Accent Type and Phrase Boundary Estimation Using Acoustic and Language Models for Automatic Prosodic Labeling

Tomoki Koriyama[†], *Hiroshi Suzuki*[†], *Takashi Nose*[‡], *Takahiro Shinozaki*[†], *Takao Kobayashi*[†]

[†]Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Japan
[‡]Graduate School of Engineering, Tohoku University, Japan
{koriyama, takao.kobayashi}@ip.titech.ac.jp

10.21437/Interspeech.2014-193

Abstract

This paper proposes an automatic prosodic labeling technique for constructing speech database used for speech synthesis. In the corpus-based Japanese speech synthesis, it is essential to use annotated speech data with prosodic information such as phrase boundaries and accent types. However, manual annotation is generally time-consuming and expensive. To overcome this problem, we propose an estimation technique of accent types and phrase boundaries from speech waveform and its transcribed text using both language and acoustic models. We use conditional random field (CRF) for the language model, and HMM for the acoustic model which has shown to be effective in prosody modeling in speech synthesis. By introducing HMM, continuously changing features of F0 contours are modeled well and this results in higher estimation accuracy than conventional techniques that use simple polygonal line approximation of F0 contours.

Index Terms: prosody, accent type, accent phrase boundary, HMM, CRF

1. Introduction

Prosodic labeling is an essential process for statistical prosody modeling in the corpus-based speech synthesis framework. Although an increasing amount of well-annotated speech data enhances the naturalness of synthetic speech, this leads to a problem of requiring manual prosodic labeling, which is generally time-consuming and expensive. Manual labeling has another problem that annotation performance depends on transcribers [1]. One of the approaches to overcome these problems is to prepare pre-annotated transcription and ask speakers to follow this transcription. However, this would force speakers to make utterances with unusual prosody that might be unnatural for the speakers. Therefore, it is important to develop automatic labeling of prosodic information.

For Japanese speech, which is a pitch accent language, predominant prosodic attributes are an accent type and its phrase boundary. There have been several approaches to estimating those attributes [2–4]. In [4], Suzuki et al., proposed prediction of accent type and phrase boundaries from an input text using conditional random field (CRF), where input features include part of speech (POS) and word frequency. Although this technique is promising for TTS systems, it is not sufficient to adopt the technique to prosodic labeling, because prosody varies with speakers and speaking situations even if an input text is the same.

In this context, there are techniques that take fundamental frequency (F0) information into account. Accent type estimation was performed by CRF in [2], where F0 contours of accent phrases were represented by a five-point approximation and clustered into a specific number of classes, and then the obtained cluster index was used as an input feature of the CRF. Accent phrase boundaries were estimated in [3], where F0 contours of short segments around word boundaries were represented by a three-point approximation and modeled by Gaussian models. Although these methods showed that incorporating F0 features enhances the estimation accuracy, F0 approximation with a few points stands for the risk of making actual F0 information too simple.

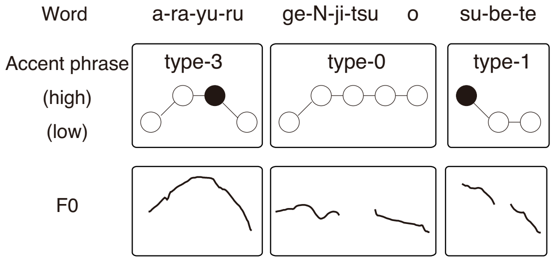


Figure 1: Example of accent type and phrase.

In this study, we propose a new approach to simultaneous estimation of the accent types and phrase boundaries for prosodic labeling of Japanese. Specifically, we incorporate hidden Markov models (HMMs) to modeling of acoustic features including F0 and CRF to linguistic features such as accent type and phrase boundary. In the proposed approach, N-best candidates of accent type and phrase boundary sequences are predicted from an input transcription using language models. Then the probability of acoustic features extracted from an input speech signal is calculated using acoustic models for each N-best candidate. The optimal sequence is estimated using the probabilities obtained from the language and acoustic models. Compared with the conventional methods using F0 approximation, HMM can model continuously changing F0 contours effectively as shown in prosody modeling in HMM-based speech synthesis [5].

