Improvements to HMM-Based Speech Synthesis Based on Parameter Generation with Rich Context Models

Shinnosuke Takamichi[†], *Tomoki Toda*[†], *Yoshinori Shiga*[‡], *Sakriani Sakti*[†], *Graham Neubig*[†] *and Satoshi Nakamura*[†]

[†] Graduate School of Information Science, Nara Institute of Science and Technology, 8916–5 Takayama-cho, Ikoma, Nara, 630-0192, Japan

[‡] National Institute of Information and Communications Technology, 3–5 Hikari–dai, Seika–cho, Soraku–gun, Kyoto, 619–0289, Japan.

Abstract

In this paper, we improve parameter generation with rich context models by modifying an initialization method and further apply it to both spectral and F_0 components in HMM-based speech synthesis. To alleviate over-smoothing effects caused by the traditional parameter generation methods, we have previously proposed an iterative parameter generation method with rich context models. It has been reported that this method yields quality improvements in synthetic speech but there are still limitations. This is because 1) this generation method still suffers from the over-smoothing effect, as it uses the parameters generated by the traditional method as an initial parameters, which strongly affect on the finally generated parameters and 2) it is applied to only the spectral component. To address these issues, we propose 1) an initialization method to generate less smoothed but more discontinuous initial parameters that tend to yield better generated parameters, and 2) a parameter generation method with rich context models for the F_0 component. Experimental results show that the proposed methods yield significant improvements in quality of synthetic speech.

Index Terms: HMM-based speech synthesis, rich context models, GMM, context clustering, over-smoothing, MSD-HMM

1. Introduction

The corpus-based approach [1] to Text-To-Speech (TTS) is currently the most popular, and has two main synthesis techniques: sample-based synthesis such as unit selection synthesis [2, 3], and statistical parametric synthesis such as Hidden Markov Model (HMM)-based speech synthesis [4]. In unit selection synthesis, although high-quality speech is synthesized by the direct use of waveform segments [5], voice characteristics of the generated speech are fully dependent on the original voice. On the other hand, HMM-based speech synthesis uses well-formulated statistical parametric representation of speech parameters. Although many merit such as a flexible control of the voice characteristics [6, 7, 8] is yielded, the generated speech parameters tend to be over-smoothed, and synthetic speech sounds muffled compared with natural speech [9].

To alleviate this over-smoothing effect, some hybrid methods that lie between those two methods have been proposed [10, 11, 12]. Maximum likelihood (ML)-based unit selection synthesis [10] uses waveform segments retrieved from the speech corpus to maximize the HMM likelihood. However, the use of waveform segments makes it impossible to flexibly control voice characteristics of synthetic speech. As another hybrid method using the formulated parametric representation, rich context modeling that represents each waveform segment with a probability distribution of individual speech component parameters such as spectrum and F_0 has been proposed [11]. In synthesis, the joint probability distribution of all speech components corresponding to one waveform segment is selected. However, this method still loses the flexible control due to necessity of using a strong constraint among different speech components in synthesis.

As a hybrid method that preserves the flexibility of HMM-based speech synthesis, we have proposed a parameter generation method using rich context models [13]. The probability distributions corresponding to individual waveform segments are reformulated as GMMs separately for each speech component. A speech parameter trajectory at each component is generated based on the ML criterion using an iterative process. This generation method has been applied to the spectral component. However, while quality improvements in synthetic speech have been confirmed, the synthetic speech still sounds muffled. Because the parameter sequence generated by the iterative generation process strongly depends on the initial parameter sequence, it can be expected that a setting of a suitable initial parameter sequence will yield further improvements in the quality of synthetic speech. As another approach for quality improvements, a F_0 parameter generation method with rich context models is expected. For F_0 component, a statistical model using rich context models and the parameter generation method need to consider voiced/unvoiced region. It can also be expected for the F_0 component that the setting of the initial parameter sequence will affect quality improvements.

