IEICE TRANS. INF. & SYST., VOL.E90–D, NO.5 MAY 2007

PAPER

A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis

Tomoki TODA[†a)] *and* Keiichi TOKUDA[††b)], *Members*

SUMMARY This paper describes a novel parameter generation algorithm for an HMM-based speech synthesis technique. The conventional algorithm generates a parameter trajectory of static features that maximizes the likelihood of a given HMM for the parameter sequence consisting of the static and dynamic features under an explicit constraint between those two features. The generated trajectory is often excessively smoothed due to the statistical processing. Using the over-smoothed speech parameters usually causes muffled sounds. In order to alleviate the over-smoothing effect, we propose a generation algorithm considering not only the HMM likelihood maximized in the conventional algorithm but also a likelihood for a global variance (GV) of the generated trajectory. The latter likelihood works as a penalty for the over-smoothing, i.e., a reduction of the GV of the generated trajectory. The result of a perceptual evaluation demonstrates that the proposed algorithm causes considerably large improvements in the naturalness of synthetic speech.

key words: *HMM-based speech synthesis, speech parameter generation, maximum likelihood criterion, over-smoothing effect, global variance*

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], [2] has caused the dramatic improvements of TTS. That approach has enabled us to construct a TTS system without professional expertise, which is indispensable for constructing the system with consistent and reasonable quality in the rule-based approach [3]. So far, many generic synthesis methods have been established.

There are two main techniques of corpus-based speech synthesis, i.e., sample-based synthesis and statistical synthesis. The sample-based synthesis such as unit selection [4], [5] directly uses acoustic inventories selected from a speech corpus for synthesizing a speech waveform.*One of main advantages of the unit selection is that high-quality speech keeping original voice characteristics is synthesized by concatenating natural acoustic units. However, since the desired units with target attributes are not always in the corpus, other units with attributes similar to the target ones need to be used instead. The concatenation of such units often causes audible discontinuities. Signal processing is useful for alleviating those discontinuities. However, even if state-of-the-art speech analysis-synthesis methods [9], [10] are employed, they might cause other artificial sounds due to an inherent problem of speech analysis: the difficulty of estimating an accurate vocal tract response from sparse frequency components observed at only F_0 harmonic points. Consequently, a speech corpus that widely covers attributes of texts to be synthesized is necessary for consistently achieving high-quality synthetic speech while avoiding the analysis-synthesis processing [11], [12]. The size of such a corpus is quite huge because there are enormous contextual factors especially affecting prosodic characteristics. It is indeed laborious and expensive to prepare such a huge-sized speech corpus. Moreover, the voice quality variation is not negligible in speech recording for a long time [13]. Although the sample-based method is adopted in most of TTS systems respecting the naturalness of synthetic speech [14]–[16], it is essentially hard to flexibly synthesize various voices with rich speech characteristics.

On the other hand, the statistical synthesis methods such as Context Oriented Clustering (COC) [17] use averaged acoustic inventories statistically extracted from the speech corpus. Synthetic speech based on those inventories has smooth and consistent quality. Although the speech analysis-synthesis process is essential in this framework, the artificial sounds caused by it are alleviated by sharing observed frequency components at multiple acoustic segments [19].** Moreover, the averaging process improves the robustness against data sparseness because unseen acoustics are generated with an interpolation of the inventories having similar attributes to the target [20]. However, synthetic speech sounds evidently muffled compared with natural speech because detailed characteristics of speech signals are removed in the statistical processing. In general, this method is inferior to the sample-based method in view of the voice quality of synthetic speech.

As one of the statistical synthesis methods, we focus on the HMM-based speech synthesis method [21], [22]. This method has many advantages as follows: 1) it is well known that the HMM is suitable for modeling a time sequence of speech parameters, 2) we can apply many techniques for

Manuscript received July 11, 2006.
Manuscript revised December 11, 2006.

[†]The author is with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
[††]The author is with the Graduate School of Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.
a) E-mail: tomoki@is.naist.jp
b) E-mail: tokuda@ics.nitech.ac.jp

DOI: 10.1093/ietisy/e90–d.5.816

*There are several systems employing unit selection for predicting a target F_0 contour as well as generating waveform segments or spectral segments [6]–[8].

