INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

An Evaluation of Parameter Generation Methods with Rich Context Models in HMM-Based Speech Synthesis

Shinnosuke Takamichi[1], *Tomoki Toda*[1], *Yoshinori Shiga*[2],
Hisashi Kawai[2], *Sakriani Sakti*[1], *Satoshi Nakamura*[1]

[1] Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan
[2] National Institute of Information and Communications Technology, Japan
shinnosuke-t@is.naist.jp, tomoki@is.naist.jp

Abstract

In this paper, we propose parameter generation methods using rich context models in HMM-based speech synthesis as yet another hybrid method combining HMM-based speech synthesis and unit selection synthesis. In the traditional HMM-based speech synthesis, generated speech parameters tend to be excessively smoothed and they cause muffled sounds in synthetic speech. To alleviate this problem, several hybrid methods have been proposed. Although they significantly improve quality of synthetic speech by directly using natural waveform segments, they usually lose flexibility in converting synthetic voice characteristics. In the proposed methods, rich context models representing individual acoustic parameter segments are reformed as GMMs and a speech parameter sequence is generated from them using the parameter generation algorithm based on the maximum likelihood criterion. Since a basic framework of the proposed methods is still the same as the traditional framework, the capability of flexibly modeling acoustic features remains. We conduct several experimental evaluations of the proposed methods from various perspectives. The experimental results demonstrate that the proposed methods yield significant improves in quality of synthetic speech.

Index Terms: HMM-based speech synthesis, over-smoothing, rich context model, parameter generation

1. Introduction

Many attempts at developing a technique for converting text into speech, i.e., Text-To-Speech (TTS), have been studied for several decades. It is no doubtful that the corpus-based approach [1] has yielded dramatic improvements of TTS. In the corpus-based approach, there are two main synthesis techniques, i.e., sample-based synthesis and statistical parametric synthesis. The sample-based synthesis such as unit selection [2] directly uses acoustic inventories selected from a speech corpus for synthesizing a speech waveform. One of the main advantages of the unit selection is that high quality speech keeping original voice characteristics is synthesized by concatenating natural acoustic segments. However, characteristics of the generated speech are fully dependent on original voices.

On the other hand, the statistical parametric synthesis methods, such as HMM-based speech synthesis, use averaged acoustic inventories extracted from the speech corpus. In HMM-based speech synthesis, spectrum, pitch, and duration are modeled simultaneously in a unified framework of HMMs. In synthesis these parameters are generated from HMMs under a maximum likelihood (ML) criterion by using dynamic features [3]. One of the biggest advantages of this method is the capability of flexibly controlling voice characteristics, e.g., speaker-individuality control [4][5], and speaking-style control [6]. However, the generated speech parameters tend to be over-smoothed, and synthetic speech evidently sounds muffled compared with natural speech because detailed characteristics of speech parameters are often removed in the statistical process.

In order to alleviate this over-smoothing effect, hybrid methods between those two main methods have been proposed. Suitable waveform segments are searched out from the speech corpus to maximize the HMM likelihood [7]. The use of waveform segments dramatically improves speech quality. However, it loses a strong advantage of HMM-based speech synthesis of controlling voice characteristics. As one of the hybrid approaches having better flexibility than unit selection, the use of rich context models to represent each waveform segment with probability distributions of individual speech component parameters, such as spectrum and F_0, has been proposed [9]. In synthesis part, the probability distributions of all components corresponding to one waveform segment are selected in each HMM-state and speech parameters are generated from them. This method also yields significant improvements in speech quality. However, efficient and flexible acoustic modeling in the original HMM-based speech synthesis is lost since this method needs to use a strong constraint among different components in the selection of their probability distributions.

In this paper, we propose parameter generation methods for another hybrid approach using the rich context models to keep flexible property in synthesis part. The trained rich context models are reformed as a Gaussian mixture model (GMM) in each HMM-state. A speech parameter trajectory in each component is separately generated from the corresponding GMMs using the ML criterion. The proposed methods also enables to effectively use probability distributions of individual components from different waveform segments as in the original HMM-based speech synthesis. We conducted several experimental evaluations from various perspectives to demonstrate the effectiveness of the proposed parameter methods.

This paper is organized as follows. In section 2, traditional HMM-based speech synthesis framework is briefly reviewed. In section 3, the rich context modeling is described. In section 4, the proposed parameter generation methods are described. In section 5, the experimental evaluation results and some discussions are given. Section 6 presents conclusions of this paper.