In this paper, we propose a technique to properly initialize parameters so that the parameter generation method using rich context models produces higher-quality speech. A less-smoothed but highly discontinuous parameter sequence is generated as an initial parameter sequence from probability distributions over-fitted to individual segments. We experimentally show that the use of this initial parameter sequence yields significant quality improvements of synthetic speech. Moreover, we propose an F_0 parameter generation method using rich context models based on the Multi-Space Distribution HMM (MSD-HMM) [14]. The experimental results demonstrate that further quality improvements are achieved by applying the proposed parameter generation method to both spectral and F_0 components.

2. HMM-based Speech Synthesis

In HMM-based speech synthesis, various contextual factors are used to capture both segmental and prosodic features. Since combinations of the contextual factors increase exponentially and the number of them is enormous, one context label (called a ”full context”) usually corresponds to only one acoustic segment in the training data.

To robustly train context-dependent HMMs, different full

Copyright © 2013 ISCA

context labels are tied together in a decision tree [15] constructed based on the Minimum Description Length (MDL) criterion [16], which is given by

$$l^{(C)} = \frac{1}{2}\sum_{c=1}^{C}\Gamma(c)\log|\boldsymbol{\Sigma}_c| + aCD\log\Gamma(0), \quad (1)$$

where c is a leaf node index, C is the total number of leaf nodes, a is a parameter to control C, D is the number of feature dimensions, $\boldsymbol{\Sigma}_c$ is the covariance matrix of leaf node c, and $\Gamma(c)$ and $\Gamma(0)$ are state occupancy counts in leaf node c and the root node, respectively. The output probability density function b_c is calculated in each leaf node. Different decision trees are constructed for individual speech components [15].

Spectral component: Spectral parameters are modeled by a continuous HMM. Its state output probability is given by

$$b_c(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \quad (2)$$

where $\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta\Delta\boldsymbol{c}_t^\top\right]^\top$ is a feature vector including static features $\boldsymbol{c}_t$ and dynamic features $\Delta\boldsymbol{c}_t$, $\Delta\Delta\boldsymbol{c}_t$, and $\boldsymbol{\mu}_c$ is the mean vector in the c-th leaf node. The Gaussian distribution with mean vector $\boldsymbol{\mu}_c$ and covariance matrix $\boldsymbol{\Sigma}_c$ is denoted as $\mathcal{N}(\cdot; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.

F_0 component: F_0 is modeled by an MSD-HMM [14]. Its state output probability is given by

$$b_c(\boldsymbol{o}_t) = \begin{cases} w_c\mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), & l_t = \mathrm{V} \\ 1 - w_c, & l_t = \mathrm{U} \end{cases}, \quad (3)$$

where l_t is a discrete voicing label that is either voiced V or unvoiced U at frame t, and w_c is the weight of the voiced space of leaf node c. Note that l_t is observable together with $\boldsymbol{o}_t$.

In synthesis, full context labels to be synthesized are clustered with the decision trees and the output probability density functions at corresponding leaf nodes are selected to form a sentence HMM. Then, a time sequence of the static feature vectors $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$ is generated by maximizing the HMM likelihood under a constraint on the relationship between static and dynamic features ($\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$, where $\boldsymbol{W}$ is the weighting matrix for calculating the dynamic features) [17]. While this method has many advantages in terms of flexibility, it has the well known problem that over-smoothing of the generated speech parameters causes significant degradation in speech quality.

3. Parameter Generation Method with Rich Context Models

3.1. Formulation of GMM Using Rich Context Models

Rich context models provide one way to alleviate the over-smoothing effect while preserving robustness of parameter estimation. In the rich context models, a mean vector is trained for each full context label and a covariance matrix is tied over different full context labels belonging to each leaf node of the decision tree [11]. The output probability density function (rich context model) of the continuous HMM for the m-th full context label in the c-th leaf node is given by

$$b_{c,m}(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c). \quad (4)$$

After training the rich context models in the same manner as in the conventional method, the output probability density in each leaf node is modeled by a GMM composed of all rich context models in the same leaf node [13] as follows :

$$b_c(\boldsymbol{o}_t) = \sum_{m=1}^{M_c}\omega_m\mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c), \quad (5)$$