**An averaging process for F_0 contours is also effective for synthesizing consistently smooth speech [18].

Copyright © 2007 The Institute of Electronics, Information and Communication Engineers

HMM-based speech recognition to speech synthesis, 3) because the HMM is mathematically tractable, voice characteristics of synthetic speech are easily controlled by modifying acoustic statistics in the manner supported mathematically. It is essential in speech synthesis to realize a smooth speech parameter trajectory based on the HMM in which the trajectory is represented with discrete state sequences. The HMM-based synthesis method directly generates the trajectory from a given HMM so that the likelihood of the HMM for a parameter sequence consisting of static and dynamic features is maximized under the constraint on the explicit relationship between those two features [23]. Although that method allows the smooth trajectory with appropriate characteristics in the maximum likelihood sense, there still remains the over-smoothing problem as mentioned above. Using multiple mixtures for modeling state output probability density alleviates the over-smoothing effect [24], [25] but it also causes another problem of over-training due to an increase of the number of model parameters, which often causes discontinuities on the generated trajectory when synthesizing texts with unseen attributes. For realizing a TTS system satisfying both flexibility and naturalness, it is important to improve the basic quality of the HMM-based speech synthesis while keeping its advantages.

This paper focuses on a global variance (GV), the amount of a total variance of parameters over a time sequence, as one of characteristics which are difficult to be reconstructed in the conventional framework of the HMM-based speech synthesis. The conventional parameter generation allows the trajectory close to a mean vector sequence of the HMM. Although this process reasonably reduces the generation error, it often makes GVs of the generated trajectories much smaller than those of natural ones. In order to generate the trajectories with appropriate GVs, we propose a parameter generation algorithm considering the GV. This paper applies the proposed algorithm to both spectral and F_0 parameter generation processes in the HMM-based speech synthesis. Experimental results show that the proposed algorithm causes significant improvements of the naturalness of synthetic speech.

Section 2 describes the conventional parameter generation algorithm. Section 3 describes the proposed algorithm considering the GV. Section 4 describes experimental evaluations. Finally, we summarize this paper in Sect. 5.

2. Conventional Parameter Generation Algorithm

2.1 HMM Likelihood

Let assume a D-dimensional static feature vector $\boldsymbol{c}_t = [c_t(1), c_t(2), \cdots, c_t(d), \cdots, c_t(D)]^\top$ at frame t. We use a speech parameter vector $\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta^{(1)}\boldsymbol{c}_t^\top, \Delta^{(2)}\boldsymbol{c}_t^\top]^\top$ consisting of not only the static feature vector but also dynamic feature vectors $\Delta^{(1)}\boldsymbol{c}_t$, $\Delta^{(2)}\boldsymbol{c}_t$, which are calculated by

$$\Delta^{(n)}\boldsymbol{c}_t = \sum_{\tau=-L_-^{(n)}}^{L_+^{(n)}} w^{(n)}(\tau)\boldsymbol{c}_{t+\tau}, \quad n = 1, 2. \tag{1}$$

Parameter vectors at all frames over an utterance is regarded as a time sequence vector. The sequence vectors of $\boldsymbol{o}_t$ and $\boldsymbol{c}_t$ are written as

$$\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \cdots, \boldsymbol{o}_t^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top, \tag{2}$$

$$\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top, \tag{3}$$

respectively. The relationship between those two vectors is represented as

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}, \tag{4}$$

where $\boldsymbol{W}$ is the $3DT$-by-DT matrix written as

$$\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_t, \cdots, \boldsymbol{W}_T]^\top \otimes \boldsymbol{I}_{D\times D}, \tag{5}$$

$$\boldsymbol{W}_t = \left[\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}\right], \tag{6}$$

$$\boldsymbol{w}_t^{(n)} = \Big[\overbrace{0, \cdots, 0}^{1\text{st}}, \overbrace{w^{(n)}(-L_-^{(n)})}^{(t-L_-^{(n)})\text{-th}}, \cdots, \overbrace{w^{(n)}(0)}^{(t)\text{-th}}, \cdots, \overbrace{w^{(n)}(L_+^{(n)}), 0, \cdots, 0}^{T\text{-th}}\Big]^\top, \quad n = 0, 1, 2, \tag{7}$$

$L_-^{(0)} = L_+^{(0)} = 0$, and $w^{(0)}(0) = 1$.

