INTEGRATION OF SPEAKER AND PITCH ADAPTIVE TRAINING FOR HMM-BASED SINGING VOICE SYNTHESIS

Kanako Shirota, Kazuhiro Nakamura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, Japan

ABSTRACT

A statistical parametric approach to singing voice synthesis based on hidden Markov models (HMMs) has been growing in popularity over the last few years. The spectrum, excitation, vibrato, and duration of the singing voice in this approach are simultaneously modeled with context-dependent HMMs and waveforms are generated from the HMMs themselves. Since HMM-based singing voice synthesis systems are "corpus-based," the HMMs corresponding to contextual factors that rarely appear in the training data cannot be well-trained. However, it may be difficult to prepare a large enough quantity of singing voice data sung by one singer. Furthermore, the pitch included in each song is imbalanced, and there is the vocal range of the singer. In this paper, we propose "singer adaptive training" which can solve the data sparseness problem. Experimental results demonstrated that the proposed technique improved the quality of the synthesized singing voices.

Index Terms— singing voice synthesis, hidden Markov model, speaker adaptive training, pitch adaptive training

1. INTRODUCTION

A statistical parametric approach to speech synthesis based on hidden Markov models (HMMs) has been growing in popularity over the last few years [1]. Context-dependent HMMs are estimated from speech databases in this approach and speech waveforms are generated from the HMMs themselves. This framework makes it possible to model different voice characteristics, speaking styles, or emotions without recording large speech databases. For example, adaptation [2], interpolation [3], and eigenvoice [4] techniques have been applied to this system, demonstrating that voice characteristics can be modified. The singing voice synthesis system has also been proposed using the HMM-based approach [5, 6].

The quality of the synthesized singing voices strongly depends on the training data because the HMM-based singing voice synthesis systems are "corpus-based." Therefore, HMMs corresponding to contextual factors that rarely appear in training data cannot be well-trained. Although databases including various contextual factors should be used in the HMM-based singing voice synthesis systems, covering all possible contextual factors is almost impossible since singing voices involve a huge number of contextual factors, e.g., pitch, tempo, keys, beat, dynamics, lyrics, note positions, durations, etc. Thus, a large quantity of singing voice data sung by a certain singer is necessary to train a model of the singer. However, it may be difficult to prepare a large enough quantity of the target singers' voice data. Furthermore, pitch needs to be properly covered particularly for HMM-based singing voice synthesis, since it has a great impact on the quality of the synthesized singing voices[1]. Nevertheless the pitch included in each song is imbalanced, and there is the vocal range of the singer. Consequently, a technique is required to obtain the high quality synthesized singing voices by using a small amount of singing voice data.

To solve the data sparseness problem for each speaker, "speaker adaptive training" [7] has been proposed in HMM-based text-to-speech synthesis. In this technique, the model to synthesize high quality speech can be trained using a small amount of data uttered by the target speaker, and the database that consists of several speakers' speech. In addition, to solve the data sparseness problem for pitch, "pitch adaptive training" [10] has been proposed in HMM-based singing voice synthesis. In this technique, all pitch can be synthesized by modeling differences between fundamental frequency (F_0) sequences extracted from waveforms and the pitch of musical notes. In this paper, we propose "singer adaptive training," in which "speaker adaptive training" and "pitch adaptive training" are integrated to obtain the high quality synthesized singing voices from a small amount of the target singers' data.

The rest of this paper is organized as follows. Section 2 gives an overview of the HMM-based singing voice synthesis system. Section 3 describes "speaker adaptive training" for HMM-based text-to-speech synthesis. Section 4 explains "pitch adaptive training" for HMM-based singing voice synthesis. Section 5 details "singer adaptive training" for HMM-based singing voice synthesis. Section 6 describes experimental conditions and the results of objective and subjective experiments, and Section 7 presents conclusions.

[1] F_0 modeling for HMM-based singing voice synthesis also has a great impact. Many techniques have been proposed [8, 9].

