Parameter Generation Methods with Rich Context Models for High-Quality and Flexible Text-To-Speech Synthesis

Shinnosuke Takamichi, *Nonmember, IEEE*, Tomoki Toda, *Member, IEEE*, Yoshinori Shiga, *Nonmember, IEEE*, Sakriani Sakti, *Member, IEEE*, Graham Neubig, *Nonmember, IEEE*, and Satoshi Nakamura *Senior Member, IEEE*.

***Abstract*—In this paper, we propose parameter generation methods using rich context models as yet another hybrid method combining Hidden Markov Model (HMM)-based speech synthesis and unit selection synthesis. Traditional HMM-based speech synthesis enables flexible modeling of acoustic features based on a statistical approach. However, the speech parameters tend to be excessively smoothed. To address this problem, several hybrid methods combining HMM-based speech synthesis and unit selection synthesis have been proposed. Although they significantly improve quality of synthetic speech, they usually lose flexibility of the original HMM-based speech synthesis. In the proposed methods, we use rich context models, which are statistical models that represent individual acoustic parameter segments. In training, the rich context models are reformulated as Gaussian Mixture Models (GMMs). In synthesis, initial speech parameters are generated from probability distributions over-fitted to individual segments, and the speech parameter sequence is iteratively generated from GMMs using a parameter generation method based on the maximum likelihood criterion. Since the basic framework of the proposed methods is still the same as the traditional framework, the capability of flexibly modeling acoustic features remains. The experimental results demonstrate: (1) the use of approximation with a single Gaussian component sequence yields better synthetic speech quality than the use of EM algorithm in the proposed parameter generation method, (2) the state-based model selection yields quality improvements at the same level as the frame-based model selection, (3) the use of the initial parameters generated from the over-fitted speech probability distributions is very effective to further improve speech quality, and (4) the proposed methods for spectral and F_0 components yields significant improvements in synthetic speech quality compared with the traditional HMM-based speech synthesis.**

***Index Terms*—HMM-based speech synthesis, rich context model, GMM, parameter generation, over-smoothing**

I. Introduction

TEXT-To-Speech (TTS) is a technology that converts any text into speech, and it plays an important role in many speech applications. Many TTS techniques have been studied for several decades. Recently, TTS systems are constructed nearly automatically using pre-recorded speech. In general, this type of TTS is called corpus-based TTS [1]. The corpus-based approach has yielded dramatic improvements of TTS as researchers have been able to easily share common knowledge and findings. In the corpus-based approach, there are two main synthesis techniques: sample-based synthesis and statistical parametric synthesis. Sample-based synthesis such as unit selection [2], [3] directly uses acoustic inventories selected from a speech corpus for synthesizing speech waveforms. As shown in Fig. 1, one of the main advantages of unit selection is the ability to synthesize that high-quality speech keeping original voice characteristics by concatenating natural acoustic segments [4]. However, characteristics of the generated speech are fully dependent on original voices.

On the other hand, statistical parametric synthesis methods, such as Hidden Markov Model (HMM)-based speech synthesis [5], 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 the maximum likelihood (ML) criterion by using temporally dynamic features. One of the biggest advantages of this method is the capability of flexibly modeling and controlling acoustic features, e.g., speaker-individuality control [6], [7] and speaking-style control [8]. However, the generated speech parameters tend to be over-smoothed, and synthetic speech sounds muffled compared with natural speech because detailed characteristics of speech parameters are often smoothed out in the statistical process [9]. Consequently, quality of speech synthesized by HMM-based speech synthesis is still significantly lower than that synthesized by unit selection [10].

To address some problems of the sample-based synthesis method and the statistical parametric synthesis method, such as difficulties of automatically tuning cost functions for selecting waveform segments in unit selection synthesis or the over-smoothing effect in HMM-based speech synthesis, some hybrid methods combining these two methods have been proposed [11], [12], [13], as also shown in Fig. 1. ML-based unit selection synthesis [11] is one of the hybrid methods to improve quality of synthetic speech. Suitable waveform segments are searched out from the speech corpus to maximize the likelihood of an HMM that is automatically trained using a speech corpus. The use of waveform segments dramatically improves speech quality compared with that in HMM-based speech synthesis. However, it loses a strong advantage of

S. Takamichi, T Toda, S Sakti, G Neubig, and S Nakamura are with the Graduate School of Information Science, Nara Institute of Science and Technology, 8916–5 Takayama-cho, Ikoma, Nara, 630-0192, Japan.

Y. Shiga is with the National Institute of Information and Communications Technology, 3–5 Hikari–dai, Seika–cho, Soraku–gun, Kyoto, 619–0289, Japan.

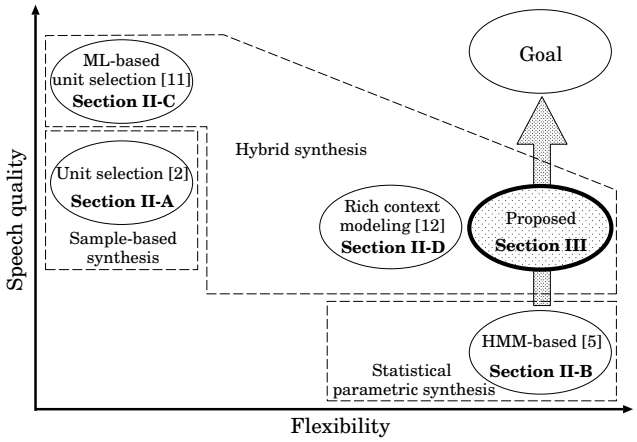


Fig. 1. Relationship diagram of the conventional methods and our proposed method.

HMM-based speech synthesis, the ability to control voice characteristics. As one hybrid approach having better flexibility than unit selection, rich context models that represent each waveform segment with probability distributions of individual speech component parameters (spectrum and F_0) have been proposed [12]. In synthesis, the probability distributions of all components corresponding to one waveform segment are selected in each HMM-state and speech parameters are generated from them in the same manner as the original HMM-based speech synthesis. This method also yields significant improvements in speech quality. However, the efficient and flexible acoustic modeling of the original HMM-based speech synthesis is lost, as this method needs to use a strong constraint among different components in the selection of their probability distributions.

In this paper, we propose ML-based parameter generation methods using rich context models as a hybrid method that preserves the flexibility of the HMM-based speech synthesis. The trained rich context models are reformulated as a Gaussian Mixture Models (GMM) in each HMM state. The speech parameter trajectory in each component is separately and iteratively generated from the corresponding GMMs using the ML criterion. The proposed methods also enable effective use of probability distributions of individual components from different waveform segments as in the original HMM-based speech synthesis. Therefore, they have better flexibility of modeling acoustic features compared with the other hybrid methods as shown in Fig. 1. In the iterative generation process, a less-smoothed but highly discontinuous parameter sequence is generated as an initial parameter sequence from probability distributions over-fitted to individual segments, and then is iteratively refined by maximizing the HMM likelihood to achieve a less-smoothed and continuous parameter sequence. We conduct several experimental evaluations of the proposed methods applied to the spectral and F_0 components. The experimental results demonstrate that the proposed methods yield significant improvements in quality of synthetic speech.

This paper is organized as follows. In **Section II**, several TTS techniques are briefly reviewed. In **Section III**, the proposed parameter generation methods with rich context models are described. In **Section IV**, the experimental evaluation results are given. **Section V** presents conclusions.