2. Traditional Framework

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 enormously large, one context label (called ”full context”) usually corresponds to only one acoustic segment in training data. To robustly train context-dependent HMMs, different full context labels are tied together in a decision tree structure and an output probability density function b_c is calculated in each leaf node of the decision tree in training part [10], which is given by

$$b_c\left(\boldsymbol{o}_t\right) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right), \tag{1}$$

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 feature, $\boldsymbol{c}_t$, and dynamic features, $\Delta\boldsymbol{c}_t,\ \Delta\Delta\boldsymbol{c}_t$. The Gaussian distribution is denoted as $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right)$, where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are a mean vector and a covariance matrix in the c-th leaf node, respectively.

In synthesis part, full context labels to be synthesized are clustered with the decision tree and 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 explicit relationship between static and dynamic features ($\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$) [8]. It is well known that the generated speech parameters are over-smoothed and this over-smoothing effect causes significant degradation in synthetic speech quality.

3. Rich Context Modeling

In the traditional approach, a single Gaussian distribution is used to model multiple acoustic segments belonging to the same leaf nodes in the decision tree. Consequently its mean vector is excessively smoothed and it becomes one of the factors causing the over-smoothing effect. On the other hand, the use of multiple acoustic segments is essential to robustly estimate its covariance matrix. To alleviate the over-smoothing effect while keeping robustness of parameter estimation high enough, in the rich context model 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 [9]. The output probability density function for the m-th full context label in the c-th leaf node is given by

$$b_{c,m}\left(\boldsymbol{o}_t\right) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right). \tag{2}$$

The total number of different mean vectors is equivalent to the number of full context labels in training data. The total number of different covariance matrices is equivalent to the number of leaf nodes in the decision tree.

In training part, the context-clustered probability density parameters are trained in a traditional way. Then, they are untied and only their mean vectors are further updated in every full context label using forward-backward algorithm while tying the covariance matrices over full context labels in each leaf node. In synthesis part, probability density functions for all components are jointly selected based on KL divergence between traditional context-clustered model and the rich context model while avoiding selecting the probability density functions for individual components from different acoustic segments. This process is regarded as unit selection but each acoustic segment is represented by probability density functions in individual components. Finally, speech parameter trajectories are generated from the selected probability density functions.

4. Parameter Generation Method Using Rich Context Models

We propose parameter generation methods for selecting the rich context models based on the ML criterion. The proposed methods keep flexibility in acoustic modeling of different speech component parameters in the traditional framework, which is lost in the conventional synthesis method described in section 3. The training and synthesis processes are shown in Figure 1.

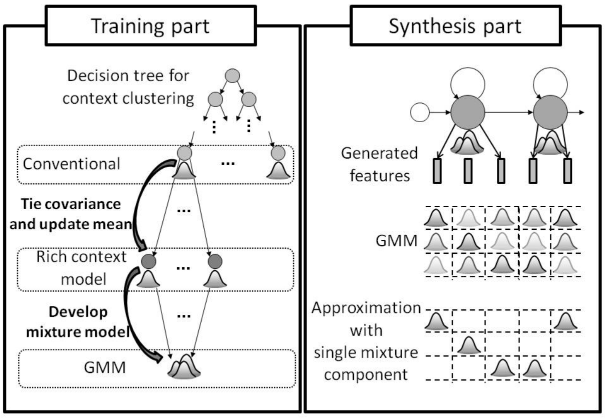

Training part
Decision tree for context clustering
Conventional
Tie covariance and update mean
Rich context model
Develop mixture model
GMM
Synthesis part
Generated features
GMM
Approximation with single mixture component

Figure 1: Training and synthesis processes in proposed methods

4.1. Representation as GMM

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 developed using all rich context models in the same leaf node as follows:

$$b_c\left(\boldsymbol{o}_t\right) = \sum_{m=1}^{M_c} \omega_m \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right), \tag{3}$$

where ω_m is the mixture component weight of the m-th rich context model, and the total number of mixture components is M_c. We can calculate the ML estimate of ω_m based on the occupancy counts given by forward-backward algorithm but in this paper we set it to an equivalent value ($\omega_m = 1/M_c$) over different mixture components since we have found this weight setting yields slight quality improvements in synthetic speech.