A likelihood of a given continuous mixture HMM $\boldsymbol{\lambda}$ for the parameter sequence vector $\boldsymbol{o}$ is written as

$$P(\boldsymbol{o}|\boldsymbol{\lambda}) = \sum_{\text{all } \boldsymbol{Q}} P(\boldsymbol{o}, \boldsymbol{Q}|\boldsymbol{\lambda}), \tag{8}$$

where

$$\boldsymbol{Q} = \{(q_1, i_1), (q_2, i_2), \cdots, (q_t, i_t), \cdots, (q_T, i_T)\} \tag{9}$$

is the state and mixture sequence, i.e., (q, i) indicates the i-th mixture of state q.

2.2 Parameter Sequence Generation Based on Maximum Likelihood Criterion

We determine the parameter sequence of static features $\boldsymbol{c}$ that maximizes the HMM likelihood. In order to reduce computation cost, the current HMM-based speech synthesis system [21], [22] determines the sub-optimum state sequence $\hat{\boldsymbol{q}} = \{\hat{q}_1, \hat{q}_2, \cdots, \hat{q}_t, \cdots, \hat{q}_T\}$ independently of $\boldsymbol{o}$ as follows

$$\hat{\boldsymbol{q}} = \operatorname{argmax} P(\boldsymbol{q}|\boldsymbol{\lambda}), \tag{10}$$

where $P(\boldsymbol{q}|\boldsymbol{\lambda})$ is a likelihood of the state duration model, which is incorporated into HMMs for modeling temporal structure of a speech parameter sequence appropriately. When using a single Gaussian for modeling each state output probability density, the sub-optimum state and mixture sequence $\hat{\boldsymbol{Q}}$ is also determined. Under such conditions, we maximize the following log-scaled likelihood with respect to $\boldsymbol{c}$,

$$\log P(\boldsymbol{o}|\hat{\boldsymbol{Q}}, \boldsymbol{\lambda}) = -\frac{1}{2}\boldsymbol{o}^\top \hat{\boldsymbol{U}}^{-1}\boldsymbol{o} + \boldsymbol{o}^\top \hat{\boldsymbol{U}}^{-1}\hat{\boldsymbol{\mu}} + K, \tag{11}$$

where

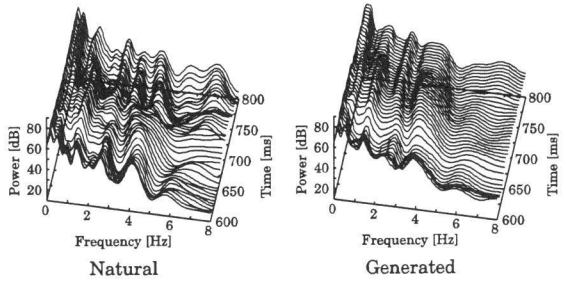


Fig. 1 An example of natural and generated spectral segments for a phoneme sequence "a-/n a u/+n."

$$\hat{\boldsymbol{\mu}} = \left[\boldsymbol{\mu}^{\top}_{\hat{q}_1,\hat{i}_1}, \boldsymbol{\mu}^{\top}_{\hat{q}_2,\hat{i}_2}, \cdots, \boldsymbol{\mu}^{\top}_{\hat{q}_t,\hat{i}_t}, \cdots, \boldsymbol{\mu}^{\top}_{\hat{q}_T,\hat{i}_T}\right]^{\top}, \tag{12}$$

$$\hat{\boldsymbol{U}}^{-1} = \mathrm{diag}\left[\boldsymbol{U}^{-1}_{\hat{q}_1,\hat{i}_1}, \boldsymbol{U}^{-1}_{\hat{q}_2,\hat{i}_2}, \cdots, \boldsymbol{U}^{-1}_{\hat{q}_t,\hat{i}_t}, \cdots, \boldsymbol{U}^{-1}_{\hat{q}_T,\hat{i}_T}\right], \tag{13}$$

$\boldsymbol{\mu}_{q_t,i_t}$ and $\boldsymbol{U}_{q_t,i_t}$ are the $3D$-by-1 mean vector and the $3D$-by-$3D$ covariance matrix, respectively, associated with i_t-th mixture of state q_t. The constant K is independent of $\boldsymbol{o}$. Under the condition (4), we determine $\boldsymbol{c}$ that maximizes the likelihood as follows

$$\boldsymbol{c} = \left(\boldsymbol{W}^{\top}\hat{\boldsymbol{U}}^{-1}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^{\top}\hat{\boldsymbol{U}}^{-1}\hat{\boldsymbol{\mu}}. \tag{14}$$

As shown in the above equation, this algorithm is not a frame-based process but a trajectory-based one, i.e., simultaneously generating static vectors at all frames. Instead of determining the sub-optimum state and mixture sequence, we can also determine $\boldsymbol{c}$ by directly maximizing $P(\boldsymbol{o}|\boldsymbol{\lambda})$ with the EM algorithm [25].