II. Text-To-Speech Techniques

A. Unit Selection Synthesis

Unit selection synthesis directly uses acoustic inventories selected from a speech corpus for synthesizing a speech waveform. In synthesis, an optimal set of acoustic segments is selected with the target information predicted by text analysis. A target cost capturing the degradation of naturalness such as caused by prosodic differences, and a concatenation cost capturing the degradation caused by concatenating acoustic segments are often used as standard selection measures. The optimal set is selected to minimize the cost function C_{us} summarizing the target cost and the concatenation cost as follows:

$$C_{\mathrm{us}} = w_{\mathrm{t}} \sum_{n=1}^{N} C_{\mathrm{t}}\left(t_n, u_n\right) + w_{\mathrm{c}} \sum_{n=2}^{N} C_{\mathrm{c}}\left(u_{n-1}, u_n\right), \tag{1}$$

where t_n and u_n are the n-th target and candidate acoustic segments, respectively, $C_{\mathrm{t}}\left(t_n, u_n\right)$ and $C_{\mathrm{c}}\left(u_{n-1}, u_n\right)$ are the target cost function evaluating the difference between t_n and u_n and the concatenation cost function evaluating the discontinuity at a joint point between u_{n-1} and u_n, respectively, w_{t} and w_{c} are the weights of the target and concatenation cost functions, respectively. The weight of each cost function is often determined manually on the basis of the result of perceptual experiments [14].

One of the main advantages of unit selection synthesis is that high-quality speech keeping the original voice characteristics is synthesized by concatenating natural acoustic segments. However, voice characteristics of the generated speech are fully dependent on the original voice. Therefore, it is difficult to flexibly control voice characteristics.

B. HMM-based Speech Synthesis

Various contextual factors need to be considered to model speech parameters in speech synthesis. Because combinations of the contextual factors increase exponentially and the number of them is enormous, one context label usually corresponds to only one acoustic segment in training data. In HMM-based speech synthesis, to robustly train context-dependent HMMs, different full context labels are tied together in a decision tree [15]. In general, the decision tree for context clustering is constructed based on the Minimum Description Length (MDL) criterion [16], which is given by

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

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, $\mathbf{\Sigma}_c$ is a covariance matrix of leaf node c, and $\Gamma\left(c\right)$ and $\Gamma\left(0\right)$ are state occupancy counts in leaf node c and that in the root node, respectively.

After the tree-based context clustering, the output probability density function b_c (clustered model) is calculated for each leaf node. Different decision trees are constructed for individual speech components (spectrum, F_0, and duration) [15].

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

$$b_c\left(\boldsymbol{o}_t\right) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right), \tag{3}$$

where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the mean vector and covariance matrix of leaf node c. The Gaussian distribution with a mean vector $\boldsymbol{\mu}_c$ and a covariance matrix $\boldsymbol{\Sigma}_c$ is denoted as $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right)$. In HMM-based speech synthesis, a feature vector is defined as $\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta\Delta\boldsymbol{c}_t^\top\right]^\top$, which includes static feature $\boldsymbol{c}_t$, and dynamic features $\Delta\boldsymbol{c}_t$, $\Delta\Delta\boldsymbol{c}_t$. These dynamic features are computed from $\boldsymbol{c}_t$ by

$$\Delta\boldsymbol{c}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} \omega_\tau^{(1)} \boldsymbol{c}_{t+\tau}, \tag{4}$$

$$\Delta\Delta\boldsymbol{c}_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} \omega_\tau^{(2)} \boldsymbol{c}_{t+\tau}, \tag{5}$$

where $\omega_\tau^{(n)}$, $L_-^{(n)}$, and $L_+^{(n)}$ are the n-th order weight coefficient and frame lengths for computing dynamic features.

F_0 **component:** F_0 parameters are modeled by a Multi-Space Distribution HMM (MSD-HMM) [17]. Its state output probability is given by

$$b_c\left(\boldsymbol{o}_t\right) = \begin{cases} w_c \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right), & l_t = \mathrm{V} \\ 1 - w_c, & l_t = \mathrm{U} \end{cases}, \tag{6}$$

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, respectively. Note that l_t is observable together with $\boldsymbol{o}_t$.

As additional speech parameters, aperiodic parameters are often used and are modeled with continuous HMMs.

In synthesis, full context labels to be synthesized are clustered with decision trees, and the output probability density functions at corresponding leaf nodes are selected to form a sentence HMM. After determining state durations $\boldsymbol{q} = \left[q_1, \cdots, q_T\right]^\top$, a time sequence of 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 ($\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$), as follows:

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{o}|\hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) \tag{7}$$

$$= \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c}|\hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right), \tag{8}$$

where $\boldsymbol{W}$ is the weighting matrix for calculating the dynamic features [18], $\boldsymbol{\lambda}$ is the parameter set of the HMM, and $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top$ is a feature vector sequence.

One of the well-known approaches for quality improvements in synthetic speech is the use of Global Variance (GV) [9]. GV is defined as the variance of the static feature vectors over an utterance, which is calculated as:

$$\boldsymbol{v}\left(\boldsymbol{c}\right) = \left[v\left(1\right), \cdots, v\left(D_c\right)\right]^\top, \tag{9}$$

$$v\left(d\right) = \frac{1}{T}\sum_{t=1}^{T}\left(c_t\left(d\right) - \overline{c}\left(d\right)\right)^2, \tag{10}$$

$$\overline{c}\left(d\right) = \frac{1}{T}\sum_{t=1}^{T} c_t\left(d\right), \tag{11}$$

where D_c is the number of dimensions of the static feature vectors. The GV likelihood $P\left(\boldsymbol{v}\left(\boldsymbol{c}\right)|\boldsymbol{\lambda}_v\right)$ is modeled by a single Gaussian distribution with the mean vector $\boldsymbol{\mu}_v$ and the covariance matrix $\boldsymbol{\Sigma}_v$ as follows:

$$P\left(\boldsymbol{v}\left(\boldsymbol{c}\right)|\boldsymbol{\lambda}_v\right) = \mathcal{N}\left(\boldsymbol{v}\left(\boldsymbol{c}\right); \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v\right) \tag{12}$$

where $\boldsymbol{\lambda}_v$ is a parameter set of the GV. The GV likelihood is estimated using natural speech parameter sequences. In parameter generation methods considering the GV, a speech parameter sequence $\boldsymbol{c}$ is generated to maximize product of the HMM likelihood and the GV likelihood, which is given by

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c}|\hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{v}\left(\boldsymbol{c}\right)|\boldsymbol{\lambda}_v\right)^{w_{\mathrm{GV}}}, \tag{13}$$

where w_{GV} is the weight of GV likelihood.

One of the biggest advantages of HMM-based speech synthesis is the capability to flexibly control voice characteristics. However, the speech parameters generated by HMM-based speech synthesis are over-smoothed because the detailed characteristics of speech parameters are often removed in the statistical process. This over-smoothing effect, which causes significant degradation in synthetic speech quality is alleviated by considering the GV. However, it is known that this method often causes artificial sounds, such as clicks, pops, and short high-pitched whines. Therefore, another approach to alleviate the over-smoothing effect without causing any artifacts is required.

C. Hybrid Synthesis with Waveform Concatenation

In order to avoid manually tuning cost functions used in unit selection synthesis, ML-based unit selection synthesis [11] has been proposed to combine unit selection synthesis and HMM-based speech synthesis. In training, after the standard HMMs are trained in the same process as in HMM-based speech synthesis, two additional statistical models called the phone duration model and the concatenation model are trained. The phone duration model represents the duration of each phone-sized acoustic segment. On the other hand, the concatenation model is defined as the differential of acoustic features between the first frame of the current phone and the last frame of the previous phone.

In synthesis, the optimal set of acoustic segments is selected from the speech database to maximize the cost function combining the likelihoods of HMM, the phone duration model, and the concatenation model. The cost function C_{ML} is represented in the same form as that of unit selection, which is given by

$$C_{\mathrm{ML}} = \sum_{n=1}^{N} C_{\mathrm{t}}\left(u_n\right) + \sum_{n=2}^{N} C_{\mathrm{c}}\left(u_{n-1}, u_n\right), \tag{14}$$

The use of acoustic segments dramatically improves speech quality. Moreover, introducing an automatically determined cost function is an effective way to avoid the manual setting of each cost function. However, the waveform generation process using waveform concatenation loses the flexibility of HMM-based speech synthesis.