2. Accent of Japanese

Japanese is known to have a pitch accent and usually described by high/low (H/L) pitch level in mora unit. An example of accent types and phrases is shown in Fig. 1. Accent in Japanese is a lexical property and the meaning of a word is changed when an accent type is altered. Accent phrase is a simple prosodic phrase and typically corresponds to break index level “2” in J-ToBI labeling scheme [6]. In Tokyo dialect, which we treat in this study, every N-mora phrase is classified into one of N accent types that are denoted by type-0, type-1, $\cdots$, and type-$N-1$. The number shows the position of the accent nucleus in which mora F0 falls rapidly. Type-0 implies the phrase does not have the accent nucleus.

Accents are not always determined uniquely from the input text. One of the reasons is that accents accompanied by multiple words are uttered in more than one way. The way of producing utterance generally varies depending on the speakers and speaking situations. Another reason is because of word concatenation. Word concatenation causes the change of accent nucleus, which is known as word accent sandhi formulated by Sagisaka et al. [7, 8]. However, the presence of word accent sandhi also varies in different speakers and situations.

3. Related work

There are similar but non-simultaneous approaches to accent type and accent phrase boundary estimation [2, 3]. These methods use polygonal line approximation of F0 contours for F0 modeling. In [2], each F0 contour of accent phrase was represented by a five-point approximation and clustered into a certain number of classes. Accent type sequence is modeled by CRF, where the clustered indexes of F0 contour $\boldsymbol{C}$ and linguistic features such as POS and the result of accent sandhi rules are used as the input feature of CRF. Thus, the optimal accent type sequence $\tilde{\boldsymbol{L}}$ is determined by the following equation

$$\tilde{\boldsymbol{L}} = \arg\max p(\boldsymbol{L}|\boldsymbol{C}, \boldsymbol{B}, \boldsymbol{W}) \quad (1)$$

where $\boldsymbol{B}$ and $\boldsymbol{W}$ correspond to accent phrase boundary sequence and word feature sequence, respectively.

In [3], F0 contours of short speech segments of about 300ms neighboring word boundary are represented by two lines by a three-point approximation. The lines' features for boundary and non-boundary are modeled by a single Gaussian, respectively. Let $\boldsymbol{F}$ be the approximated F0 features and the optimal accent phrase boundary sequence $\tilde{\boldsymbol{B}}$ is estimated by

$$\tilde{\boldsymbol{B}} = \arg\max p(\boldsymbol{F}|\boldsymbol{B}, \boldsymbol{W})^{\beta} p(\boldsymbol{B}|\boldsymbol{W}) \quad (2)$$

where β is a weight for adjusting the effect of the F0 features, $p(\boldsymbol{F}|\boldsymbol{B}, \boldsymbol{W})$ and $p(\boldsymbol{B}|\boldsymbol{W})$ are the probabilities modeled by Gaussians and CRF, respectively.

Another approach is to incorporate HMMs into prosodic phrase boundary estimation for Mandarin speech synthesis [9, 10]. In this method, the likelihood of acoustic features obtained by context dependent HMMs are used for estimating phrase boundaries. Although this approach is similar to our approach, they ignore the probability of language model, which is expected to be important for Japanese accent type estimation.

4. Methods

4.1. Model outline

When we perform prosodic labeling manually, we usually put accent phrase boundaries and accent types simultaneously because both depend on each other. For example, if there are two accent nuclei in a segment, the segment must have at least one phrase segment. Therefore, we propose a unified framework of simultaneous estimation of accent type and phrase boundary.

Assuming that speech waveform and a set of information about corresponding pronunciations, POSs, and pauses is provided. Let

$$\boldsymbol{B} = (b_1, b_2, \ldots, b_{N-1}) \quad (3)$$

Table 1: Features for the construction of the language models.

Accent type model
of moras,
of words,
POS of the first word,
predicted accent type by word accent sandhi rule, for preceding/current/succeeding accent phrase
Accent phrase boundary model
of moras,
POS for current/succeeding word,

be an accent phrase boundary sequence of a sentence, where N is the number of words in the sentence. $b_i (i = 1, \ldots, N-1)$ is a binary-valued variable that represents whether an accent phrase boundary exists or not between the i- and $i+1$-th words. In other words, there is an accent phrase boundary when $b_i = 1$, but not when $b_i = 0$. Moreover, let

$$\boldsymbol{L} = (l_1, l_2, \cdots, l_K) \quad (K \leq N) \quad (4)$$

denote an accent type sequence of a sentence, where K is the number of accent phrases in the sentence. The variable l_k has a categorical value represented by $0, \ldots, M_k - 1$ where M_k is the number of moras included in the k-th accent phrase. In addition, we define $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_N)$ as a word feature sequence, where $\boldsymbol{w}_i$ consists of POS and the number of moras, and $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ as an acoustic feature sequence which has T frames.