4.2. Parameter Generation

Given a state sequence $\boldsymbol{q} = \left[q_1, \cdots, q_T\right]^\top$, which is determined in a traditional way, the HMM likelihood is written as

$$P\left(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}\right) = \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{o}, \boldsymbol{m}|\boldsymbol{q}, \boldsymbol{\lambda}\right), \tag{4}$$

where $\boldsymbol{m} = \left[m_1, \cdots, m_T\right]^\top$, $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top$, and $\boldsymbol{\lambda}$ are a mixture component sequence (i.e., a rich context model sequence), a feature vector sequence, and an HMM parameter set, respectively. The static feature vector sequence is determined by maximizing the HMM likelihood under the constraint ($\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$) as in the traditional parameter generation process [8] as follows:

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}} \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{o}, \boldsymbol{m}|\boldsymbol{q}, \boldsymbol{\lambda}\right). \tag{5}$$

4.2.1. EM algorithm

The ML estimate of $\boldsymbol{c}$ is determined with expectation-maximization (EM) algorithm. First, an initial static feature vector sequence $\boldsymbol{c}^{(0)}$ is determined. Then, the following auxiliary function is maximized by iteratively updating the posterior probability $P\left(\boldsymbol{m}|\boldsymbol{W}\boldsymbol{c}^{(i)}, \boldsymbol{q}, \boldsymbol{\lambda}\right)$ given a current estimate $\boldsymbol{c}^{(i)}$ in E-step and a new estimate $\hat{\boldsymbol{c}}^{(i+1)}$ while fixing it constant in

M-step :

$$Q\left(\boldsymbol{c}^{(i)}, \boldsymbol{c}^{(i+1)}\right) = \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{W}\boldsymbol{c}^{(i)}, \boldsymbol{q}, \boldsymbol{\lambda}\right) \ln P\left(\boldsymbol{W}\boldsymbol{c}^{(i+1)}, \boldsymbol{m}|\boldsymbol{q}, \boldsymbol{\lambda}\right). \quad (6)$$

4.2.2. Approximation with a single mixture component sequence

We approximate the HMM likelihood given in Eq. (4) with a single mixture component sequence as follows:

$$\sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{o}, \boldsymbol{m}|\boldsymbol{q}, \boldsymbol{\lambda}\right) \simeq P\left(\boldsymbol{o}, \boldsymbol{m}|\boldsymbol{q}, \boldsymbol{\lambda}\right). \quad (7)$$

After determining the initial static feature vector sequence $\boldsymbol{c}^{(0)}$, the single mixture component sequence and the static feature vector sequence is iteratively updated as follows :

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

$$\hat{\boldsymbol{c}}^{(i+1)} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c}|\hat{\boldsymbol{m}}^{(i+1)}, \boldsymbol{q}, \boldsymbol{\lambda}\right). \quad (9)$$

4.2.3. Discussion

One rich context model usually corresponds to one HMM-state acoustic segment. Therefore, the proposed processes are strongly related to unit selection. In the proposed methods, the HMM likelihood for the static features and that for the dynamic features are regarded as a target cost and a concatenation cost, respectively [11]. The synthesis process with EM algorithm is similar to the process of selecting multiple acoustic segments and mixing them to generate speech parameters [12]. On the other hand, the synthesis process with a single mixture component sequence is similar to the process of selecting a single acoustic segment sequence to generate speech parameters [2].

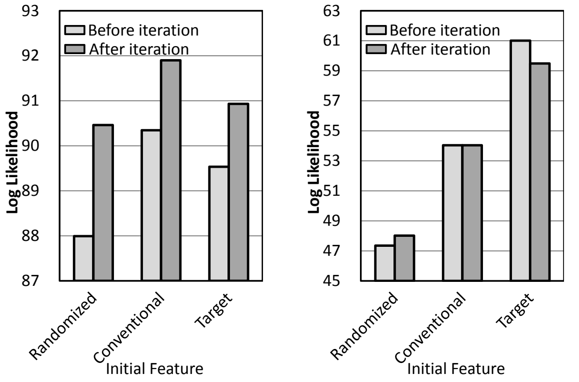
The proposed methods don't have to use the constraint used in the conventional selection method of the rich context models. Therefore, the proposed methods keep an advantage of flexible acoustic modeling in the traditional HMM-based speech synthesis. For instance, it is possible to separately search the best rich context model sequences for different speech component parameters to more widely cover a joint acoustic space. It is also straightforward to apply different speech parameter generation methods to individual speech component parameters.