D. Hybrid Synthesis with Parameter Generation

As one of the hybrid approaches having better flexibility than the standard unit selection synthesis or the ML-based unit selection synthesis, the use of rich context models to represent each acoustic segment with probability distributions of individual speech component parameters, such as spectrum and F_0 has been proposed [12]. In the traditional HMM-based speech synthesis, 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 the model parameters, in particular its covariance matrix. Although the use of GMMs as each state output probability distribution reduces the over-smoothing effect [18], its reduction effect is limited. To alleviate the over-smoothing effect while preserving robustness of the parameter estimation, in 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 [12].

Spectral component: The output probability density function $b_{c,m}$ of the rich context model for the m-th full context label in the c-th leaf node is given by

$$b_{c,m}\left(\boldsymbol{o}_t\right) = \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right). \tag{15}$$

F_0 **component:** Same as the spectral parameters, a mean vector of the Gaussian distribution in voiced space is updated as follows:

$$b_{c,m}\left(\boldsymbol{o}_t\right) = \begin{cases} w_c \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right) & l_t = \mathrm{V} \\ 1 - w_c, & l_t = \mathrm{U} \end{cases}. \tag{16}$$

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, parameters of the clustered models are estimated in the traditional manner of HMM-based speech synthesis. Then, they are untied and only their mean vectors are further updated in every full context label using the Baum-Welch algorithm while tying the covariance matrices over full context labels in each leaf node.

In synthesis, full context labels to be synthesized are clustered with the decision trees and the clustered models at corresponding leaf nodes are determined as a target $\boldsymbol{g} = \{g_1, \cdots, g_N\}$ where g_n represents the clustered model in the n-th state. Then, a sequence of the rich context models $\boldsymbol{r} = \{r_1, \cdots, r_N\}$ is selected to minimize the following Kullback-Leibler Divergence (KLD) considering both spectral and F_0 components where r_n represents the rich context model in the n-th state

$$\mathcal{D}_{\mathrm{KL}}\left(\boldsymbol{g}, \boldsymbol{r}\right) = \sum_{n=1}^{N} \mathcal{D}_{\mathrm{KL}}\left(g_n, r_n\right) T_n, \tag{17}$$

$$\mathcal{D}_{\mathrm{KL}}\left(g_n, r_n\right) = \mathcal{D}_{\mathrm{KL}}^{(\mathrm{SP})}\left(g_n, r_n\right) + \mathcal{D}_{\mathrm{KL}}^{(\mathrm{F0})}\left(g_n, r_n\right), \tag{18}$$

where $\mathcal{D}_{\mathrm{KL}}\left(\cdot\right)$ is the total KLD, $\mathcal{D}_{\mathrm{KL}}^{(\mathrm{SP})}\left(g_n, r_n\right)$ and $\mathcal{D}_{\mathrm{KL}}^{(\mathrm{F0})}\left(g_n, r_n\right)$ are KLDs for spectral and F_0 components respectively, and T_n is a state duration in the n-th state.

The rich context models for spectral and F_0 components are selected simultaneously using a constraint among different components (spectrum and F_0). This selection process can be regarded as unit selection, where each acoustic unit is represented as a joint probability distribution of both spectrum and F_0. This approach also yields significant improvements in speech quality. However, efficient and flexible acoustic modeling in the original HMM-based speech synthesis is lost by the use of the strong constraint in the model selection.

III. Parameter Generation Methods with Rich Context Models Based on the ML Criterion

A. Reformulation of GMM Using Rich Context Models

An overview of the proposed method is shown in Fig. 2. In the proposed method, the rich context models are trained for each leaf node after training conventional clustered models. In synthesis, after determining the leaf nodes corresponding to full context labels to be synthesized, the rich context models must be selected from a large number of models in the leaf nodes. The proposed methods introduce a model selection process based on the ML criterion. After training the rich context models in the same manner as in the conventional method described in **Section II-D**, the output probability density in each leaf node is modeled using a GMM developed with all rich context models in the same leaf node as follows:

Spectral component:

$$b_c\left(\boldsymbol{o}_t\right) = \sum_{m=1}^{M_c} \omega_m \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right), \tag{19}$$

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 the 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.

F_0 **component:** The F_0 component is calculated as

$$b_c\left(\boldsymbol{o}_t\right) = \begin{cases} \sum_{m=1}^{M_c} w_{c,m} \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_c\right) & l_t = \mathrm{V} \\ 1 - w_c, & l_t = \mathrm{U} \end{cases}, \tag{20}$$

where $\omega_{c,m}$ is the mixture component weight in voiced space of the m-th rich context model. We set it to an constant value ($\omega_{c,m} = \omega_c / M_c$) based on our previous findings as mentioned above.

B. Parameter Generation Methods

Given a state sequence $\hat{\boldsymbol{q}}$ [15], which is determined in the traditional HMM-based speech synthesis, the HMM likelihood is written as

$$P\left(\boldsymbol{o}|\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right) = \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{o},\boldsymbol{m}|\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right), \tag{21}$$

where $\boldsymbol{m} = \{m_1, \cdots, m_T\}$ is a mixture component sequence. 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 [18] as follows:

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\text{argmax}} \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{o},\boldsymbol{m}|\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right). \tag{22}$$

1) EM algorithm: The ML estimate of $\boldsymbol{c}$ is determined with the 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 $P\left(\boldsymbol{m}|\boldsymbol{W}\boldsymbol{c}^{(i)},\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right)$ given a current estimate $\boldsymbol{c}^{(i)}$ in the E step and a new estimate $\hat{\boldsymbol{c}}^{(i+1)}$, while fixing it constant in the M step:

$$Q\left(\boldsymbol{c}^{(i)},\boldsymbol{c}^{(i+1)}\right) = \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{W}\boldsymbol{c}^{(i)},\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right) \ln P\left(\boldsymbol{W}\boldsymbol{c}^{(i+1)},\boldsymbol{m}|\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right) \tag{23}$$

This equation is solved with the conventional generation method using the HMM-GMM [18].

2) Approximation with Single Gaussian: We approximate the HMM likelihood given in Eq. (21) with a single mixture component sequence as follows:

$$P\left(\boldsymbol{o}|\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right) = \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{o},\boldsymbol{m}|\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right) \simeq P\left(\boldsymbol{o},\hat{\boldsymbol{m}}|\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right). \tag{24}$$

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

$$\hat{\boldsymbol{m}}^{(i+1)} = \underset{\boldsymbol{m}}{\text{argmax}}\, P\left(\boldsymbol{m}|\boldsymbol{W}\hat{\boldsymbol{c}}^{(i)},\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right), \tag{25}$$

$$\hat{\boldsymbol{c}}^{(i+1)} = \underset{\boldsymbol{c}}{\text{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{c}|\hat{\boldsymbol{m}}^{(i+1)},\hat{\boldsymbol{q}},\boldsymbol{\lambda}\right). \tag{26}$$

Eq. (26) is solved in the same manner as traditional HMM-based speech synthesis.

C. Initialization Method with Tree-based Context Clustering

In the proposed parameter generation methods, we need to determine an initial parameter sequence. One of the straightforward ways is to use the parameter sequence generated by the clustered models of the traditional HMM-based speech synthesis. However, these initial parameters are not effective in improving quality. As we will show in **Section IV-D**, speech parameters generated by the proposed methods are strongly dependent on the setting of the initial parameters. Although the transitions of this initial parameter sequence are continuous, the parameter sequence is over-smoothed. Consequently, the parameters finally generated through the iterative generation process tends to still be over-smoothed.