Here we consider the problem of estimating the optimal accent type sequence $\tilde{\boldsymbol{L}}$ and accent phrase boundary sequence $\tilde{\boldsymbol{B}}$ simultaneously by maximizing the following posterior:

$$(\tilde{\boldsymbol{L}}, \tilde{\boldsymbol{B}}) = \arg\max_{\boldsymbol{L}, \boldsymbol{B}} p(\boldsymbol{L}, \boldsymbol{B}|\boldsymbol{O}, \boldsymbol{W}). \quad (5)$$

This posterior can be reformulated as follows:

$$\begin{aligned}(\tilde{\boldsymbol{L}}, \tilde{\boldsymbol{B}}) &= \arg\max_{\boldsymbol{L}, \boldsymbol{B}} \frac{p(\boldsymbol{O}, \boldsymbol{L}, \boldsymbol{B}|\boldsymbol{W})}{p(\boldsymbol{O}|\boldsymbol{W})} \\ &= \arg\max_{\boldsymbol{L}, \boldsymbol{B}} p(\boldsymbol{O}|\boldsymbol{B}, \boldsymbol{L}, \boldsymbol{W}) p(\boldsymbol{L}|\boldsymbol{B}, \boldsymbol{W}) p(\boldsymbol{B}|\boldsymbol{W}) \\ &\approx \arg\max_{\boldsymbol{L}, \boldsymbol{B}} p(\boldsymbol{O}|\boldsymbol{B}, \boldsymbol{L}, \boldsymbol{W})^{\alpha} p(\boldsymbol{L}|\boldsymbol{B}, \boldsymbol{W}) p(\boldsymbol{B}|\boldsymbol{W})\end{aligned} \quad (6)$$

where α is a weight for controlling the effect of the acoustic feature sequence. Therefore, we can estimate the optimal accent type sequence $\tilde{\boldsymbol{L}}$ and accent phrase boundary sequence $\tilde{\boldsymbol{B}}$ by calculating the probabilities of the sequences of acoustic features, accent types, and accent phrase boundaries, namely $p(\boldsymbol{O}|\boldsymbol{L}, \boldsymbol{B}, \boldsymbol{W})$, $p(\boldsymbol{L}|\boldsymbol{B}, \boldsymbol{W})$, and $p(\boldsymbol{B}|\boldsymbol{W})$ for all possible combinations of $\boldsymbol{L}$ and $\boldsymbol{B}$. We refer to the models for $p(\boldsymbol{O}|\boldsymbol{L}, \boldsymbol{B}, \boldsymbol{W})$ and $p(\boldsymbol{L}|\boldsymbol{B}, \boldsymbol{W})p(\boldsymbol{B}|\boldsymbol{W})$ as the acoustic and language models, respectively.

4.2. Language model

CRF is a discriminative model that is widely used for the tasks of annotating labels on input sequences. Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_N)$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_N)$ be input and output sequences, respectively. We denote $\phi_f(\boldsymbol{x}, \boldsymbol{y})$ as the frequency count of a feature $f \in \boldsymbol{F}$. The probability distribution

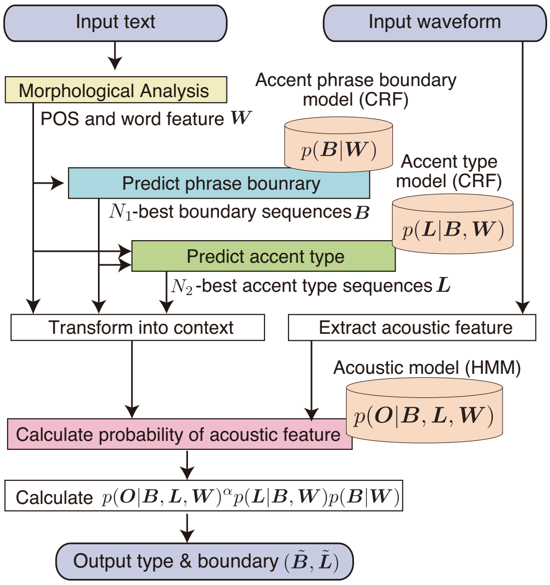

Input text
Input waveform
Morphological Analysis
Accent phrase boundary model (CRF)
POS and word feature W
$p(B|W)$
Accent type model (CRF)
Predict phrase bounrary
N_1-best boundary sequences B
$p(L|B,W)$
Predict accent type
N_2-best accent type sequences L
Transform into context
Extract acoustic feature
Acoustic model (HMM)
$p(O|B,L,W)$
Calculate probability of acoustic feature
Calculate $p(O|B,L,W)^{\alpha}p(L|B,W)p(B|W)$
Output type & boundary $(\tilde{B},\tilde{L})$