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

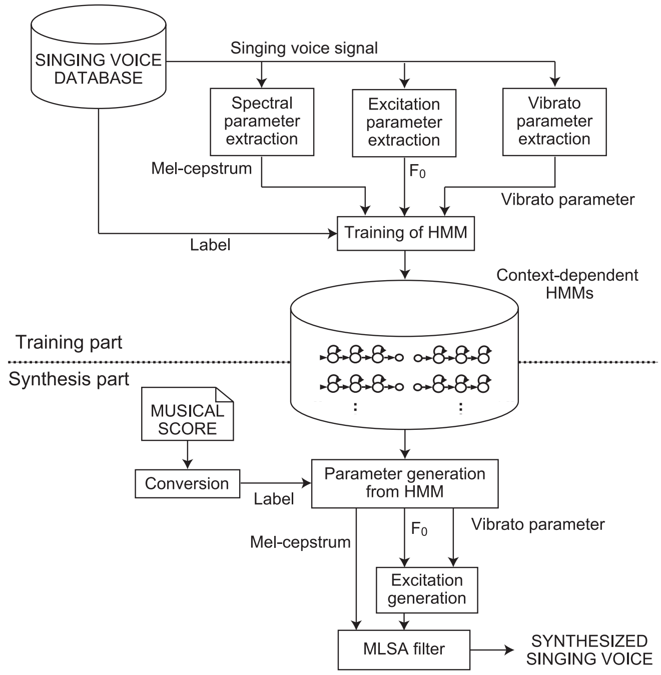

Fig. 1. *Overview of the HMM-based singing voice synthesis system.*

2. HMM-BASED SINGING VOICE SYNTHESIS SYSTEM

The HMM-based singing voice synthesis system is quite similar to the HMM-based text-to-speech synthesis system. However, they have distinct differences. Figure 1 gives an overview of the HMM-based singing voice synthesis system [5, 6], which consists of training and synthesis parts. The spectrum (e.g., mel-cepstral coefficients), excitation, and vibrato are extracted from a singing voice database in the training part and are then modeled with context-dependent HMMs. Context-dependent models of state durations are also estimated simultaneously. In the synthesis part, an arbitrarily given musical score including the lyrics to be synthesized is first converted into a context-dependent label sequence. Second, in accordance with the label sequence, a state sequence corresponding to the song is constructed by concatenating the context-dependent HMMs. Third, the state durations of the song HMM are determined with respect to the state duration models. Fourth, the spectrum, excitation, and vibrato parameters are generated by an algorithm to generate the speech parameters [11]. Finally, a singing voice is synthesized directly from the generated spectrum, excitation, and vibrato parameters by using a mel log spectrum approximation (MLSA) filter.

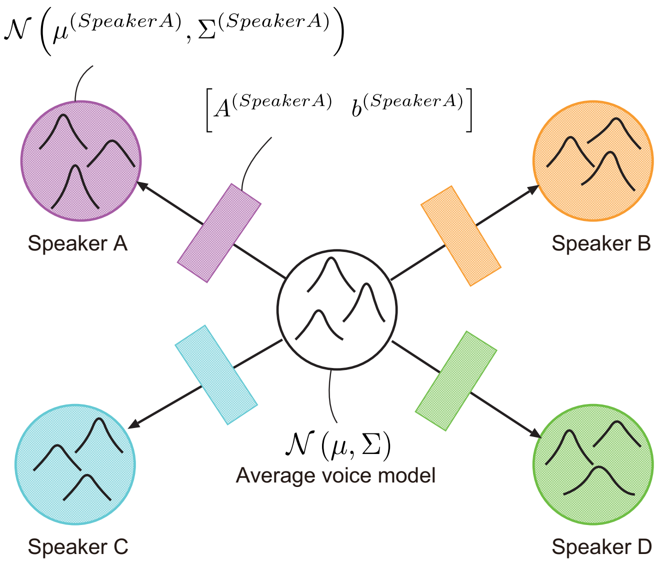
$\mathcal{N}\left(\mu^{(Speaker A)}, \Sigma^{(Speaker A)}\right)$
$\left[A^{(Speaker A)} \quad b^{(Speaker A)}\right]$
Speaker A
Speaker B
$\mathcal{N}(\mu, \Sigma)$
Average voice model
Speaker C
Speaker D