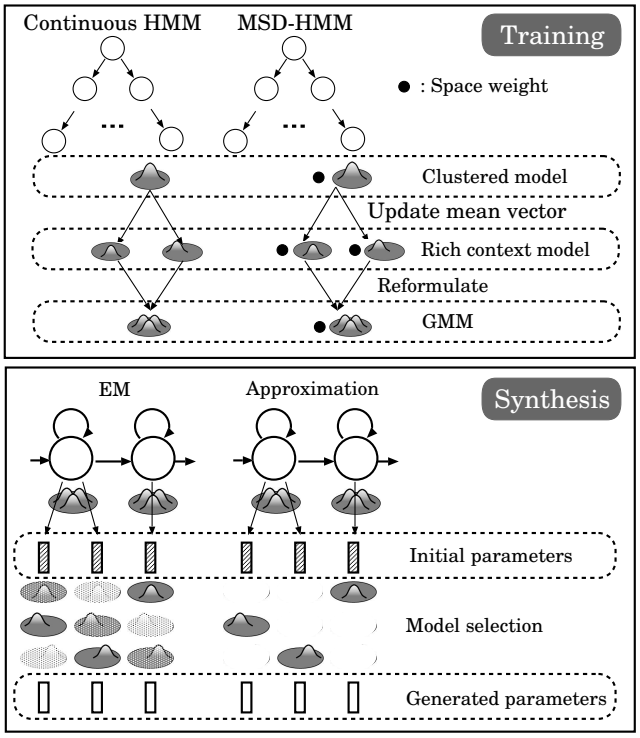


Fig. 2. An overview of the proposed generation methods.

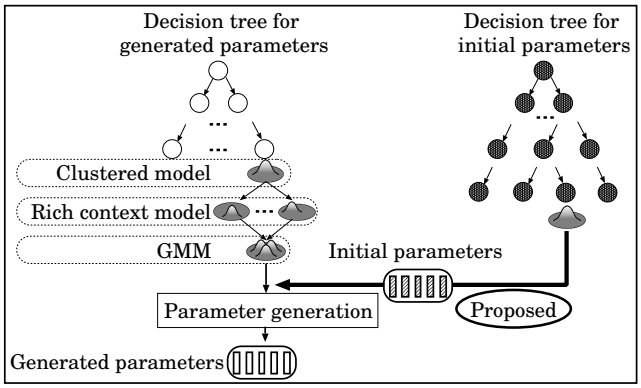


Fig. 3. An overview of the proposed initialization method.

To generate a less-smoothed initial parameter sequence, we propose an initialization method with over-fitted models generated by tree-based context clustering. As shown in Fig. 3, a large-sized tree for context clustering is constructed by decreasing the parameter a of the MDL criterion shown in Eq. (2). Note that the sufficient statistics to build this tree are the same as those used in calculating rich context models, which are extracted using the conventional clustered models. Therefore, its tree structure is slightly different from that of the decision tree for rich context models, which is the same as that for the conventional clustered models and is built using different sufficient statistics before developing the conventional clustered models.

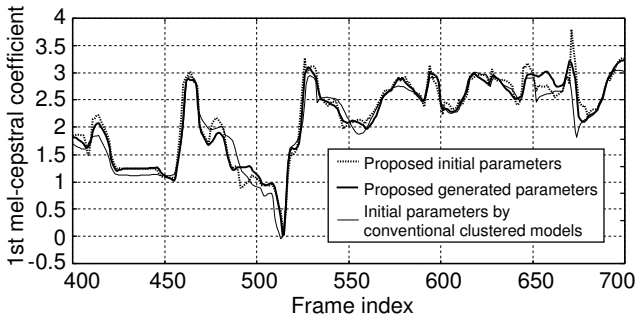


Fig. 4. An example of initial and generated mel-cepstral parameter sequence.

In this tree, the mean vector and covariance matrix of the probability density function are calculated from only a few acoustic inventories determined by context factors. Therefore, compared to a parameter sequence generated by the conventional clustered model, a less-smoothed initial parameter sequence is generated. It is expected that this initial parameter sequence helps to select a less-smoothed model sequence in the parameter generation process with rich context models. On the other hand, the use of a larger-sized decision tree causes over-fitting problems. In particular, the initial parameter sequence significantly suffers from discontinuous transitions. In the parameter generation with rich context models, this problem of discontinuity is addressed by the use of tied covariance matrices in the rich context models and the model selection based on the likelihoods for both static and dynamic features. An example of initial and generated parameter sequences by the proposed method is shown in Fig. 4. As a comparison, an initial parameter sequence generated by the conventional clustered models is also shown in this figure. We can see that discontinuous transitions in the initial parameter sequence are alleviated in the generated parameter sequence.

For the F_0 contour generation, the voiced/unvoiced interval is determined by that of the initial parameter sequence determined by the weight of the voiced space of the clustered models in the larger-sized decision tree.

D. Discussions

One rich context model usually corresponds to one HMM-state acoustic segment. Therefore, the proposed processes are strongly related to unit selection synthesis. In the proposed method, the HMM likelihood for the static features and that for the dynamic features are regarded as a target cost and a concatenation cost, respectively [19], [20].

The synthesis process with the EM algorithm is similar to the process of selecting multiple acoustic segments and mixing them to generate speech parameters [21]. 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 parameter generation methods don't have to use the constraint used in the conventional selection method of the rich context model and still keep the acoustic modeling framework the same as that of traditional HMM-based speech synthesis. Therefore, the proposed methods preserve the advantage of flexible acoustic modeling provided by traditional HMM-based speech synthesis. For instance, it is possible to separately search for the best rich context model sequences for different speech component parameters to more widely cover a joint acoustic space. Moreover, the probabilistic representation with GMMs in the proposed method makes it possible to straightforwardly use various techniques proposed for HMM-GMMs, such as flexible control of the model complexity according to individual speech components using traditional information criteria to develop a scalable system, and model adaptation directly using the conventional techniques, such as maximum likelihood linear regression [22]. It also has a potential to optimize the covariance matrices of the rich context models are also optimized in the proposed method by directly using the conventional methods, such as cross validation [23]. On the other hand, it is not straightforward to apply these techniques to the framework based on the conventional selection method of the rich context models, and therefore, some modifications would be required. It is also straightforward to apply different speech parameter generation methods to individual speech component parameters such as the conventional method with or without the GV [9].

In the proposed parameter generation 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. In the state-based model selection for the F_0 component, the voiced/unvoiced region in each HMM state is determined by the ratio of the number of voiced frames to that of unvoiced frames.

IV. Experimental Evaluation

A. Experimental Conditions

We trained a context-dependent phoneme Hidden Semi-Markov Model (HSMM) [24] for a Japanese female speaker. We used 450 sentences for training and 53 sentences for evaluation from phonetically balanced 503 sentences including in ATR Japanese speech database [25]. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 0th-through-24th mel-cepstral coefficients were extracted as spectral parameters and log-scaled F_0 and 5 band-aperiodicity [26], [27] were extracted as excitation parameters by the STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) analysis system [28]. The feature vector consisted of spectral and excitation parameters and their delta and delta-delta features. Five-state left-to-right HSMMs were used. In synthesis, global variance (GV) [9] was not considered. Both conventional clustered models and the rich context models were constructed. Then, the rich context models were reformulated as GMMs using the proposed methods. Table I shows the numbers of leaf nodes in the conventional clustered models and the rich context models. The average numbers of the mixture components were 186.3 for the spectral component and 60.7 for the F_0 component.

We conducted five kinds of experimental evaluation. In the first evaluation, we compared the two proposed parameter

generation methods described in **Section III-B**. In the second evaluation, we compared frame-based and state-based model selection to investigate the effect of the model selection unit. In the third evaluation, we investigated the effects of the initial parameter sequence on the generated parameter sequence. In the fourth evaluation, we investigated the effectiveness of the proposed initialization method described in **Section III-C**. In the last evaluation, we applied proposed methods to both spectral and F_0 components to confirm the effectiveness of proposed methods. Conventional clustered models were used for duration and aperiodic components in all evaluations.