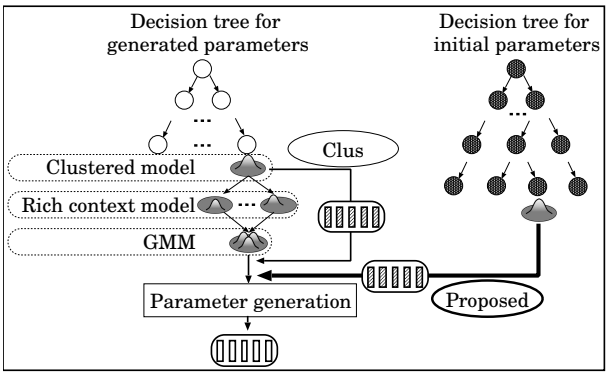

Figure 1: An overview of the proposed initialization method.

where ω_m is the mixture component weight of the m-th rich context model ($\omega_m = 1/M_c$), and the total number of mixture components is M_c.

3.2. Parameter Generation Method [13]

In synthesis, the parameter trajectory is generated to maximize HMM likelihood. Given a state sequence $\boldsymbol{q} = [q_1, \cdots, q_T]^\top$, which is determined in the traditional way [15], we approximate the HMM likelihood with a single mixture component sequence $\hat{\boldsymbol{m}} = [m_1, \cdots, m_T]$ as follows:

$$\sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{o}, \boldsymbol{m}|\boldsymbol{q}, \boldsymbol{\lambda}) \simeq P(\boldsymbol{o}, \hat{\boldsymbol{m}}|\boldsymbol{q}, \boldsymbol{\lambda}). \quad (6)$$

After setting the initial static feature vector sequence $\boldsymbol{c}^{(0)}$ to a parameter sequence generated from the clustered models in the traditional manner, the single mixture component sequence and the static feature vector sequence are iteratively updated as follows :

$$\hat{\boldsymbol{m}}^{(i+1)} = \underset{\boldsymbol{m}}{\operatorname{argmax}}\, P\left(\boldsymbol{m}|\boldsymbol{W}\hat{\boldsymbol{c}}^{(i)}, \boldsymbol{q}, \boldsymbol{\lambda}\right), \quad (7)$$

$$\hat{\boldsymbol{c}}^{(i+1)} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c}|\hat{\boldsymbol{m}}^{(i+1)}, \boldsymbol{q}, \boldsymbol{\lambda}\right). \quad (8)$$

One rich context model usually corresponds to one HMM-state acoustic segment. The model selection process (Eq. (7)) is similar to selecting a single acoustic segment sequence to generate speech parameters. The likelihoods for both static and dynamic features used in this selection could be regarded as target and concatenation costs in unit selection [18, 19].

3.3. Effect of Initial Parameter Sequence

This iterative parameter generation process easily falls into local optima, and thus the generated speech parameter sequence strongly depends on the initial parameter sequence. In the previous work, the initial parameter sequence was generated by conventional clustered models. Initial parameters generated in this way are continuous but excessively smoothed, resulting in over-smoothing even in the finally generated speech parameter sequence. Consequently, the resulting synthetic speech still sounds muffled. On the other hand, if we are able to provide a more effective initialization, it will greatly improve the overall speech quality.

4. Improved Parameter Generation Method for Spectral and F0 Components

4.1. Better Initialization

To generate a less-smoothed initial parameter sequence, we propose an initialization method with tree-based context clustering. As shown in Fig. 1, a large-sized tree for context clustering is

constructed by decreasing parameter a in Eq. (1). Note that the sufficient statistics to build this tree are the same as those used in calculating rich context models, which is calculated using the conventional clustered models.