Fig. 2. *Overview of the "speaker adaptive training."*

3. SPEAKER ADAPTIVE TRAINING FOR HMM-BASED TEXT-TO-SPEECH SYNTHESIS

HMM-based speech synthesis systems heavily depend on training data in performance. Therefore, HMMs corresponding to contextual factors that rarely appear in the training data cannot be well-trained. To obtain a model to synthesize high quality speech using a small amount of data uttered by a target speaker, "speaker adaptive training" [7] has been proposed in HMM-based text-to-speech synthesis. Figure 2 gives an overview of the "speaker adaptive training." In this method, an average voice model and transformation matrices are estimated using the training data of several speakers' speech. The difference between the training speakers' model and the average voice model is assumed to be expressed in the "speaker adaptive training" algorithm as a simple linear regression function of the mean vectors and the covariance matrices of state output distributions. The transformation matrix is estimated by a speaker adaptation technique using constrained maximum likelihood-linear regression (CMLLR) [12], and the model of each speaker is obtained by applying the transformation matrix to the estimated average voice model. A mean vector $\hat{\boldsymbol{\mu}}_i^{(f)}$ and a covariance matrix $\hat{\boldsymbol{\Sigma}}_i^{(f)}$ in state i for training speaker f are defined as

$$\hat{\boldsymbol{\mu}}_i^{(f)} = \boldsymbol{A}_i^{(f)} \boldsymbol{\mu}_i + \boldsymbol{b}_i^{(f)}, \qquad (1)$$

$$\hat{\boldsymbol{\Sigma}}_i^{(f)} = \boldsymbol{A}_i^{(f)} \boldsymbol{\Sigma}_i \boldsymbol{A}_i^{(f)\top}, \qquad (2)$$

where $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, and $\left[\boldsymbol{A}_i^{(f)} \;\; \boldsymbol{b}_i^{(f)}\right]$ correspond to a mean vector and a covariance matrix of the average voice model, and a transformation matrix that indicates the difference between the model of the training speaker f and the average voice model, respectively. The model to synthesize the high quality speech can be obtained from a small amount of target speakers' data by using "speaker adaptive training."

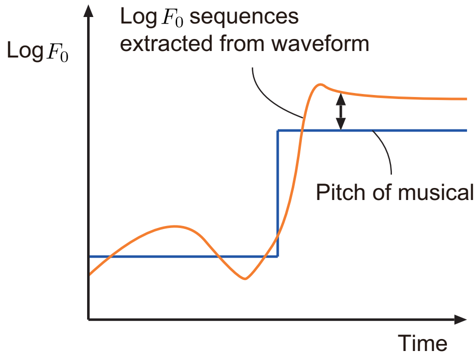

Fig. 3. *The differences between log F_0 sequences extracted from the waveform and the pitch of musical note.*

4. PITCH ADAPTIVE TRAINING FOR HMM-BASED SINGING VOICE SYNTHESIS

The HMM-based singing voice synthesis system is quite similar to the HMM-based text-to-speech synthesis system. Therefore, databases including various contextual factors should also be used for training. However, the data is sparse because singing voices involve numerous contextual factors (e.g., pitch, tempo, key, beat, and dynamics) in addition to them used in text-to-speech synthesis. Specifically, pitch should be properly covered since generated F_0 trajectories greatly affect the quality of the synthesized singing voices. However, the pitch included in the singing voices is imbalanced, and there is the vocal range of the singer. Therefore, it is difficult to prepare a database including all pitch. To solve the data sparseness problem for pitch, "pitch adaptive training" [10] in HMM-based singing voice synthesis has been proposed. The differences between log F_0 sequences extracted from the waveform and the pitch of a musical note shown in Fig. 3 are modeled in "pitch adaptive training." A mean $\hat{\boldsymbol{\mu}}_i^{(p)}$ of static features of log F_0 in state i of pitch p is defined as

$$\hat{\boldsymbol{\mu}}_i^{(p)} = \boldsymbol{\mu}_i + \boldsymbol{b}_i^{(p)}, \tag{3}$$

where $\boldsymbol{\mu}_i$ is a parameter of HMMs representing a mean of the differences between log F_0 extracted from the waveform and pitch of a musical note, $\boldsymbol{b}_i^{(p)}$ is log F_0 of a musical note in state i, and $\begin{bmatrix} \boldsymbol{I} & \boldsymbol{b}_i^{(p)} \end{bmatrix}$ is a transformation matrix for each pitch. Here, $\boldsymbol{I}$ expresses an identity matrix. Since the transformation matrices are fixed by the musical score, "pitch adaptive training" only estimates the parameters of HMMs. The model that can synthesize all pitch can be obtained by using "pitch adaptive training."

