phrase boundary or not. In this case, the n-gram model of  $P(\mathcal{C})$  can be written as

$$P(\mathcal{C}) \approx \prod_{i=1}^M P(c_i|c_{i-n+1}, \dots, c_{i-1}), \quad (5)$$

where  $M$  represents total number of words in a utterance,  $c_i$  stands for the boundary type after the  $i$ -th word and  $n$  is the order of the n-gram model. The conditional distribution in (5) can be estimated using the training set with manual phrase boundary labels.

- *Context-dependent modeling.* Here, the dependency between  $\mathcal{C}$  and  $\mathcal{C}_g$  is considered. Therefore, the n-gram model (5) for the task of phrase boundary labeling becomes

$$P(\mathcal{C}|\mathcal{C}_g) \approx \prod_{i=1}^M P(c_i|c_{i-n+1}, \dots, c_{i-1}, c_{gi}), \quad (6)$$

where  $c_{gi}$ , which denotes the known context descriptions of the  $i$ -th word, replaces  $\mathcal{C}_g$  in the conditions to simply the model structure.

## 2.3. Decoding

To predict the phrase boundary labels at decoding stage, the network which contains all possible phrase boundary labeling results as paths must be constructed at first. Taking the utterance "auto insurance was overhauled last year" as example, the simplified network is shown as Figure 2. Actually, the experimental network is more complicated because phonemes are used as the basic unit for decoding. In Figure 2, "B" denotes there is phrase boundary after current word and "N" denotes opposite meaning, i.e., "auto/B" and "auto/N" denote the words with same phonetic symbols but different prosodic boundary types. "Sil", "Sp" and "Brth" stands for the segments of silence, short pause, and breath, which could appear at phrase boundaries according to the training database.

Similar to ASR [10], the Viterbi decoding algorithm is adopted to find the most probable path through the network under MAP criterion. This path gives the prediction results of

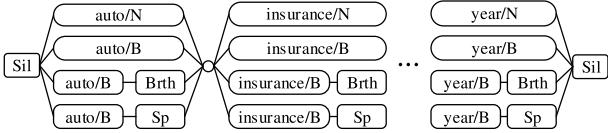


Figure 2: An example of the custom-designed network for labeling prosodic phrase boundaries.

	train	test	total
duration (minutes)	93.2	94.1	187.3
# utterances	1,000	1,000	2,000
# words	12,019	12,053	24,072
# phrase boundaries	1,962	2,005	3,967

Table 1: Two subsets of the Blizzard Challenge 2007 database used in our experiments.

phrase boundary labels. In our experiments, we found that the weight between the spectral component and the F0 component of  $P(O|C, C_g)$  is important to the performance of the final decoding results and should be tuned by experiments.

### 3. Experiments

#### 3.1. Experimental conditions

For fair comparison with other works [2, 3, 4], an objective evaluation was performed on the Boston University Radio News Corpus (BURNC) [11], which was recorded at the WBUR radio studio during broadcasting. The BURNC database consisted of a small number of news stories, repeated by many speakers. We used the recording materials of a female speaker (*f2b*) in the BURNC database for experiments, which contained 123 utterances, 9,090 words and 2,062 phrase boundaries. The duration was 56.1 minutes in total. In accord with [4], the performance of automatic phrase boundary labeling on this database was evaluated using ten-fold cross-validation in our experiment. Furthermore, in order to evaluate the performance of our proposed method on actual speech synthesis databases, another experiment was conducted on the Blizzard Challenge 2007 (BC2007) database [8]. This database contained 6,579 utterances of about 8 hours read by an American male native speaker. We annotated the prosodic structures of all utterances in this database manually. 2,000 utterances were randomly selected from the database and were further divided into a training set and a test set for automatic phrase boundary labeling. Table 1 shows some statistics of the two subsets used in this experiment.

In the BURNC and BC2007 databases, ToBI (tones and break indices) [12] was adopted as the prosody annotation convention, which gave a partition of break levels. In this work, we regarded the break index “4” as the phrase boundary to be annotated, just in accordance with [2, 4]. Meanwhile, to be consistent with [4], it was assumed that the sentence boundary and punctuation information was not available. Under this assumption, to achieve automatic phrase boundary labeling became a supervised two-class classification task, i.e., to predict whether there was a phrase boundary after each word given the speech waveforms and text information such as phonetic transcription, part-of-speech (POS), and so on.

Table 2 lists the context features used in the model training of CD-HMMs. The prosodic boundary type after current word was regarded as the target variable which was known in training set and needed to be predicted for test set. Acoustic features

Category	Context information
Phoneme	Identifiers of the current and next phoneme
	Position and number of phonemes in syllable
Syllable	Position and number of syllables in word
Word	Position of the current word in utterance
POS	Part-of-speech of the current word
Boundary	Boundary type after the current word

Table 2: The context features used in CD-HMM training.