In this tree, both the mean vector and the covariance matrix of each probability density function are calculated from only a few acoustic inventories determined by context factors. Therefore, the initial parameter estimate is less-smoothed than that generated by the conventional clustered model. It can be expected that this initial parameter sequence will help to select a less-smoothed model final sequence after the iterative parameter generation process. On the other hand, the use of a larger-sized decision tree causes over-fitting problems. In particular, the initial parameter sequence suffers from many discontinuous transitions causing the synthetic speech to sound harsh. However, these discontinuous transitions are alleviated by the use of tied covariance matrices and model selection considering the HMM likelihoods for not only static feature but also dynamic features during the iterative process described in section 3.2.

4.2. Implementation for F0 Component

The rich context models of the F_0 component are trained in the same manner as in the continuous HMMs. which is given by

$$b_{c,m}(\boldsymbol{o}_t) = \begin{cases} w_c \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right), & l_t = \mathrm{V} \\ 1 - w_c, & l_t = \mathrm{U} \end{cases}, \quad (9)$$

where the weight of the voiced space is also tied over the rich context models. A GMM in the voiced space is composed of Gaussian distributions in voiced space of all rich context models in the same leaf node

$$b_c(\boldsymbol{o}_t) = \begin{cases} \sum_{m=1}^{M_c} w_{c,m} \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right), & l_t = \mathrm{V} \\ 1 - w_c, & l_t = \mathrm{U} \end{cases}, \quad (10)$$

where $\omega_{c,m}$ is the mixture component weight of the m-th rich context model in voiced space ($\omega_{c,m} = \omega_c / M_c$). We can calculate the ML estimate of $\omega_{c,m}$ based on the occupancy counts but we set it to an constant value based on our previous findings that the constant weight setting is effective in the spectral component.

In the parameter generation, the initial parameter sequence is determined by the clustered models with a large-sized decision tree and unvoiced/voiced decision is performed based on these models. Then, the selection of rich context models and the generation of a parameter sequence is iteratively performed.

5. Experimental Evaluations

5.1. Experimental Conditions

In the experiments, we trained a context-dependent phoneme Hidden Semi-Markov Model (HSMM) [20] for a Japanese female speaker. We used 450 sentences for training and 53 sentences for evaluation from the ATR Japanese speech database [21]. Speech signals were sampled at 16 kHz. The 0th-through-24th mel-cepstral coefficients were extracted as spectral parameters and log-scaled F_0 and 5 band-aperiodicity [22] were extracted as excitation parameters by STRAIGHT [23]. The feature vector consisted of spectral and excitation parameters and their delta and delta-delta features. 5-state left-to-right HSMMs were used. In synthesis, global variance (GV) [24] was not considered.

We conducted 3 kinds of experimental evaluation. First, we investigate the effectiveness of the proposed initialization method by applying it to the spectral component and comparing it with the conventional method described in Section 3. Second, we investigate effectiveness of the proposed F_0 parameter generation method with the rich context models. Finally, we investigate the effectiveness of applying the proposed parameter generation to both spectral and F_0 components. Conventional clustered models were used for duration and aperiodic components in all evaluations.

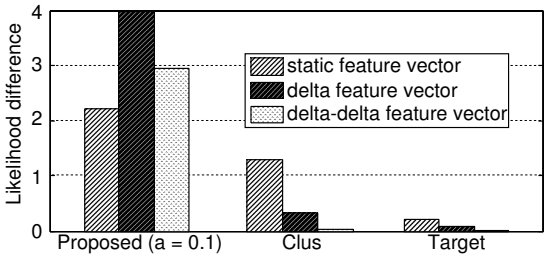

Figure 2: Differences of HMM likelihood between before and after iteration.

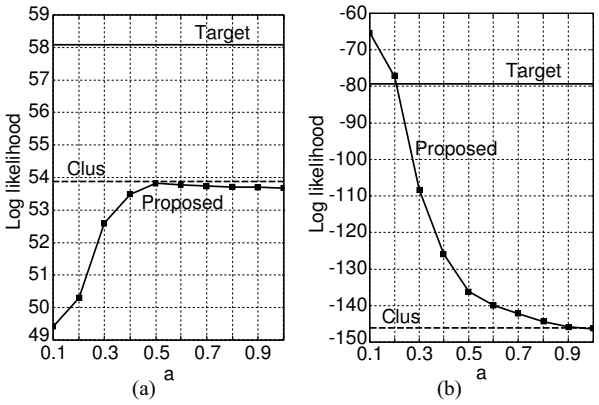

Figure 3: Results of objective evaluations of (a) HMM likelihood of selected rich context models for natural parameters, (b) GV likelihood for generated parameters.