2.3 Over-Smoothing of Generated Parameters

Figure 1 shows an example of spectral segments of natural speech and synthetic speech by the HMM-based speech synthesis. The generated spectra are excessively smoothed compared with the natural ones. The statistical modeling removes the details of spectral structures. This smoothing surely causes error reduction of the spectral generation. However, it also causes the degradation of naturalness of synthetic speech because those removed structures are still necessary for synthesizing high-quality speech.

3. Proposed Parameter Generation Algorithm Considering Global Variance

We start to find which characteristics of parameter trajectories are removed statistically. As one of them, this paper focuses on their variance features.

3.1 Global Variance (GV)

A GV of the static features over a time sequence is calculated by

$$\boldsymbol{v}(\boldsymbol{c}) = [v(1), v(2), \cdots, v(d), \cdots, v(D)]^{\top}, \tag{15}$$

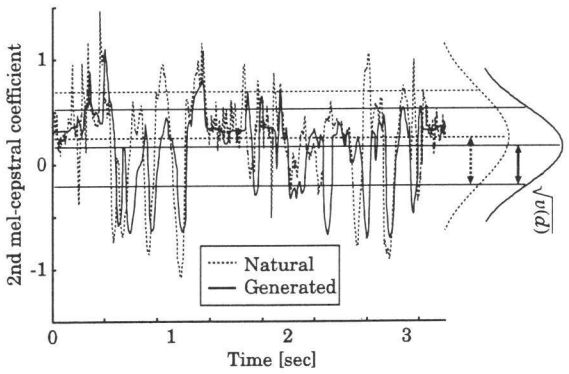


Fig. 2 Natural and generated mel-cepstrum sequences. A square root of the GV of each sequence is shown as a bidirectional arrow.

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}(c_t(d) - \overline{c}(d))^2, \tag{16}$$

$$\overline{c}(d) = \frac{1}{T}\sum_{\tau=1}^{T} c_\tau(d). \tag{17}$$

This paper calculates the GV utterance by utterance.

Figure 2 shows a time sequence of the 2nd mel-cepstral coefficients extracted from natural speech and that generated from the HMM. It can be observed that the GV of the generated mel-cepstra is smaller than that of the natural ones. The maximum likelihood criterion makes the generated trajectory close to a mean vector sequence of the HMM. The GV reduction is often observed when using the HMM of which each state output probability distribution is trained with multiple inventories from different contexts.

3.2 Proposed Likelihood

The proposed method considers not only the HMM likelihood for the static and dynamic feature vectors but also the likelihood on the GV. Specifically, instead of maximizing the likelihood (8), we maximize the following likelihood represented as a product of the two likelihoods,

$$P(\boldsymbol{o}|\boldsymbol{\lambda}, \boldsymbol{\lambda}_v) = \sum_{\text{all } \boldsymbol{Q}} P(\boldsymbol{o}, \boldsymbol{Q}|\boldsymbol{\lambda})^{\omega} P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{\lambda}_v), \tag{18}$$

where $P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{\lambda}_v)$ is modeled by a single Gaussian with the mean vector $\boldsymbol{\mu}_v$ and the covariance matrix $\boldsymbol{U}_v$ as follows:

$$P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{\lambda}_v) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{U}_v|}} \exp\left(-\frac{1}{2}(\boldsymbol{v}(\boldsymbol{c}) - \boldsymbol{\mu}_v)^{\top}\boldsymbol{U}_v^{-1}(\boldsymbol{v}(\boldsymbol{c}) - \boldsymbol{\mu}_v)\right). \tag{19}$$

The Gaussian distribution $\boldsymbol{\lambda}_v$ and the HMM $\boldsymbol{\lambda}$ are independently trained from the speech corpus. The constant ω denotes the weight for controlling a balance between the two likelihoods. This paper sets ω to the ratio of the number of dimensions between vectors $\boldsymbol{v}(\boldsymbol{c})$ and $\boldsymbol{o}$, i.e., $1/(3T)$. Note that the proposed likelihood is a function of $\boldsymbol{c}$.