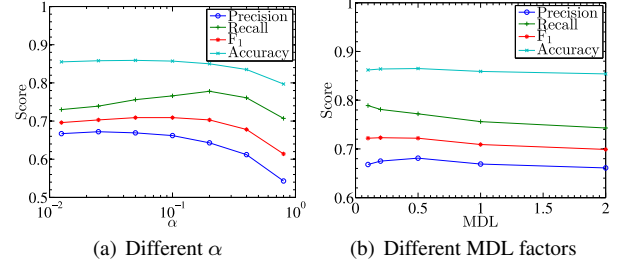


Figure 3: Performance of phrase boundary labeling using different  $\alpha$  and MDL factors on the BURNC database.

and other context features which derived from text analysis and manual annotation were regarded as input variables.

In our experiments, the speech waveforms were digitized at 16 kHz sampling rate using a 16 bit A/D. The acoustic features were extracted by STRAIGHT [13], including 40-order line spectral pairs (LSP) and F0 with their delta, delta-delta components. A 5-state left-to-right HMM structure was adopted to train the context-dependent acoustic models, where each state was assumed to obey a Gaussian distribution.

The prior model of phrase boundary was estimated using either context-independent (CI)  $n$ -gram model or context-dependent (CD)  $n$ -gram model. The  $n$ -gram models were estimated on the training set using Hidden Markov Model Toolkit (HTK) [14]. In our experiments, the CI-unigram model, CI-bigram model, CD-unigram, and CD-bigram model were built and compared. In the CD-unigram and CD-bigram models, the POS of each word was used as the  $c_{gi}$  in (6).

#### 3.2. Parameter tuning

Two system parameters were tuned on the BURNC database using the baseline method [5], which utilized only the acoustic likelihood part of the decoding criterion to label phrase boundary positions.

First, the weight factor  $\alpha$  between the F0 and spectral components of the likelihood function was tuned based on F-score ( $F_1$ ) which was calculated as the harmonic mean of precision and recall. Figure 3(a) shows the results of precision, recall,  $F_1$  and accuracy using ten-fold cross-validation with different  $\alpha$ , where  $\alpha$  and  $1 - \alpha$  denoted the weights of the F0 and spectrum components respectively. We can see that the highest  $F_1$  were obtained when  $\alpha$  was 0.05, which was consistent with the results obtained by [15] on another Mandarin corpus. Subsequently,  $\alpha = 0.05$  was adopted in the following experiments.

Second, the MDL factor used in the decision-tree-based model clustering is important to the generalization ability of the estimated acoustic models. Figure 3(b) draws the precision, recall,  $F_1$  and accuracy of phrase boundary labeling using ten-fold cross-validation with different MDL factors. An examination on the figure shows that the acoustic model achieved the best

Database	Method	Pre.	Rec.	F <sub>1</sub>	Acc.
BURNC	<i>Acoustic</i>	68.1	<b>77.2</b>	72.2	86.5
	<i>CI-unigram</i>	80.6	71.5	75.7	89.5
	<i>CI-bigram</i>	80.8	72.9	76.5	89.8
	<i>CD-unigram</i>	81.7	74.3	77.7	90.3
	<i>CD-bigram</i>	<b>85.2</b>	74.8	<b>79.6</b>	<b>91.3</b>
BC2007	<i>Acoustic</i>	58.7	<b>85.3</b>	69.6	87.3
	<i>CD-bigram</i>	<b>81.3</b>	80.7	<b>81.0</b>	<b>93.6</b>

Table 3: The performance of phrase boundary labeling on two English databases.

performance when MDL factor was 0.5 in view of both F<sub>1</sub> and accuracy. Similarly, the following experiments used 0.5 as the MDL factor for decision-tree-based model clustering.

### 3.3. Results

An objective evaluation was conducted to test the performance of the proposed method. The F<sub>1</sub> and accuracy between the predicted phrase boundary positions and the manually labeled ones were chosen as the main measurements. Table 3 lists the results of different annotation methods, where *Acoustic* is the baseline method using only acoustic likelihood at decoding time, and the other four methods are our proposed ones with different forms of prior distributions as introduced in Section 2.2.2. It can be found that the proposed method utilizing both acoustic models and prior distributions of prosodic labels worked better than the baseline method. Furthermore, context-dependent modeling achieved better performance than context-independent modeling on the BURNC database when utilizing POS of words as context features. The previous work of Rosenberg [4] used both acoustic and syntactic features to train an AdaBoost-based classifier for phrase boundary labeling. This method achieved the state-of-the-art performances (an F-score of 76.1% and an accuracy of 91.5% using ten-fold cross-validation) on the full BURNC database. From Table 3, we can see that our results on the BURNC database is competitive with an F-score of 79.6% and an accuracy of 91.3%. As shown in the last two rows in Table 3, our proposed method using CD-bigram prior distributions also outperformed the baseline method on the BC2007 database with an gain of F-score from 69.6% to 81.0%, just in accordance with the results on the BURNC database.

Figure 4 shows an example of an utterance in the BC2007 database with the Praat TextGrid display<sup>1</sup>. This figure consists of five tiers. Tier 1 shows the speech waveform; tier 2 shows the spectrogram with F0 and intensity contours; tiers 3 shows word-level transcriptions; tier 4 shows POS tags of each word; tier 5 shows the ground-truth phrase boundary types for each word given by manual annotation, where “B” denotes there is a phrase boundary after current word and “N” has opposite meaning. In this example, the baseline method, i.e., *Acoustic* in Table 3, incorrectly annotated phrase boundaries after both words “replied” and “Jeanne” by considering only acoustic cues. Our proposed method, i.e., *CD-bigram* in Table 3, gave correct results for this utterance because the probability of having two consecutive phrase boundaries before and after a noun, i.e., word “Jeanne”, was low according to the estimated CD-bigram prior distributions.