5. INTEGRATION OF SPEAKER AND PITCH ADAPTIVE TRAINING

Both "speaker adaptive training" and "pitch adaptive training" are technique to synthesize high quality speech and singing voices from limited data by using training data effectively.

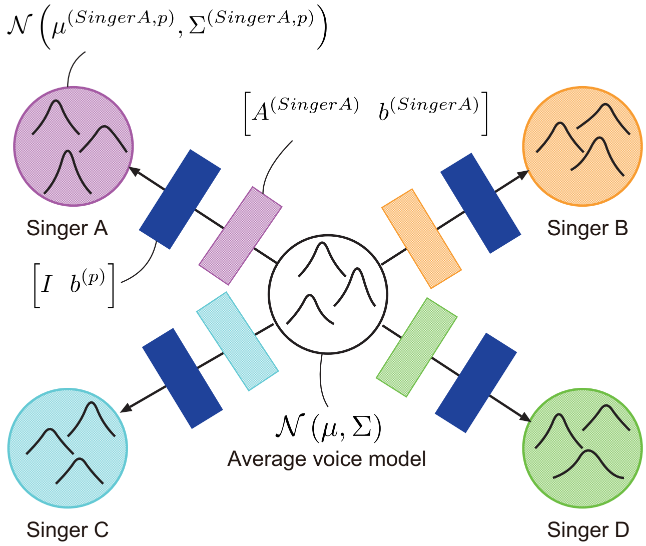
$\mathcal{N}\left(\mu^{(SingerA,p)}, \Sigma^{(SingerA,p)}\right)$
$\begin{bmatrix} A^{(SingerA)} & b^{(SingerA)} \end{bmatrix}$
Singer A
Singer B
$\begin{bmatrix} I & b^{(p)} \end{bmatrix}$
$\mathcal{N}(\mu, \Sigma)$
Average voice model
Singer C
Singer D

Fig. 4. *Overview of the "singer adaptive training."*

This paper proposes "singer adaptive training" integrating "speaker adaptive training" and "pitch adaptive training." Figure 4 gives an overview of the "singer adaptive training." A mean vector $\hat{\boldsymbol{\mu}}_i^{(f,p)}$ and a covariance matrix $\hat{\boldsymbol{\Sigma}}_i^{(f,p)}$ in state i of pitch p for training singer f are defined as

$$\hat{\boldsymbol{\mu}}_i^{(f,p)} = \boldsymbol{A}_i^{(f)}\boldsymbol{\mu}_i + \boldsymbol{b}_i^{(f)} + \boldsymbol{b}_i^{(p)}, \tag{4}$$

$$\hat{\boldsymbol{\Sigma}}_i^{(f,p)} = \boldsymbol{A}_i^{(f)}\boldsymbol{\Sigma}_i\boldsymbol{A}_i^{(f)\top}, \tag{5}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ correspond to a mean vector and a covariance matrix of the average voice model that represents the difference between log F_0 extracted from the waveform and pitch of a musical note. Furthermore, $\begin{bmatrix} \boldsymbol{A}_i^{(f)} & \boldsymbol{b}_i^{(f)} \end{bmatrix}$ and $\begin{bmatrix} \boldsymbol{I} & \boldsymbol{b}_i^{(p)} \end{bmatrix}$ are transformation matrices for "speaker adaptive training" and "pitch adaptive training" respectively. Note that "singer adaptive training" is applied to all features, although $\boldsymbol{b}_i^{(p)}$ is fixed to $\boldsymbol{0}$ except a static feature of log F_0.