5.2. Effectiveness of Initialization Method

5.2.1. Confirmation of Alleviating Discontinuous Transition

First, we performed a preliminary experiment to confirm whether or not the iterative parameter generation effectively alleviates the discontinuous transitions in the initial parameter sequence, we evaluated 3 settings of the initial parameter sequences: 1) Clus: generated from the conventional clustered models, 2) Proposed ($a = 0.1$): generated with a large-sized decision tree ($a = 0.1$), and 3) Target: natural target speech parameter sequence as a reference. The difference of HMM likelihoods for the generated parameters between the initially selected rich context model sequence and the finally selected one was calculated for each static and dynamic features in the spectral parameters.

The result of the likelihood differences yielded by the iterative parameter generation is shown in Fig. 2. We can see that the HMM likelihood for dynamic features of "$a = 0.1$" increases more than that of the other initial parameter sequence. From this result, we can see that the discontinuous transitions in the initial parameter sequence are alleviated by the iterative parameter generation.

5.2.2. Objective Evaluation

To investigate the effectiveness of the proposed initialization method, we evaluated 3 settings of initial parameters: 1) Clus, 2) Proposed ($a = 0.1, 0.2, \cdots, 1.0$), and 3) Target. The rich

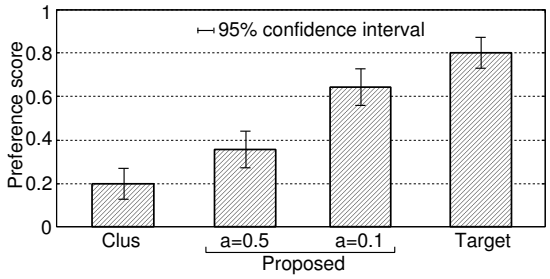

Figure 4: Preference scores on speech quality (Spectrum).

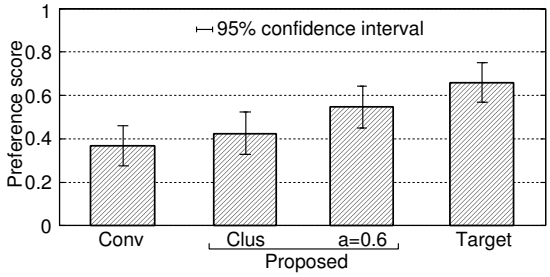

Figure 5: Preference scores on speech quality (F_0).

context model sequences selected by the parameter generation method using these initial parameter settings were evaluated with two criteria that are well-known to measure the improvement in speech quality: the HMM likelihood of the selected rich context models for the natural speech parameters, and the GV likelihood [24] for the parameter sequence generated by the selected rich context models.

The result of HMM likelihood is shown in Fig. 3(a) and that of GV likelihood is shown in Fig. 3(b). It is observed from Fig. 3(a) that the HMM likelihood of "Proposed" slightly increases as the parameter a decreases from 1.0 to 0.5, and it rapidly decreases as the parameter a decreases further. We can see that the HMM likelihood at $a = 0.5$ is almost the same as that of "Clus" but it is significantly lower than that of "Target." On the other hand, It is observed from Fig. 3(b) that the GV likelihood of "Proposed" rapidly increases as the parameter a decreases, and its value at $a = 0.1$ is higher than that of "Target."

5.2.3. Subjective Evaluation

A preference test (AB test) on speech quality was conducted using 4 types of synthetic speech, "Clus," "Proposed ($a = 0.1$)," "Proposed ($a = 0.5$)," and "Target." Every pair of these 4 types of synthetic speech was presented to 7 listeners in random order. Listeners were asked which sample sounds better in terms of speech quality.