B. Comparison of Proposed Parameter Generation Methods

1) Comparison of Proposed Methods: To evaluate two proposed generation methods, we compared synthetic speech generated by the conventional clustered model (Conv), the proposed generation method with EM algorithm (Proposed (GMM)), that with a single mixture component sequence (Proposed (Single)), and the single mixture component sequence selected by the natural speech parameters as a reference (Target). The initial parameter sequence in the proposed generation methods was generated by "Conv." Note that the proposed generation methods are applied to only the spectral component, and the clustered model is used for the F_0 component. 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. Note that natural state duration determined by the state-level forced alignment with the conventional context-clustered models was used to clarify the effectiveness of the proposed methods in a better setting.

The result is shown in Fig. 5. The proposed methods significantly improve speech quality. Moreover, the use of a single mixture component sequence yields better speech quality than the use of the EM algorithm. We can also see that the best rich context model sequence, which is approximated with "Target," is difficult to select using the likelihood measure. In the following experiments, the parameter generation method using approximation with a single mixture component sequence was used as the proposed parameter generation method.

TABLE I
NUMBERS OF LEAF NODES IN THE CONVENTIONAL CLUSTERED MODELS AND THE RICH CONTEXT MODELS.

Component	Model	State	Number of leaf nodes
Spectrum	Clustered models	1	151
		2	160
		3	171
		4	105
		5	141
	Rich Context Models	1 - 5	27118
F_0	Clustered models	1	497
		2	473
		3	580
		4	374
		5	310
	Rich Context Models	1 - 5	27118

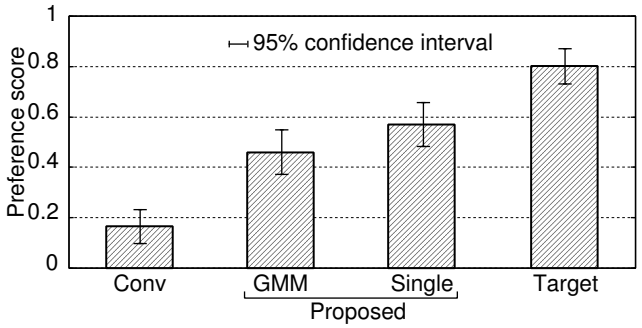


Fig. 5. Preference score on speech quality for comparing two proposed generation methods.

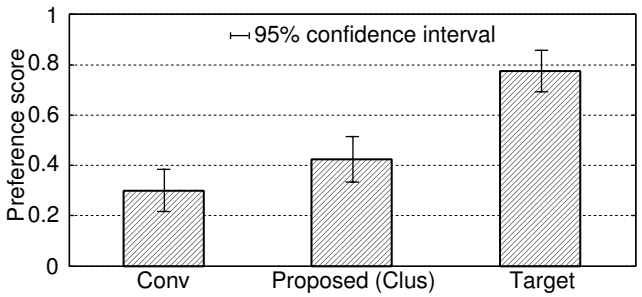


Fig. 6. Preference scores on speech quality for investigating quality improvements under generated duration.

2) Evaluation in Generated Duration: To investigate the effectiveness of the proposed generation method under the generated duration, we compared synthetic speech: 1) Conv: generated from conventional clustered models, 2) Proposed (Clus): generated using the parameter sequence of "Conv" as the initial parameters in the proposed generation method, 3) Target: generated using natural target speech parameters as the initial parameters in the proposed generation method. A preference test (AB test) by seven listeners on speech quality was conducted in the same manner as in **Section IV-B1**. Note that the proposed method is applied to only spectral parameters.

The result of the preference test is shown in Fig. 6. It is observed that the proposed generation method yields only slight improvements in synthetic speech. On the other hand, we can find that the difference between "Proposed (Clus)" and "Target" is larger than that in Fig. 5. From this result, we can find that the state duration affects the quality improvements in the proposed generation method. Synthetic speech using the generated state duration sounds more muffled compared with that using the natural state duration. It is expected that this quality degradation is caused by the quality differences of the initial parameter sequences because we have found that similar quality differences between the generated state duration and the natural state duration are also observed in synthetic speech using the conventional clustered models.

Although we did not conducted similar experiments for the F_0 component in this section, we believe that it will show the same results in the F_0 component.

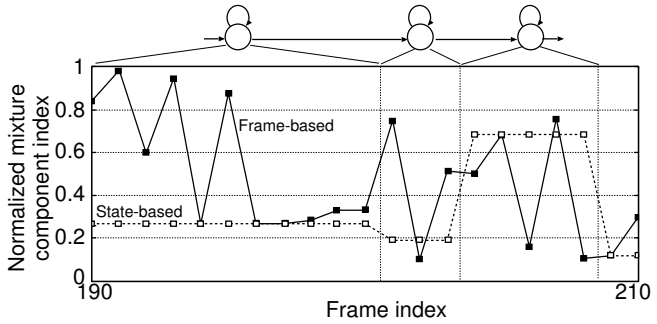
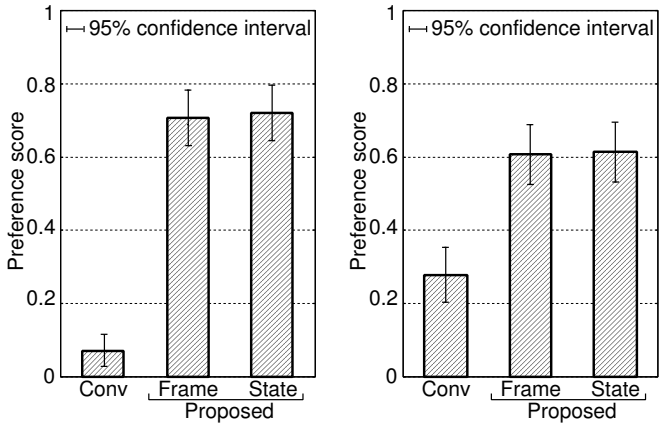


Fig. 7. An example of mixture component sequences selected by frame-based and state-based model selection (the selected mixture component index is normalized by the total number of mixture components.).



(a) Spectrum (b) F_0

Fig. 8. Preference scores on speech quality for comparing the selection unit

C. Comparison of Model Selection Unit

To investigate the effect of the model selection unit in the proposed generation method, i.e., frame-based selection or state-based selection, we compared synthetic speech generated by the conventional clustered model (Conv), the proposed method with a single mixture component sequence selected frame by frame (Proposed (Frame)), and that selected state by state (Proposed (State)). The natural state duration was used to investigate the effects in a better setting. The proposed generation method was applied to each spectral and F_0 component, and the natural speech parameter sequence was used as initial parameters. The conventional clustered models were used for the component that the proposed methods were not applied to. A preference tests by 7 listeners on speech quality were conducted for spectral and F_0 component in the same manner as in the **Section IV-B**. We confirmed that the mixture component sequences selected by these two methods were different from each other as shown in Fig. 7.

The result for spectral component is shown in Fig. 8(a), and that for the F_0 component is shown in Fig. 8(b). We can see that there is no significant difference between the frame-based selection and the state-based selection in spectral components. Moreover from Fig. 8(b), the same result is observed in the F_0 component even though the U/V intervals are also different from each other. From these results, the state-based selection is also effective for improving synthetic speech quality in both spectral and F_0 components. We can also see that the difference between "Conv" and "Proposed (Frame)" is larger in the spectral component than that in the F_0 component. This result shows that the improvements yielded by the rich context models for the spectral component is larger than those for the F_0 component.