Figure 2: Flow of the proposed technique of accent type and phrase boundary estimation.

of $\boldsymbol{y}$ given by $\boldsymbol{x}$ is defined by

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \sum_{f\in \boldsymbol{F}} \theta_f \phi_f(\boldsymbol{x},\boldsymbol{y})}{\sum_{\boldsymbol{y}\in \boldsymbol{Y}} \exp \sum_{f\in \boldsymbol{F}} \theta_f \phi_f(\boldsymbol{x},\boldsymbol{y})} \quad (7)$$

where $\boldsymbol{Y}$ is a set of all possible output sequences, and θ_f is a weight for feature f, which is tuned in training for maximizing $p(\boldsymbol{y}|\boldsymbol{x})$.

In this study, CRF is used for an accent type model and an accent phrase boundary model. Table 1 shows the list of input features used for the language models. The features for the accent type and phrase boundary model are based on [2] and [3], respectively. The word accent sandhi rule formulated by [7] is included in the features. In this study, we use CRF++ [11] for training and estimation.

4.3. Acoustic model

The acoustic model used in the proposed technique is an HMM-based one that has been used for HMM-based speech synthesis. Since F0 cannot be observed in unvoiced regions, F0 sequences are modeled by a multi-space probability distribution HMM (MSD-HMM) [12]. The acoustic feature vector for HMM includes not only F0 but also spectral features such as mel-cepstrum so that spectral features support modeling of state transition in unvoiced regions. As the context, which is essential in the HMM-based speech synthesis framework, we use information obtained by the word feature sequence $\boldsymbol{W}$, accent phrase boundary sequence $\boldsymbol{B}$, and accent type sequence $\boldsymbol{L}$. HMM parameters are shared by decision trees because the combination of contextual factors is diverse.

4.4. Estimation procedure

As described in Sect. 4.1, the optimal accent type and phrase boundary sequence can be chosen by calculating probabilities for all possible combinations. However, the combination of prosody is so diverse[1] that it is practically inappropriate to compute the probabilities for all possible combinations. Hence, we use an N-best method to drop some combinations in this study. The outline of the proposed technique is shown in Fig. 2. In this procedure, first, POS information is extracted from an input text. N_1-best candidates of accent phrase boundary sequences are predicted from POS information using the accent phrase boundary model. After that, N_2-best candidates of accent type sequences are predicted for each candidate of accent phrase boundary sequence. Consequently, we have $(N_1 \times N_2)$ candidates and the probability of the acoustic feature extracted from an input waveform is calculated for each candidate. Finally, we choose the optimal candidate using (6).

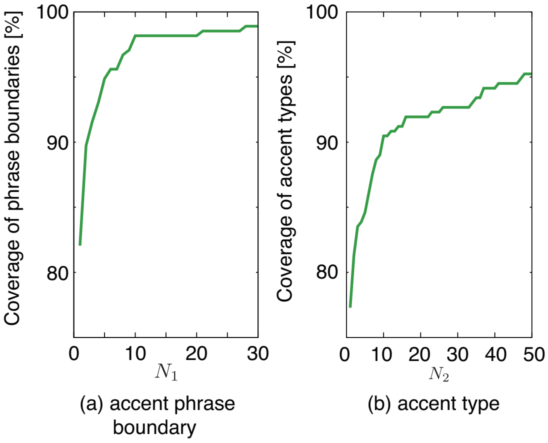


Figure 3: Coverages of correct accent phrase boundaries and accent types.

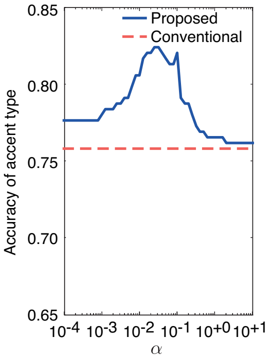
5. Experiments

5.1. Experimental conditions

We used six male speakers' speech data included in ATR Japanese speech database set B [13] for the experiments. We evaluated the proposed technique under two conditions: speaker-closed and speaker-open cases. The speaker-closed means that a certain amount of annotated data has been given already and additional data is labeled. In the speaker-open case, we trained the models without any prosodic labels of the target speaker. 53 sentences of one male speaker (MHT) were used for evaluation. We used 450 sentences of the target speaker for model training for the speaker-closed case, and used 2250 utterances of five male speakers for the speaker-open case. The five male speakers did not include the target speaker. The CRFs and HMMs are modeled separately.