The result of the preference test is shown in Fig. 4. The proposed initialization method significantly improves speech quality compared with the conventional initialization method "Clus." We can also see that the score of "Proposed ($a = 0.1$)" is higher than "Proposed ($a = 0.5$)." This tendency is the same as observed in the GV likelihood shown in Fig. 3 (b).

5.3. Effectiveness of Proposed F0 Generation

To investigate the effectiveness of the proposed generation method for F_0, we compared speech quality of speech generated by the proposed method with that generated by conventional parameter generation algorithm with clustered models. We evaluated 4 kinds of synthetic speech: 1) Conv: generated from clustered models, 2) Clus: generated using the parameter sequence of "Conv" as the initial parameters in the proposed method, 3) Proposed: generated using the large-sized decision tree ($a = 0.6$) for the initial parameter generation in the proposed method, 4) generated using natural target speech parameters as the initial parameters in the proposed method. We set the parameter a to 0.6 because it was observed that the GV likelihood is highest at this parameter setting in preliminary experiment. A preference test (AB test) on speech quality was conducted by 6 listeners in the same manner as in section 5.2.

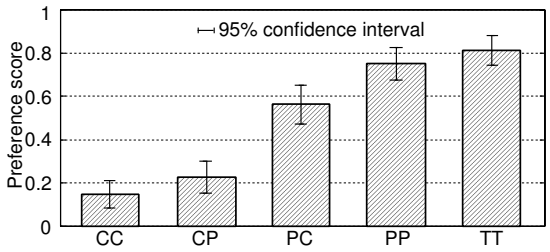

Figure 6: Preference scores on speech quality (Full synthesis).

Table 1: Synthetic speech samples used for "Full synthesis" evaluation. "Target" is generated by parameter generation with rich context models using natural speech parameter sequence as initial parameter.

method	Spectrum	F_0
CC	Conventional	Conventional
CP	Conventional	Proposed ($a = 0.6$)
PC	Proposed ($a = 0.1$)	Conventional
PP	Proposed ($a = 0.1$)	Proposed ($a = 0.6$)
TT	Target	Target

The result of the preference test is shown in Fig. 5. It is observed the score of "Proposed ($a = 0.6$)" is higher than that of "Conv." The proposed parameter generation method is also effective even for the F_0 component.

5.4. Evaluation in Full Synthesis

To investigate the effectiveness of all proposed methods, we evaluated 5 kinds of synthetic speech shown in Table 1. A preference test (AB test) on speech quality was conducted by 8 listeners in the same manner as in the section 5.2.

The result of the preference test is shown in Fig. 6. It is observed that a larger speech-quality improvement is yielded by the proposed method for the spectral component than for the F_0 component. Moreover, a further improvement is yielded by applying the proposed method to both spectral and F_0 components, and the resulting speech quality shown as "PP" is close to "TT." From this result, we can see that the proposed parameter generation with rich context models for spectral and F_0 components is very effective to improve quality in synthetic speech and close to the upper bound of initialization using target speech parameters.

6. Summary

In this paper, we improved a parameter generation method with rich context models by introducing a new parameter initialization method and applied it to both spectral and F_0 components in HMM-based speech synthesis. The experimental results demonstrated that the proposed method yields significant improvements in synthetic speech quality. As future work, we will study adaptation techniques using rich context models.

Acknowledgements: Part of this work was supported by JSPS KAKENHI Grant Number 24240032, and was executed under the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

7. References

[1] Y. Sagisaka, “Speech Synthesis by Rule Using an Optimal Selection of Non-Uniform Synthesis Units,” Proc. ICASSP, pp. 679–682, 1988.

[2] N. Iwahashi, N. Kaiki, and Y. Sagisaka, “Speech Segment Selection for Concatenative Synthesis Based on Spectral Distortion Minimization,” IEICE Trans., Fundamentals, Vol. E76-A, No. 11, pp. 1942–1948, 1993.