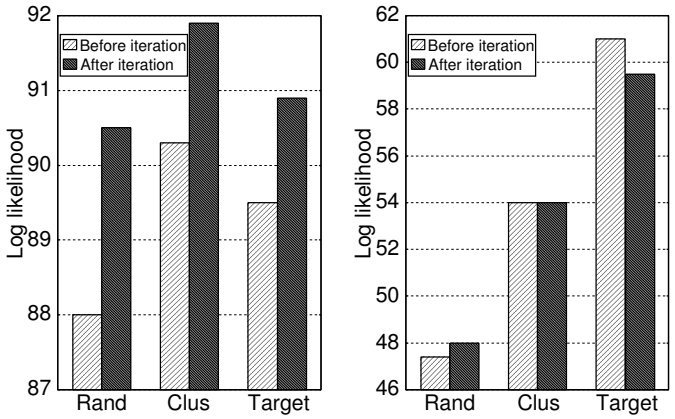
D. Investigation of Dependency on Initial Parameter Sequence

To investigate the dependency on the setting of initial parameter sequence on the finally generated speech parameter sequence after the proposed parameter generation, we evaluated three settings of the initial parameter sequence: 1) Rand: generated from rich context models randomly selected from individual leaf nodes, 2) Clus: generated from the conventional context-clustered models, and 3) Target: natural target speech parameters. The initially selected rich context model sequence and the finally selected rich context model sequence were evaluated with the model likelihood for the generated speech parameters and that for natural speech parameters. This evaluation was conducted for each spectral and F_0 components under the natural state duration.

The result of HMM likelihood for the generated parameters for the spectral components is shown in Fig. 9(a), that for the natural parameters in the spectral components is shown in Fig. 9(b), and those for the F_0 component are shown in Fig. 10(a) and Fig. 10(b), respectively. Because the HMM likelihood for the generated parameters is the criterion for the parameter generation, it is reasonable that the likelihood for the generated speech parameters increases through iteration in both components as shown in Fig. 9(a) and Fig. 10(a). On the other hand, the likelihood for the natural speech parameters does not always increase through iteration and its value strongly depends on the initial parameter sequence as shown in Fig. 9(b) and Fig. 10(b). We can also see that the likelihood differences in Fig. 9(b) and Fig. 10(b) are much larger than those in Fig. 9(a) and Fig. 10(a). These results suggest that the proposed generation method strongly depends on the initial parameter.

E. Effectiveness of Initialization Method

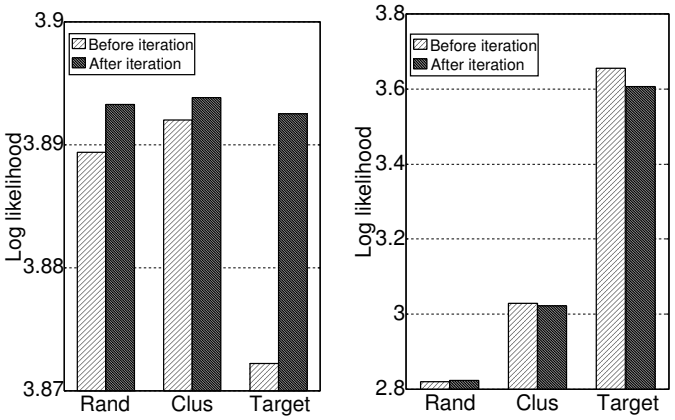
1) Confirmation of Alleviating Discontinuous Transition: Before investigating the effectiveness of the proposed initialization method, we conducted a preliminary experiment to confirm whether or not the proposed iterative parameter generation method effectively alleviates the discontinuous transition in the initial parameter sequence. We evaluated three settings of the initial parameter sequences: 1) Clus: initial parameters generated from the conventional clustered models, 2) $a = 0.1$: initial parameters generated with a large-sized decision tree ($a = 0.1$), and 3) Target: natural target speech parameter sequence as a target reference. The difference of HMM likelihoods for the generated parameters between the initially selected rich context model sequence and the finally



(a) For the generated parameters.

(b) For the natural speech parameters.

Fig. 9. HMM likelihood of selected rich context models (spectrum).



(a) For the generated parameters.

(b) For the natural speech parameters.

Fig. 10. HMM likelihood of selected rich context models (F_0).

selected rich context model sequence was calculated for each static and dynamic features in the spectral parameters.

The result of the likelihood differences caused by the iterative parameter generation is shown in Fig. 11. It was observed that the HMM likelihood for dynamic features of "$a = 0.1$" increases more than that in the other initial parameter sequences. From this result, we can see that the discontinuous transitions in the initial parameter sequence are alleviated by the iterative parameter generation.

2) Objective Evaluation of Initialization Method: To investigate the tree size used to generate the initial parameter sequence, we evaluated 3 settings of the initial parameters: 1) Clus: initial parameters generated from the conventional clustered models, 2) Proposed: initial parameters generated with a large-sized decision tree ($a = 0.1, 0.2, \cdots, 1.0$), and 3) Target: the natural target speech parameter sequence as a

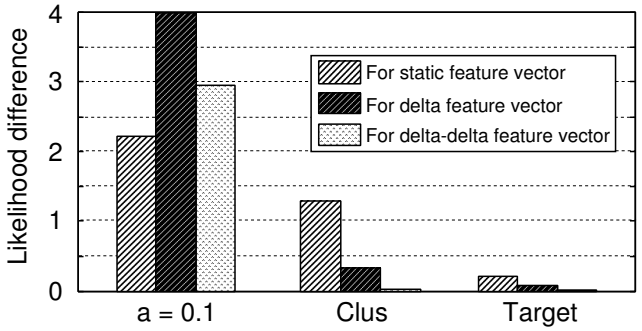
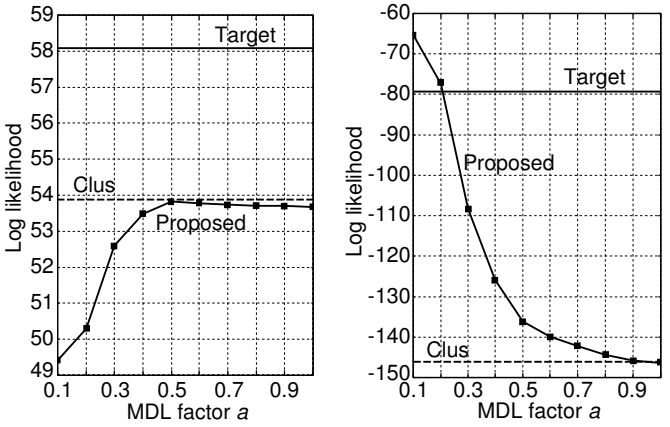


Fig. 11. Differences of HMM likelihood between before and after iteration.

target reference for each spectral and F_0 components. The rich context model sequences finally selected by the parameter generation method using these initial parameter settings were evaluated with the HMM likelihood for the natural speech parameters. The tree size was calculated as the ratio of the number of leaf nodes of the decision trees compared to the number of full context models. Moreover, the parameter sequences generated by the selected rich context models (i.e., those generated by the proposed parameter generation method) are evaluated with both the GV likelihood [9] and U/V error rate. The U/V error rate for the F_0 component is calculated as the ratio of the number of U/V mismatched frames in the generated parameter sequence compared to the natural parameter sequence.

The result of HMM likelihood in the spectral component is shown in Fig. 12(a), that of GV likelihood in the spectral component is shown in Fig. 12(b), and those in the F_0 component are shown in Fig. 13(a) and Fig. 13(b), respectively. Moreover, the size of decision trees used in the proposed initialization method is shown in Fig. 14, and the result of U/V error rate is shown in Fig. 15. It is observed from Fig. 12(a) that the HMM likelihood of "Proposed" very slightly increases as the parameter a decreases from 1.0 to 0.5, and it rapidly decreases as the parameter a decreases more in the spectral components. 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." The result for the F_0 component shown in Fig. 13(a) is similar to this result except that no peaks appeared as the parameter a decreases. On the other hand, It is observed from Fig. 12(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" in the spectral component. In the F_0 component, the GV likelihood of "Proposed" rapidly increases as the parameter a decreases from 1.0 to 0.6, and it rapidly decreases beyond 0.6. Moreover from Fig. 15, we can see that the U/V error rate increases as the parameter a decreases.