6. EXPERIMENTS

Objective comparison tests of likelihood and subjective comparison test of mean opinion score (MOS) were conducted to evaluate the performance of "singer adaptive training" for HMM-based singing voice synthesis.

Japanese children's songs sung by two female singers (F001 and F002) were used. Different ten songs sung by both singers were used for training. The total length for training data of singer F001 is 7.1 minutes, and that of singer F002 is 10.3 minutes. Singing voice signals were sampled at 48kHz and windowed with a 5-ms shift. The feature vectors consisted of spectrum and excitation parameters. The vibrate parameters were not used in this experiment. The spectrum parameter vectors consisted of 49 STRAIGHT [13] mel-cepstral coefficients including the zero coefficient, their delta, and delta-delta coefficients. The excitation parameter vectors consisted of log F_0, its delta, and delta-delta.

A seven-state (including beginning and ending null states), left-to-right, no-skip structure was used for the hidden semi-Markov model (HSMM) [14]. The spectrum stream was modeled with single multi-variate Gaussian distributions. The excitation streams were modeled with multi-space probability distribution HSMM (MSD-HSMM) [15]. The state durations of each model were modeled with a five-dimensional (equal to the number of emitting states in each model) multi-variate Gaussian distribution. The decision tree-based context-clustering technique was separately applied to distributions for the spectrum, excitation, and state duration. The MDL criterion [16] was used to control the size of the decision trees, and the heuristic weight for the penalty term was 3.0.

In the objective test, models of each singer trained from 20 songs in total by using "singer adaptive training" (**PROPOSED**) were compared with speaker dependent models trained from ten songs of each singer by using "pitch adaptive training" (**CONVENTIONAL**) [10] in likelihood. Ten songs included in training data and ten songs not included in that were used for the evaluation. Figures 5 and 6 show the average log likelihood per frame for the training data set (close) and the test data set (open) to compare the two methods. The results show that the **PROPOSED** method outperformed the **CONVENTIONAL** method for both data sets. These results indicate that the quality of the model was improved in the **PROPOSED** method.

In the subjective test, ten subjects were asked to evaluate the naturalness of the synthesized singing voices on a MOS with a scale from 1 (poor) to 5 (good). Ten songs not included in the training data were used for the evaluation. Fifteen randomly selected musical phrases were presented to each subject. The experiments were carried out in a sound-proof room. Figure 7 shows the subjective listening test results[2]. In Fig. 7, the **PROPOSED** method obtained higher MOSs than the **CONVENTIONAL** method. It seems that the improvement rate of singer F001 is higher than that of F002 because the speaker-dependent training data size of F001 is smaller than that of F002. These results mean that "singer adaptive training" can synthesize more natural singing voices.

7. CONCLUSIONS

In this paper, we proposed "singer adaptive training" by integrating "speaker adaptive training" for HMM-based text-to-speech synthesis and "pitch adaptive training" for HMM-based singing voice synthesis to improve the quality of the synthesized singing voices in HMM-based singing voice synthesis. A model to synthesize the high quality singing voices is obtained from a small amount of data sung by a target singer by using "speaker adaptive training," and the quality deterioration of the singing voices caused by imbalanced of the pitch included in training data is prevented by using "pitch adaptive training." In the experiments, the models trained by using "singer adaptive training" showed higher objective and subjective evaluation scores than the speaker dependent models trained by using only "pitch adaptive training." These results show that the quality of the synthesized singing voices was improved. Although two singers' singing voice data were used in the experiments, it seems that a model to synthesize the higher quality singing voices may be obtained by increasing the number of singers. Future work involve experiments on larger data sets.