[3] A. J. Hunt and A.W. Black, “Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database,” Proc. ICASSP, pp. 373–376, 1996.

[4] H. Zen, K. Tokuda, and A. Black, “Statistical Parametric Speech Synthesis,” Speech Commun., Vol. 51, No. 11, pp. 1039–1064, 2009.

[5] A. K. Syrdal, C. W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K-S. Lee, and M.J. Makashay, “Corpus-based techniques in the AT&T NextGen synthesis system,” Proc. ICSLP, Vol. 3, pp. 410–415, 2000.

[6] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, “Speaker Interpolation for HMM-based Speech Synthesis System”, J. Acoust. Soc. Jpn. (E), Vol. 21, No. 4, pp. 199–206, 2000.

[7] J. Yamagishi, and T. Kobayashi, “Average-voice-based speech Synthesis Using HSMM-based Speaker Adaptation and Adaptive Training,” IEICE Trans., Inf. and Syst., Vol. E90-D, No. 2, pp. 533–543, 2007.

[8] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, “A Style Control Technique for HMM-based Expressive Speech Synthesis,” IEICE Trans., Inf. and Syst., Vol. E90-D, No. 9, pp. 1406–1413, 2007.

[9] S. King, V. Karaiskos, “The Blizzard Challenge 2011,” Proc. Blizzard Challenge workshop, 2011.

[10] Z. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang, Y. Jiang, Z. Zhao, J. Yang, J. Chen, and G. Hu, “The USTC and iflytek Speech Synthesis Systems for Blizzard Challenge 2007,” Proc. Blizzard Challenge workshop, 2007.

[11] Z. Yan, Q. Yao, and S. K. Frank, “Rich Context Modeling for High Quality HMM-Based TTS,” Proc. INTERSPEECH, 2009.

[12] Y. Qian, Z. Yan, Y. Wu, and F. K. Soong, “An HMM Trajectory Tiling (HTT) Approach to High Quality TTS,” Proc. INTERSPEECH, 2010.

[13] S. Takamichi, T. Toda, Y. Shiga, H. Kawai, S. Sakti, and S. Nakamura, “An Evaluation of Parameter Generation Methods with Rich Context Models in HMM-Based Speech Synthesis,” Proc. INTERSPEECH, 2012.

[14] K. Tokuda, T. Masuko, B. Miyazaki, and T. Kobayashi, “Multi-Space Probability Distribution HMM,” IEICE Trans., Inf. and Syst., Vol. E85-D, No. 3, pp. 455–464, 2002.

[15] 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.

[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. 79–86, 2000.

[17] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech Parameter Generation Algorithms for HMM-based Speech Synthesis,” Proc. ICASSP, pp. 1315–1318, 2000.

[18] S. Kataoka, N. Mizutani, K. Tokuda, and T. Kitamura, “ Decision Tree Backing-off in HMM-based Speech Synthesis, ” Proc. INTERSPEECH, 2004.

[19] Z. Ling, and R. Wang, “HMM-based unit selection using frame sized speech segments,” Proc. INTERSPEECH, 2006.

[20] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Hidden Semi-Markov Model Based Speech Synthesis System,” IEICE Trans., Inf. and Syst., E90-D, No. 5, pp. 825–834, 2007.

[21] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuwahara, “A large-scale Japanese speech database,” ICSLP90, pp. 1089–1092, 1990.

[22] H. Kawahara, Jo Estill and O. Fujimura, ” Aperiodicity Extraction and Control Using Mixed Mode Excitation and Group Delay Manipulation for a High Quality Speech Analysis, Modification and Synthesis System STRAIGHT ” , MAVEBA 2001, 2001.

[23] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, “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 Commun., Vol. 27, No. 3–4, pp. 187–207, 1999.

[24] T. Toda, and K. Tokuda, “A Speech Parameter Generation Algorithm Considering Global Variance for HMM-based Speech Synthesis,” IEICE Trans., Vol. E90-D, No. 5, pp. 816–824, 2007.