3.3 Parameter Sequence Generation Considering GV Based on Maximum Likelihood Criterion

The proposed generation algorithm determines $\boldsymbol{c}$ that maximizes the likelihood (18). Namely, the parameter generation is performed under both the conventional constraint and a novel constraint on the GV of the generated trajectory. The likelihood $p(\boldsymbol{v}(\boldsymbol{c}))$ might be viewed as a penalty term for a reduction of the GV. This paper describes the generation process when $\hat{\boldsymbol{Q}}$ is determined in the same manner as described in the previous section. We can also iteratively determine $\boldsymbol{c}$ that maximizes the likelihood (18) with the EM algorithm.

The following log-scaled likelihood is maximized with respect to $\boldsymbol{c}$ under the condition of determined $\hat{\boldsymbol{Q}}$,

$$\begin{aligned} L &= \log\left[P(\boldsymbol{o}|\hat{\boldsymbol{Q}},\boldsymbol{\lambda})^{\omega}P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{\lambda}_v)\right] \\ &= \omega\left(-\frac{1}{2}\boldsymbol{c}^{\top}\boldsymbol{W}^{\top}\hat{\boldsymbol{U}}^{-1}\boldsymbol{W}\boldsymbol{c} + \boldsymbol{c}^{\top}\boldsymbol{W}^{\top}\hat{\boldsymbol{U}}^{-1}\hat{\boldsymbol{\mu}}\right) \\ &\quad -\frac{1}{2}\boldsymbol{v}(\boldsymbol{c})^{\top}\boldsymbol{U}_v^{-1}\boldsymbol{v}(\boldsymbol{c}) + \boldsymbol{v}(\boldsymbol{c})\boldsymbol{U}_v^{-1}\boldsymbol{\mu}_v + \overline{K}, \end{aligned} \tag{20}$$

where the constant $\overline{K}$ is independent of $\boldsymbol{c}$. To determine $\boldsymbol{c}$, we iteratively update $\boldsymbol{c}$ with the gradient method,

$$\boldsymbol{c}^{(i+1)\text{-th}} = \boldsymbol{c}^{(i)\text{-th}} + \alpha \cdot \delta\boldsymbol{c}^{(i)\text{-th}}, \tag{21}$$

where α is a step size parameter. The following two gradient methods are basically employed for calculating the vector $\delta\boldsymbol{c}^{(i)\text{-th}}$.

Steepest decent algorithm: When using the steepest decent algorithm, $\delta\boldsymbol{c}^{(i)\text{-th}}$ is written as

$$\delta\boldsymbol{c}^{(i)\text{-th}} = \left.\frac{\partial L}{\partial \boldsymbol{c}}\right|_{\boldsymbol{c}=\boldsymbol{c}^{(i)\text{-th}}}. \tag{22}$$

The first derivative is calculated by

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{c}} &= \omega\left(-\boldsymbol{W}^{\top}\hat{\boldsymbol{U}}^{-1}\boldsymbol{W}\boldsymbol{c} + \boldsymbol{W}^{\top}\hat{\boldsymbol{U}}^{-1}\hat{\boldsymbol{\mu}}\right) \\ &\quad + \left[\boldsymbol{v}_1'^{\top}, \boldsymbol{v}_2'^{\top}, \cdots, \boldsymbol{v}_t'^{\top}, \cdots, \boldsymbol{v}_T'^{\top}\right]^{\top}, \end{aligned} \tag{23}$$

$$\boldsymbol{v}_t' = [v_t'(1), v_t'(2), \cdots, v_t'(d), \cdots, v_t'(D)]^{\top}, \tag{24}$$

$$v_t'(d) = -\frac{2}{T}\left(c_t(d) - \overline{c}(d)\right)\boldsymbol{p}_v^{(d)\top}\left(\boldsymbol{v}(\boldsymbol{c}) - \boldsymbol{\mu}_v\right), \tag{25}$$

where $\boldsymbol{p}_v^{(d)}$ is the d-th column vector of $\boldsymbol{P}_v = \boldsymbol{U}_v^{-1}$.