In the proposed methods, the rich context models are selected frame by frame. We can also select them state by state by using an additional constraint that the same rich context model is selected within the same HMM-state. Moreover, there are some ways to determine the initial feature vector sequence. One of the reasonable ways is to use the sequence generated by the context-clustered HMMs in the traditional manner.

5. Experimental Evaluations

5.1. Experimental Conditions

We trained the context-dependent HSMM for a Japanese female speaker in a standard manner. We used 450 sentences for phonetically balanced 503 sentences from ATR Japanese speech database [13] for training, and 53 sentences for evaluation. Speech signals were sampled at 16 kHz. Mel-cepstral coefficients were extracted by STRAIGHT [14]. The shift length was set to 5 ms. The feature vector consisted of 25 mel-cepstral coefficients including the zeroth coefficient as a spectral parameter and both log-scaled F_0 and 5 band-aperiodicity as excitation parameters. We used 5-state left-to-right HSMMs. The rich context models were trained for only spectral component in this paper. In synthesis, a global variance [15] was not considered .



(a) Likelihood for generated parameter

(b) Likelihood for natural parameter

Figure 2: Effect of initial parameter sequence

We conducted two kinds of experimental evaluations. First, we investigated the proposed methods from various perspectives. To clarify the effectiveness of the proposed synthesis process in the spectral component, natural state duration determined by the state-level forced alignment with the conventional context-clustered models was used in the first evaluation. In the other evaluation, the proposed method was evaluated in fully synthesis process including duration, F_0, 5-band aperiodicity, and spectral parameter generation. Note that the conventional context-clustered models were used for duration, F_0, and 5-band aperiodicity in all evaluations.

5.2. Evaluations in Natural State Duration

5.2.1. Effect of Initial Parameter Sequence

To investigate the effect of the initial parameter sequence on the finally generated speech parameter sequence, we evaluated three settings of the initial parameter sequence; 1) Randomized: generated from rich context models randomly selected in individual leaf nodes, 2) Conventional: generated from the conventional context-clustered models, and 3) Target: generated from rich context models selected by maximizing the likelihood for natural target speech parameters. We used the approximation of a single mixture component sequence in the proposed method. An initially selected rich context model sequence and a finally selected rich context model sequence were evaluated with not only the likelihood for the generated speech parameters but also that for natural speech parameters.

The result is shown in Figure 2. It is reasonable that the likelihood for the generated speech parameters increases through the iteration as shown in Figure 2(a). On the other hand, the likelihood for the natural speech parameters does not always increase through the iteration and its value strongly depends on the initial parameter sequence as shown in Figure 2(b). We can also see that the likelihood differences in Figure 2(b) is much larger than those in Figure 2(a). From these results, it is shown that the setting of the initial parameter sequence is essential and it is difficult to find the best rich context models using the likelihood measure.

5.2.2. Comparison of Proposed Parameter Generation Methods

To evaluate proposed methods, we compared synthetic speech generated by the conventional clustered model (Conventional), the proposed method with EM algorithm (Proposed(GMM)), the proposed method with a single mixture component sequence (Proposed(single)), and the single mixture component sequence

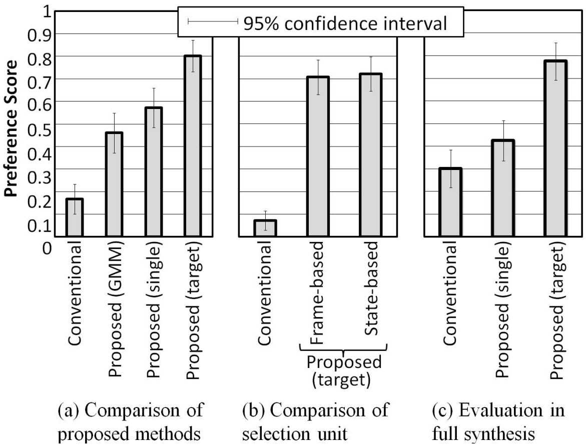


Figure 3: Preference Scores for Speech Quality

selected by the natural speech parameters as a reference (Proposed(target)). The initial parameter sequence in the proposed methods was generated by ”Conventional”. A preference test (AB test) on speech quality was conducted. Every pair of these four types of synthetic speech was presented to seven listeners in random order. Listeners were asked which sample sounds better in terms of speech quality.