We used the results of POS information of training and test data obtained using ChaSen [14] for a morphological parser with UniDic [15] as a dictionary. The errors by POS extraction were corrected manually. The acoustic feature vector consisted of F0, the 0-39th mel-cepstral coefficients, 5-band aperiodicity features, and their delta and delta-delta dynamic features, which were generally used in HMM-based speech synthesis. These features were extracted and obtained from 16kHz sam-

[1] There are 2^{M-1} possible accent phrase boundary sequences in M-word sequence. Moreover there exist $\prod_{k=1}^{K} M_k$ possible accent type sequences in each phrase boundary sequence where M_k is the number of moras in k-th accent phrase.



(a) Speaker-closed

(b) Speaker-open

Figure 4: Results of accent type estimation.

pled speech waveforms. The frame shift of the analysis was 5ms. In order to lessen the speaker dependency for speaker-open case, we employed speaker adaptive training (SAT) [16] and shared decision tree context clustering (STC) [17] for the training of speaker independent HMMs. We compared the proposed technique with accent type estimation proposed in [2] and accent phrase boundary estimation proposed in [3]. The number of clusters for the accent type estimation was fixed to 256.

5.2. Coverage of N-best candidates

First we evaluated the effects of the tuning parameters, N_1 and N_2, in the speaker-closed case. Figure 3 (a) shows the coverage of the annotated accent phrase boundaries in N_1-best candidates. From the figure, it is seen that the coverage was low when N_1 was smaller than 5 and almost converged when N_1 is more than 10. Similarly, Figure 3 (b) shows the coverage of the annotated accent types in N_2-best candidates under the condition that annotated accent phrase boundaries are given. The coverage was more than 90% when N_2 is more than 10. From these results, we use $N_1 = 10$ and $N_2 = 50$ in the following experiments.

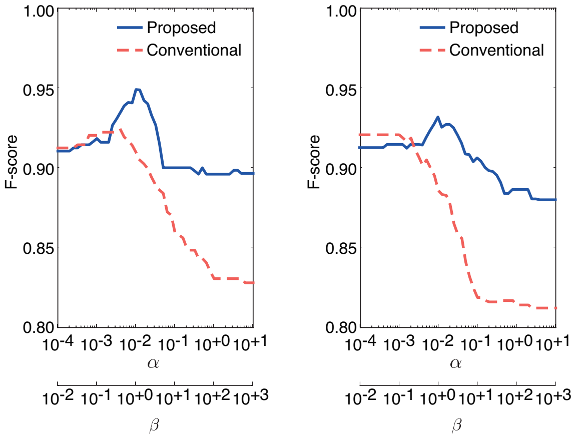
5.3. Accent type estimation

The results of comparative evaluation of accent type are shown in Fig. 4. The accuracy in Fig. 4 was measured under the condition that the annotated phrase boundaries are given. In these figures, when the weight α is a very small value like 10^{-4}, the scores are almost equivalent to those using only the language model, because the probability of acoustic feature, $p(\boldsymbol{O}|\boldsymbol{L}, \boldsymbol{B}, \boldsymbol{W})$ in (6), is almost ignored. Similarly, when the weight α is large e.g. $\alpha = 10$, the scores can be regarded as those using only the acoustic model, even though the candidates were chosen by the language model.

The accuracies of accent estimation of the proposed method were the highest at $\alpha = 2.5 \times 10^{-2}$ for the speaker-closed case and at $\alpha = 6.3 \times 10^{-2}$ for the speaker-open case. This asserts that using both acoustic and language models enhances the estimation performance as described in the related work. In addition, the highest scores of the proposed method in speaker-closed and speaker-open cases were about 6.5% and 7.3% higher than those of the conventional method, respectively.