Newton-Raphson method: If the initial trajectory $\boldsymbol{c}^{(0)\text{-th}}$ is close to the optimum one, we may also use the Newton-Raphson method using not only the first derivative but also the second derivative, i.e., the Hessian matrix. The vector $\delta\boldsymbol{c}^{(i)\text{-th}}$ is written as

$$\delta\boldsymbol{c}^{(i)\text{-th}} = -\left(\frac{\partial^2 L}{\partial\boldsymbol{c}\partial\boldsymbol{c}^{\top}}\right)^{-1}\left.\frac{\partial L}{\partial\boldsymbol{c}}\right|_{\boldsymbol{c}=\boldsymbol{c}^{(i)\text{-th}}}. \tag{26}$$

Because the Hessian matrix is not always a positive definite matrix, the following second derivative approximated with only diagonal elements is used,

$$\frac{\partial^2 L}{\partial\boldsymbol{c}\partial\boldsymbol{c}^{\top}} \simeq -\omega\cdot\mathrm{diag}\left[\boldsymbol{r}^{\top} + \boldsymbol{v}''^{\top}\right], \tag{27}$$

$$\boldsymbol{r} = [r_1, r_2, \cdots, r_t, \cdots, r_T]^{\top}, \tag{28}$$

$$\boldsymbol{v}'' = [v_1'', v_2'', \cdots, v_t'', \cdots, v_T'']^{\top}, \tag{29}$$

$$\boldsymbol{r}_t = [r_t(1), r_t(2), \cdots, r_t(d), \cdots, r_t(D)]^{\top}, \tag{30}$$

$$\boldsymbol{v}_t'' = [v_t''(1), v_t''(2), \cdots, v_t''(d), \cdots, v_t''(D)]^{\top}, \tag{31}$$

$$r_t(d) = \boldsymbol{w}^{((t-1)D+d)\top}\hat{\boldsymbol{U}}^{-1}\boldsymbol{w}^{((t-1)D+d)}, \tag{32}$$

$$\begin{aligned} v_t''(d) &= -\frac{2}{T^2}\left\{(T-1)\boldsymbol{p}_v^{(d)\top}\left(\boldsymbol{v}(\boldsymbol{c}) - \boldsymbol{\mu}_v\right)\right. \\ &\quad \left. +2p_v^{(d)}(d)\left(c_t(d) - \overline{c}(d)\right)^2\right\}, \end{aligned} \tag{33}$$

where $\boldsymbol{w}^{(d)}$ is the d-th column vector of $\boldsymbol{W}$.

There are mainly two settings of the initial trajectory $\boldsymbol{c}^{(0)\text{-th}}$. One is to use the conventional trajectory calculated by (14). The other is to use the trajectory $\boldsymbol{c}'$ linearly converted from the conventional one as follows

$$c_t'(d) = \sqrt{\frac{\mu_v(d)}{v(d)}}\left(c_t(d) - \overline{c}(d)\right) + \overline{c}(d). \tag{34}$$

The conventional trajectory maximizes the HMM likelihood $P(\boldsymbol{o}|\hat{\boldsymbol{Q}},\boldsymbol{\lambda})$, while $\boldsymbol{c}'$ maximizes the GV likelihood $P(\boldsymbol{v}(\boldsymbol{c})|\boldsymbol{\lambda}_v)$. We found that $\boldsymbol{c}'$ usually has a larger value of the proposed likelihood than the conventional one when setting the weight ω as described above. Moreover, the Newton-Raphson method consistently has the better convergence performance compared with the steepest decent algorithm when starting from $\boldsymbol{c}'$.

3.4 Effectiveness of Considering GV

Figure 3 shows an example of generated trajectories when employing the conventional generation algorithm and the proposed one. It is shown that at a certain dimension the proposed method emphasizes the trajectory movements very much in parts beyond ranges shown as a mean ± one standard deviation of a Gaussian at each HMM state while at

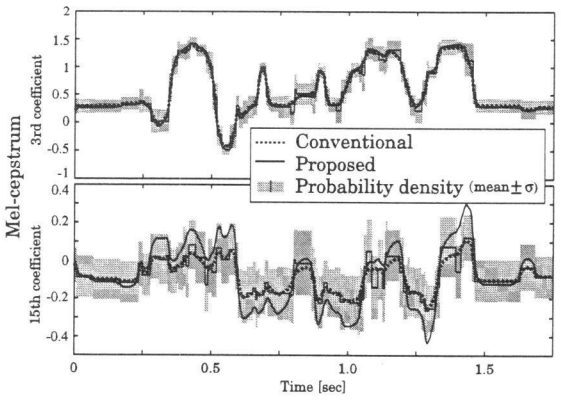


Fig. 3 An example of generated parameter trajectories with the conventional and proposed algorithms.

another dimension it keeps them almost equal to those of the conventional trajectory. Those emphasis scales vary between individual dimensions and frames, and they are automatically determined according to the proposed likelihood. This process may be regarded as statistically postfiltering.