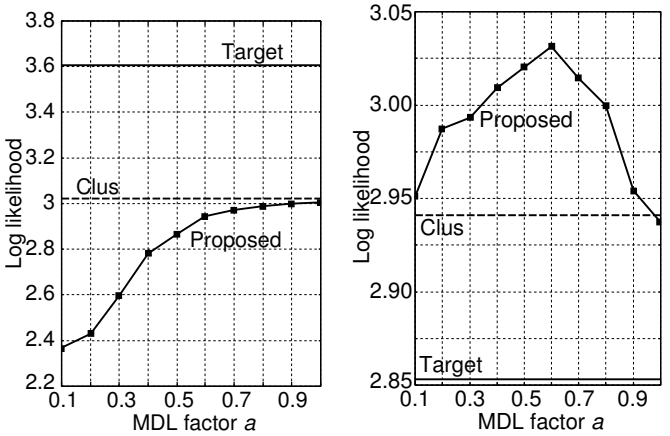
3) Subjective Evaluation of Initialization Method: To confirm the effectiveness of the proposed initialization method, two preference tests (AB test) by 7 listeners on speech quality were conducted in the same manner as in the **Section IV-B**. The evaluated synthetic speech samples are generated from the rich context models with using 1) Clus, 2) Proposed ($a = 0.1$),



(a) HMM likelihood for natural parameters

(b) GV likelihood for generated parameters

Fig. 12. Objective evaluations for initialization method (spectrum).



(a) HMM likelihood for natural parameters

(b) GV likelihood for generated parameters

Fig. 13. Objective evaluations for initialization method (F_0).

3) Proposed ($a = 0.5$), and 4) Target as the initial parameter for the spectral parameter. As described above, $a = 0.1$ is the setting of the highest GV likelihood and $a = 0.5$ is that of the highest HMM likelihood. For the F_0 component, they were 1) Conv: generated from conventional clustered model, and generated from the rich context models with using 2) Clus, 3) Proposed ($a = 0.6$), and 4) Target as the initial parameters. Note that the conventional clustered models were used for the component that the proposed methods were not applied to.

The result of the preference test in the spectral component is shown in Fig. 16, and that in the F_0 component is shown in Fig. 17. From Fig. 16, 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 that of "Proposed ($a = 0.5$)." This tendency is the same as that observed in the GV likelihood shown in Fig. 12(b). From Fig. 15 and Fig. 17, although the setting of the parameter a to maximize the GV likelihood slightly increases U/V error rate, it was still observed to be effective to improve speech quality even in the F_0 component.

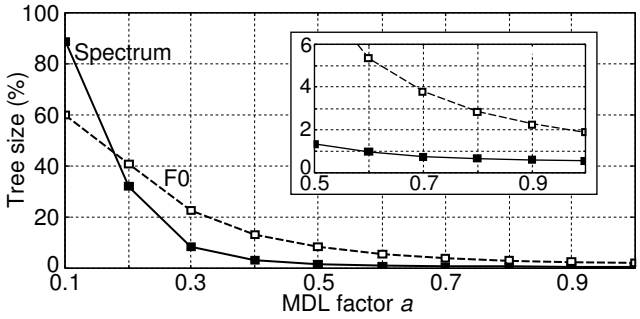


Fig. 14. Size of the decision trees for initial parameter generation.

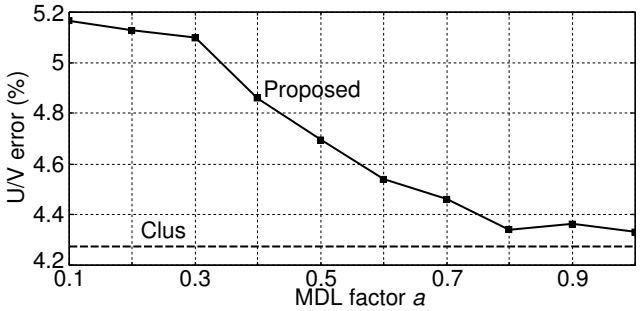


Fig. 15. Error rates of the unvoiced/voiced decision.

F. Evaluation in Full Synthesis

To investigate the effectiveness of all proposed methods, we evaluated 5 kinds of synthetic speech[1] shown in Table II. A preference test (AB test) on speech quality was conducted by 8 listeners in the same manner as in the **Section IV-B**. Note that "Target" was generated by parameter generation with rich context models using the natural speech parameter sequence as initial parameters.

The result of the preference test in full synthesis is shown in Fig. 19, and spectrograms of "Conventional," "Proposed," and the natural speech are shown in Fig. 18. It is observed that a larger speech-quality improvement was yielded by applying the proposed method to the spectral component than to the

TABLE II
SYNTHETIC SPEECH SAMPLES USED FOR "FULL SYNTHESIS" EVALUATION.

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

[1]Some samples are available from http://isw3.naist.jp/~shinnosuke-t/sample_rcm.html

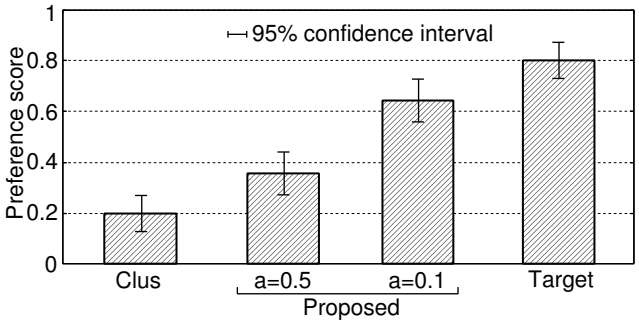


Fig. 16. Preference scores on speech quality for investigating the effectiveness of initialization method for spectral component.

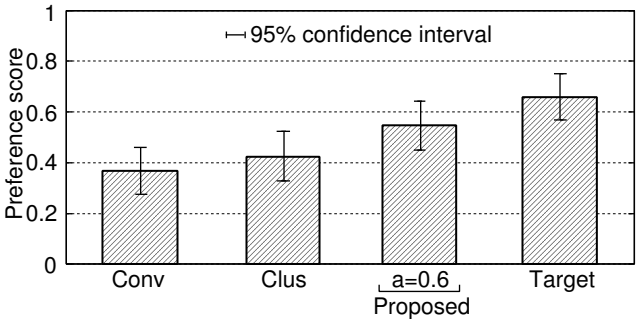


Fig. 17. Preference scores on speech quality for investigating the effectiveness of initialization method for F_0 component.

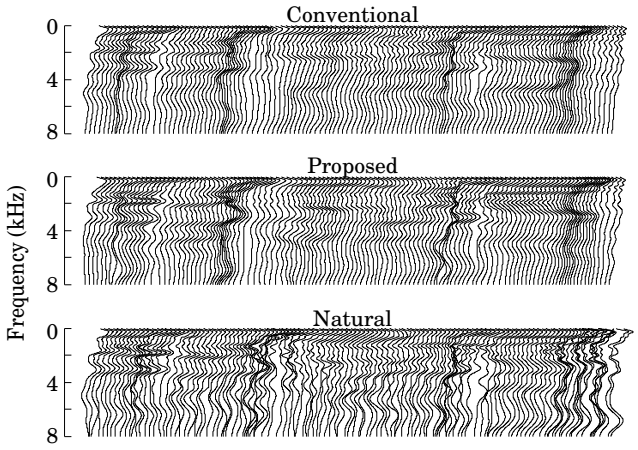


Fig. 18. An example of spectrogram of synthetic speeches ("Natural" represents the spectrograms of natural speech).

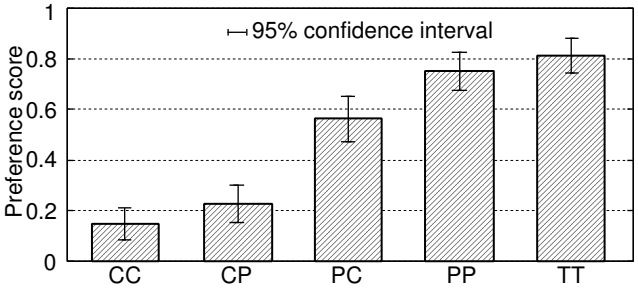


Fig. 19. Preference scores on speech quality for full synthesis.