The result is shown in Figure 3(a). The proposed methods significantly improve speech quality. Moreover, the use of a single mixture component sequence yields better speech quality than the use of EM algorithm. We can also see that the best rich context model sequence, which is well approximated with ”Proposed(target)”, is difficult to select using the likelihood measure. This result is consistent with the result shown in Figure 2(b).

5.2.3. Comparison of Selection Unit

To investigate the effect of the selection unit in the proposed method, i.e., frame-based selection or state-based selection, we compared synthetic speech generated by the conventional clustered model (Conventional), the proposed method with a single mixture component sequence selected frame by frame (Frame-based), and that selected state by state (State-based). In the selection process of the proposed method, the natural speech samples were used as the target. We confirmed that the mixture component sequences selected by these two methods were different from each other.

The result is shown in Figure 3(b). We can see that there is no significant difference between the frame-based selection and the state-based selection and the state-based selection is also effective for improving synthetic speech quality.

5.3. Evaluation in Full Synthesis

To confirm the effectiveness of the proposed method in fully synthesis process where all speech parameters were generated from the models, we compared synthetic speech generated by the conventional clustered model (Conventional), the proposed method with a single mixture component sequence selected state by state (Proposed(single)), and the single mixture component sequence selected state by state using the natural target speech parameters as a reference (Proposed(target)). The initial parameter sequence in the proposed method was generated by ”Conventional”. A preference test (AB test) on speech quality was conducted in the same manner as in Section 5.2.2.

The result is shown in Figure 3(c). It is observed that the proposed method yields significant quality improvements in synthetic speech even in fully synthesis process. We can find that the difference between ”Proposed(single)” and ”Proposed(target)” is larger in Figure 3(c) than that in Figure 3(a). We will investigate what causes this difference in future work.

6. Conclusions

In this paper, we have proposed parameter generation methods using rich context models in HMM-based speech synthesis to improve quality of synthetic speech while keeping capability of flexibly modeling acoustic features. The rich context models are reformed as the GMM and the likelihood measure is used in speech parameter generation. Various experimental results have demonstrated the effectiveness of the proposed methods. However, it has also shown that suitable rich context models are not selected by using the likelihood as a selection measure. As future work, we will improve the selection measure.

Acknowledgements: Part of this research was executed under the Commissioned Research for ”Research and Development on Medical Communication Support System for Asian Languages based on Knowledge and Language Grid” 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, 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] H. Zen, K. Tokuda, A. Black, ”Statistical parametric speech synthesis,” Speech Commun., Vol.51, No.11, pp.1039–1064, 2009.

[4] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, T. Kitamura, ”Speaker interpolation for HMM-based speech synthesis system”, J. Acoust. Soc. Jpn. (E), Vol.21, No.4, pp.199–206, 2000.

[5] J. Yamagishi, T. Kobayashi, ”Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training,” IEICE Trans. on Inf. and Syst., Vol.E90-D, No.2, pp.533–543, 2007.

[6] T. Nose, J. Yamagishi, T. Masuko, 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.

[7] Z. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang, Y. Jiang, Z. Zhao, J. Yang, J. Chen, G. Hu, ”The USTC and iflytek speech synthesis systems for Blizzard Challenge 2007,” Proc. of Blizzard Challenge workshop, 2007.

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, ”Speech parameter generation algorithms for HMM-based speech synthesis,” Proc.ICASSP, pp.1315–1318, 2000.

[9] Z. Yan, Q. Yao, S.K. Frank, ”Rich Context Modeling for High Quality HMM-Based TTS,” INTERSPEECH 2009, pp.1755–1758, 2009.

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

[11] S. Kataoka, N. Mizutani, K. Tokuda, T. Kitamura, “ Decision tree backing-off in HMM-based speech synthesis, ” INTERSPEECH 2004, WeB1403p.12, 2004.

[12] T. Mizutani, T. Kagoshima, ”Concatenative Speech Synthesis Based on the Plural Unit Selection and Fusion Method”, IEICE Trans. on Inf. and Syst., Vol. E88-D, No.11, pp.2565–2572, 2005.

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

[14] H. Kawahara, I. Masuda-Katsuse, 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 souds,” Speech Commun., Vol.27, No.3–4, pp.187–207, 1999.

[15] T. Toda, 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.