5.4. Accent phrase boundary estimation

The results of comparative evaluation of accent phrase boundary are shown in Fig. 5 as a function of weight α for the proposed method and β for the conventional one. The F-score F is given by



(a) Speaker-closed

(b) Speaker-open

Figure 5: Results of accent phrase boundary estimation.

$$R = \frac{\text{(\# of correctly estimated boundaries)}}{\text{(\# of estimated phrase boundaries)}} \tag{8}$$

$$P = \frac{\text{(\# of correctly estimated boundaries)}}{\text{(\# of annotated phrase boundaries)}} \tag{9}$$

$$F = \frac{2PR}{P + R} \tag{10}$$

where R and P correspond to recall and precision, respectively. As described in the previous section, differences of the scores at $\alpha = 10$ and $\beta = 10^3$ are almost equivalent to those of the performance of acoustic models. Hence, it is seen that the use of HMM for acoustic model outperformed the conventional method. In both speaker-closed and speaker-open cases, the F-scores are highest at $\alpha = 10^{-2}$ and the scores were 0.949 and 0.932, respectively. As is seen that the proposed methods have higher performance than the conventional ones when we compare the best scores for each method.

6. Conclusions

In this paper, we have proposed the automatic labeling technique of accent type and phrase boundaries using CRF for language models and HMM for acoustic models, respectively. From the evaluation results of accent type estimation, it is confirmed that using both acoustic and language model enhances the estimation performance and the proposed technique outperformed the conventional ones. However, the performances still depend on the weight that controls the effect of acoustic features. Therefore, future work should investigate the method of choosing an appropriate weight. In addition, future work will utilize more data for the language model and refine the acoustic model to improve the accuracy. Furthermore, we should assess the effectiveness of automatically annotated data for statistical speech synthesis system.

7. Acknowledgements

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 24300071.

8. References

[1] A. K. Syrdal and J. T. McGory, “Inter-transcriber reliability of ToBI prosodic labeling.” in *Proc. INTERSPEECH*, 2000, pp. 235–238.

[2] K. Suzuki, A. Yamamoto, K. Cho, and Y. Yamashita, “Automatic accent type labeling for spoken sentences based on statistical methods using accentuation rules (in Japanese),” *The Journal of the Acoustical Society of Japan*, vol. 66, no. 10, pp. 487–496, 2010.

[3] A. Yamamoto, K. Cho, and Y. Yamashita, “Automatic prediction of accent phrase boundaries using linguistic and f0 information (in Japanese),” *IEICE technical report. Speech*, vol. 110, no. 401, pp. 37–42, 2011.

[4] M. Suzuki, R. Kuroiwa, K. Innami, S. Kobayashi, S. Shimizu, N. Minematsu, and K. Hirose, “Accent sandhi estimation of tokyo dialect of Japanese using conditional random fields,” *IEICE Trans. Inf. & Syst. (Japanese edition)*, vol. 96, no. 3, pp. 644–654, 2013.

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

[6] J. J. Venditti, “The J_ToBI model of Japanese intonation,” *Prosodic typology: The phonology of intonation and phrasing*, pp. 172–200, 2005.

[7] Y. Sagisaka, “Accentuation rules for japanese text-to-speech conversion,” *Review of the Electrical Communication Laboratories*, vol. 32, pp. 188–199, 1984.

[8] N. Minematsu, R. Kita, and K. Hirose, “Automatic estimation of accentual attribute values of words to realize accent sandhi in Japanese text-to-speech conversion,” in *Proc. 2002 IEEE Workshop on Speech Synthesis*, 2002, pp. 107–110.

[9] 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 *Proc. ISCSLP*, 2010, pp. 374–377.

[10] C.-Y. Yang and Z.-H. Ling, “Unsupervised prosodic labeling of speech synthesis databases using context-dependent HMMs,” *IEICE Trans. on Inf. & Syst.*, vol. 97, no. 6, pp. 1449–1460, 2014.

[11] CRF++, https://code.google.com/p/crfpp/.

[12] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Multi-space probability distribution HMM,” *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[13] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[14] M. Asahara and Y. Matsumoto, “Extended models and tools for high-performance part-of-speech tagger,” in *Proceedings of the 18th conference on Computational linguistics*, 2000, pp. 21–27.

[15] Y. Den, J. Nakamura, T. Ogiso, and H. Ogura, “A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation.” in *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, 2008, pp. 1019–1024.

[16] T. Aanastasakos, “A compact model for speaker-adaptive training,” *ICSLP*, vol. 2, 1996.

[17] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “A context clustering technique for average voice models,” *IEICE Trans. Inf. & Syst.*, vol. 86, no. 3, pp. 534–542, 2003.