Furthermore, two HMM-based parametric speech synthesis systems were built using the test set of the BC2007 database

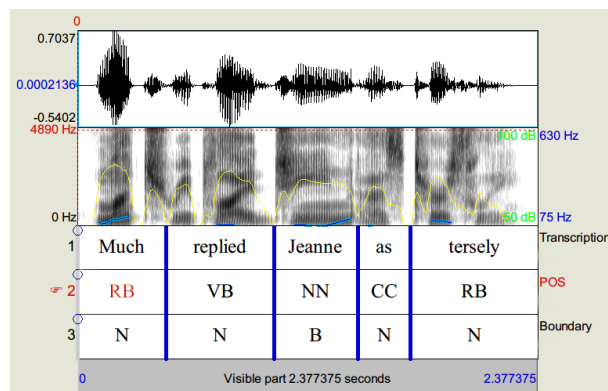


Figure 4: An example of an utterance in the BC2007 database with the Praat TextGrid display. The definitions of the tiers can be found in Section 3.3.

<i>Acoustic</i>	<i>CD-bigram</i>	N/P	<i>p</i>
31.5	42.0	26.5	0.091

Table 4: Preference scores(%) among the HMM-based synthesis systems built using two labeling methods on the BC2007 database, where N/P denotes “no preference” and *p* means the *p*-value of *t*-test between two systems.

shown in Table 1. The phrase boundary labeling results given by the *Acoustic* and the *CD-bigram* methods were adopted respectively. The systems were built following the method introduced in [16]. Finally, another 20 utterances not included in Table 1 were synthesized by these two systems. A preference test was conducted by crowdsourcing on Amazon Mechanical Turk (AMT) to compare the performance of these two systems. Each pair of synthetic speech were evaluated by 10 listeners. Table 4 shows the results. We can see that the system using our proposed phrase boundary labeling method achieved more preference than the one using the baseline method which considered only acoustic features. This superiority was significant at the 0.1 level.

## 4. Conclusions

In this paper, a method of automatic phrase boundary labeling of speech synthesis database has been proposed. CD-HMMs are trained using acoustic features and n-gram distributions are estimated from the phrase boundary labels of training set. The MAP criterion is followed to predict the phrase boundary positions by combining both acoustic models and prior distributions. Our experimental results show the effectiveness of our proposed method in improving the accuracy of phrase boundary labeling and the naturalness of synthetic speech. To evaluate the performance of this method on multi-speaker databases and to extend this idea to other prosodic labeling tasks will be our future work.

## 5. Acknowledgements

This work was supported by the National Nature Science Foundation of China (Grant No.61273032). The authors would like to thank Mr. Xiao-Hui Sun at NELSILIP, USTC for helping conduct subjective evaluation on Amazon Mechanical Turk.

<sup>1</sup><http://www.fon.hum.uva.nl/praat/>

## 6. References

- [1] E. Selkirk, "Sentence prosody: Intonation, stress, and phrasing," 1995.
- [2] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 469–481, 1994.
- [3] V. R. Sridhar, S. Bangalore, and S. S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 4, pp. 797–811, 2008.
- [4] A. Rosenberg, *Automatic detection and classification of prosodic events*. Columbia University, 2009.
- [5] C.-Y. Yang, Z.-H. Ling, and L.-R. Dai, "Unsupervised prosodic labeling of speech synthesis databases using context-dependent HMMs," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 6, pp. 1449–1460, 2014.
- [6] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, vol. 13, no. 2, pp. 260–269, 1967.
- [7] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.
- [8] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang, Y. Jiang, Z.-W. Zhao, J.-H. Yang, J. Chen *et al.*, "The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007," in *Blizzard Challenge Workshop*, 2007.
- [9] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1. IEEE, 1999, pp. 229–232.
- [10] H. Ney and S. Ortman, "Dynamic programming search for continuous speech recognition," *Signal Processing Magazine, IEEE*, vol. 16, no. 5, pp. 64–83, 1999.
- [11] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.
- [12] C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proceedings of the 1992 International Conference on Spoken Language Processing, ICSLP*, 1992, pp. 12–16.
- [13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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, no. 3, pp. 187–207, 1999.
- [14] S. Young, P. Woodland, and W. Byrne, "HTK: Hidden Markov model toolkit v1. 5," 1993.
- [15] C. Yang, L. Zhu, Z. Ling, and L. Dai, "Automatic phrase boundary labeling for a Mandarin TTS corpus using the Viterbi decoding algorithm," *Journal of Tsinghua University Science and Technology*, vol. 51, no. 9, pp. 1276–1281, 2011.
- [16] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006: an improved HMM-based speech synthesis method," in *Proc. Blizzard Challenge Workshop*, 2006.