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 in improving quality in synthetic speech.

V. Conclusion

In this paper, we proposed 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 training, the rich context models were reformulated as Gaussian Mixture Models (GMMs). In synthesis, an initial speech parameters were generated from probability distributions over-fitted to individual segments, and the speech parameter sequence was iteratively generated from GMMs using a parameter generation method based on the maximum likelihood criterion. The experimental results have demonstrated: (1) the use of approximation with a single Gaussian component sequence yields better synthetic speech quality than the use of EM algorithm in the proposed parameter generation method, (2) the state-based model selection yields quality improvements as same as the frame-based model selection. (3) the proposed initialization method is very effective to further improvement speech quality, and (4) the proposed methods for spectral and F_0 components yields significant improvements in synthetic speech quality compared with the traditional HMM-based speech synthesis.

Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Number 24240032, and 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. The authors are grateful to Prof. Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method.

References

[1] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, U.S.A, Apr. 1988, pp. 679–682.
[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. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, Atlanta, U.S.A., May 1996, pp. 373–376.
[4] A. K. Syrdal, C. W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K.-S. Lee, and M. Makashay, "Corpus-based techniques in the AT&T NextGen synthesis system," in *Proc. ICSLP*, Beijing, China, Oct 2000, pp. 410–415.
[5] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.
[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] 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.
[10] S. King and V. Karaiskos, "The blizzard challenge 2011," in *Proc. Blizzard Challenge workshop*, Turin, Italy, Sept. 2011.
[11] 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," in *Proc. Blizzard Challenge workshop*, Bonn, Germany, Aug. 2007.
[12] Z. Yan, Q. Yao, and S. K. Frank, "Rich context modeling for high quality HMM-based TTS," in *Proc. INTERSPEECH*, Brighton, U.K., Sept. 2009, pp. 1755–1758.
[13] Y. Qian, Z. Yan, Y. Wu, and F. K. Soong, "An HMM trajectory tiling (HTT) approach to high quality TTS," in *Proc. INTERSPEECH*, Chiba, Japan, Sept. 2010, pp. 422–425.
[14] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis," in *Proc. ICASSP*, Montreal, Canada, May 2004, pp. 657–660.
[15] 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*, Budapest, Hungary, Apr. 1999, pp. 2347–2350.
[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. 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.
[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1315–1318.
[19] S. Kataoka, N. Mizutani, K. Tokuda, and T. Kitamura, "Decision tree backing-off in HMM-based speech synthesis," in *Proc. INTERSPEECH*, vol. 2, Jeju, Korea, Oct. 2004, pp. 1205–1208.
[20] Z. Ling and R. Wang, "HMM-based unit selection using frame sized speech segments," in *Proc. INTERSPEECH*, Pittsburgh U.S.A., Sept. 2013.
[21] T. Mizutani and 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.
[22] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," in *Computer Speech and Language*, vol. 9, 1995, pp. 171–185.
[23] T. Shinozaki and M. Ostendorf, "Cross-validation and aggregated EM training for robust parameter estimation," in *Computer Speech and Language*, vol. 22, 2008, pp. 185–195.
[24] H. Zen, K. Tokuda, T. K. T. Masuko, and T. Kitamura, "Hidden semi-markov model based speech synthesis system," *IEICE Trans., Inf. and Syst., E90-D*, no. 5, pp. 825–834, 2007.
[25] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuawhara, "A large-scale Japanese speech database," in *ICSLP90*, Kobe, Japan, Nov. 1990, pp. 1089–1092.
[26] H. Kawahara, J. 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," in *MAVEBA 2001*, Firentze, Italy, Sept. 2001, pp. 1–6.
[27] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006, pp. 2266–2269.
[28] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

Shinnosuke Takamichi received his B.E. from Nagaoka University of Technology, Japan, in 2011 and his M.E. degree from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan, in 2013. He is currently a Ph. D. student of NAIST. His research interests include electroacoustics, signal processing, and speech synthesis. He is a student member of ASJ, and is a member of ISCA.

Tomoki Toda earned his B.E. degree from Nagoya University, Aichi, Japan, in 1999 and his M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively. He was a Research Fellow of JSPS in the Graduate School of Engineering, Nitech, Aichi, Japan, from 2003 to 2005. He was an Assistant Professor of the Graduate School of Information Science, NAIST from 2005 to 2011, where he is currently an Associate Professor. He has also been a Visting Researcher at the NICT, Kyoto, Japan, since May 2006. From March 2001 to March 2003, he was an Intern Researcher at the ATR SLC Research Laboratories, Kyoto, Japan, and then he was a Visiting Researcher until March 2006. He was also a Visiting Researcher at the LTI, CMU, Pittsburgh, USA, from October 2003 to September 2004 and at the CUED, Cambridge, UK, from March to August 2008. His research interests include statistical approaches to speech processing. He received the 18th TELECOM System Technology Award for Students and the 23rd TELECOM System Technology Award from the TAF, the 2007 ISS Best Paper Award and the 2010 ISS Young Researcher's Award in Speech Field from the IEICE, the 10th Ericsson Young Scientist Award from Nippon Ericsson K.K., the 4th Itakura Prize Innovative Young Researcher Award and the 26th Awaya Prize Young Researcher Award from the ASJ, the 2009 Young Author Best Paper Award from the IEEE SPS, the Best Paper Award (Short Paper in Regular Session Category) from APSIPA ASC 2012, the 2012 Kiyasu Special Industrial Achievement Award from the IPSJ, and the 2013 Best Paper Award (Speech Communication Journal) from EURASIP-ISCA. He was a member of the Speech and Language Technical Committee of the IEEE SPS from 2007 to 2009.

Yoshinori Shiga has been involved in speech technology research, since 1987, at various institutions including the University of Edinburgh, Edinburgh, U.K., the University of Surrey, Guildford, U.K., the Tokyo University of Science, Tokyo, Japan, Toshiba Corporation, Kawasaki, Japan, Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan, and the National Institute of Information and Communications Technology (NICT), Kyoto, Japan. Currently he is the leader of the Spoken Dialogue & Speech Synthesis Team at NICT's Spoken Language Communication Group. He holds B.E. and M.E. degrees in Electrical Engineering from the Tokyo University of Science, and a Ph.D. degree in Speech Technology from the University of Edinburgh. His research interests are in the area of speech and sound processing including speech synthesis and recognition. From 2009 to 2011, he was a Guest Professor of the Graduate School of Intercultural Studies at Kobe University, Kobe, Japan. He received the 2002 Information and Systems Society Paper Award from the Institute of Electronics, Information and Communication Engineers. He is currently a member of International Speech Communication Association (ISCA), the Acoustical Society of Japan (ASJ) and the Phonetic Society of Japan.

Sakriani Sakti received her B.E degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received "DAAD-Siemens Program Asia 21st Century" Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, Daimler-Chrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). She also served as a visiting professor of Computer Science Department, University of Indonesia (UI) in 2009-2011. Currently, she is an assistant professor of the Augmented Human Communication Lab, NAIST, Japan. She is a member of JNS, SFN, ASJ, ISCA, IEICE amd IEEE. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Graham Neubig received his B.E. from University of Illinois, Urbana-Champaign in 2005, and his M.S. and Ph.D. in informatics from Kyoto University in 2010 and 2012 respectively. From 2012, he has been an assistant professor at the Nara Institute of Science and Technology, where he is pursuing research in machine translation and spoken language processing.

Satoshi Nakamura is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He also serves as a visiting professor of Collaborative Research Unit, National Institute of Informatics. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He was a project leader of the world first network-based commercial speech-to-speech translation service for 3-G mobile phones in 2007 and VoiceTra project for iPhone in 2010. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampoli Award 2012. He organized the International Workshop of Spoken Language Translation (IWSLT 2006) and Oriental Cocosda 2008 as a general chair. He also served as the program chair of INTERSPEECH 2010. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011 and IEEE Signal Processing Magazine Editorial Board Member since April 2012.