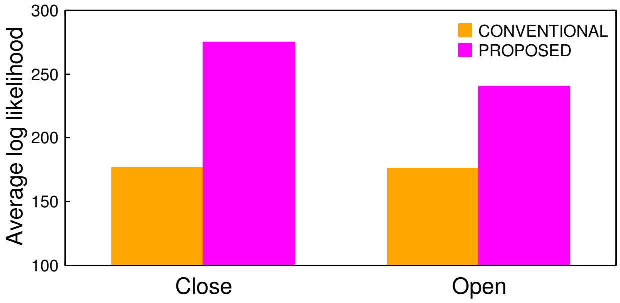

Fig. 5. *Objective evaluation results of likelihood (Singer F001).*

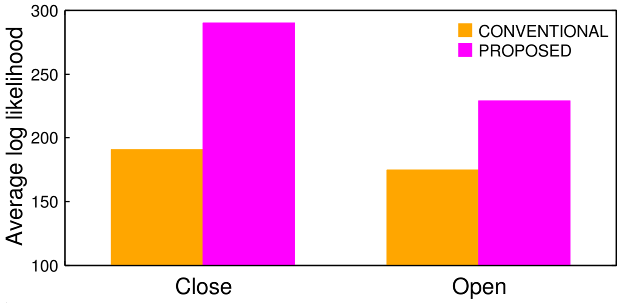

Fig. 6. *Objective evaluation results of likelihood (Singer F002).*

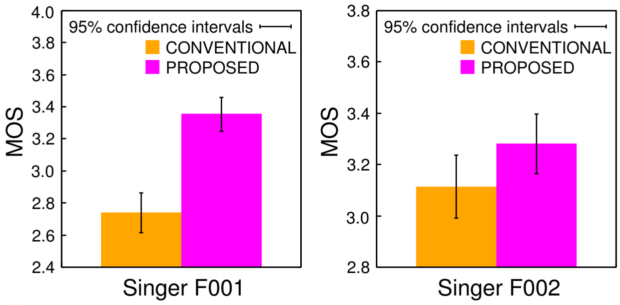

Fig. 7. Subjective evaluation results of mean opinion score.

8. ACKNOWLEDGEMENTS

This research was partly funded by Core Research for Evolutionary Science and Technology (CREST) from Japan Science and Technology Agency (JST), and the Hori Sciences and Arts Foundation.

[2] The obtained results are not comparable in absolute value across singers because these experiments were conducted independently.

9. REFERENCES

[1] 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. of Eurospeech, pp. 2347–2350, 1999.

[2] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using mllr,” in Proc. of ICASSP, pp. 805–808, 2001.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” in Proc. of Eurospeech, vol. 5, pp. 2523–2526, 1997.

[4] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Eigenvoice for HMM-based speech synthesis,” in Proc. of ICSLP, vol. 1, pp. 1269–1272, 2002.

[5] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, “An HMM-based singing voice synthesis system,” in Proc. of Interspeech, pp. 1141–1144, 2006.

[6] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, ”Recent development of the HMM-based singing voice synthesis system – Sinsy,” in Proc. the 7th ISCA Tutorial and Research Workshop on Speech Synthesis, pp. 211–216, 2010.

[7] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training,” IEICE Trans. Inf. & Syst., vol. E-90D, no. 2, pp. 533–543, 2007.

[8] Y. Qian, H. Liang, and F. K. Soong, “Generating natural F_0 trajectory with additive trees,” in Proc. of Interspeech, pp. 2126–2129, 2008.

[9] H. Zen and N. Braunschweiler, “Context- dependent additive log F_0 model for HMM-based speech synthesis,” in Proc. of Interspeech, pp. 2091–2094, 2094.

[10] K. Oura, A. Mase, Y. Nankaku, and K. Tokuda, “Pitch adaptive training for HMM-based singing voice synthesis,” in Proc. of ICASSP, pp. 5377–5380, 2012.

[11] K. Tokuda, T. Kobayashi, and S. Imai, “Speech parameter generation from HMM using dynamic features,” in Proc. of ICASSP, pp. 660–663, 1995.

[12] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” Computer Speech and Language, vol. 12, no. 2, pp. 75–98, 1998.

[13] H. Kawahara, M. K. Ikuyo, and A. Cheneigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, pp. 187–207, 1999.

[14] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “A hidden semi-Markov model-based speech synthesis system,” IEICE Trans. Inf. & Sys., vol. 90-D, no. 5, pp. 825–834, 2007.

[15] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden markov models based on multi-space probability distribution for pitch pattern modeling,” in Proc. of ICASSP, vol. 1, pp. 229–232, 1999.

[16] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” J. Acoust. Soc. Jpn. (E), vol. 21, no. 2, pp. 76–86, 2000.