One of advantages of the proposed algorithm is to keep the number of parameters almost equal to that in the conventional algorithm. In addition, since the proposed framework is based on the statistical process, it keeps many advantages of the HMM-based speech synthesis such as allowing model training or adaptation in the manner supported mathematically. For example, it is straightforward to apply the conventional context clustering techniques to the GV modeling. Although the computation cost for the proposed parameter generation is larger than that for the conventional one, the generation process still works enough fast to synthesize a speech waveform much faster than the real time [26].

4. Experimental Evaluations

4.1 Experimental Conditions

We trained context-dependent HMMs for each of four Japanese speakers (two males, MHT and MYI, and two females, FTK and FYM). We used 450 sentences of phonetically balanced 503 sentences from ATR Japanese speech database B-set [27] as training data for each speaker. Remaining 53 sentences for each were used for evaluations. Context-dependent labels were prepared from phoneme and linguistic labels included in the ATR database.

As a spectral parameter, we used 0^{th} through 24^{th} mel-cepstral coefficients obtained from the smoothed spectrum analyzed by a high quality speech analysis-synthesis method called Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) [9]. As a source parameter, we used a log-scaled F_0 automatically extracted from a waveform. Each of the spectral and F_0 parameter vectors included the static feature and their delta and delta-delta features. Frame shift was set to 5 ms.

The spectral part was modeled by the continuous HMM of which each state output probability density was modeled by a single Gaussian with a diagonal covariance matrix. For the F_0 part, we used the multi-space probability distribution HMM (MSD-HMM) [28] to model a time sequence consisting of continuous values, i.e., log-scaled F_0s, and discrete symbols that represent unvoiced frames. Static, delta, and delta-delta F_0s were treated in different streams. We constructed context-dependent HMMs for each part with a decision-tree based context clustering technique based on the minimum description length (MDL) criterion [29]. We also trained context-dependent duration models for modeling the state duration probabilities. A Gaussian distribution for modeling probability density of the GV for each part was trained with GVs extracted from individual training sentences.

In the synthesis, we prepared a sentence HMM for given input contexts by concatenating the context-dependent HMMs and then we determined sub-optimum state sequence based on the state duration model. A mel-cepstrum sequence was directly generated from a resulting sequence of probability density functions (pdfs). The generation process was independently performed at each dimension since diagonal covariance matrices were used. In the F_0 parameter generation, we firstly determined unvoiced frames based on the output probability of the unvoiced symbol from the MSD-HMM. Then, we generated an F_0 parameter sequence from a pdf sequence that didn't include the unvoiced frames. Inverse variances for the dynamic features were set to 0 at the boundaries between voiced and unvoiced frames. A simple excitation was constructed with a pulse train and white noise based on the generated F_0 parameters. Then, a speech waveform was synthesized with the Mel Log Spectrum Approximation (MLSA) filter [30] based on the generated mel-cepstra.

4.2 Objective Evaluations

4.2.1 Comparison of GV

Figure 4 shows GVs of generated mel-cepstra with the conventional and the proposed algorithms. As for the conventional algorithm, it also shows GVs of the generated mel-cepstra emphasized with postfiltering [31] when setting a filtering coefficient β to 0.4 and 0.8. The GV of the natural mel-cepstra is also shown in the figure as a reference. It can be seen that the GV of the mel-cepstra generated with the conventional algorithm ($\beta = 0.0$) is evidently small. Although postfiltering makes the GV large, the emphasized mel-cepstra has GV characteristics obviously different from those of the natural ones. On the other hand, the proposed algorithm generates the mel-cepstra of which the GV is almost equal to that of natural ones.

4.2.2 Comparison of Likelihoods

Figure 5 shows the log-scaled HMM and GV likelihoods

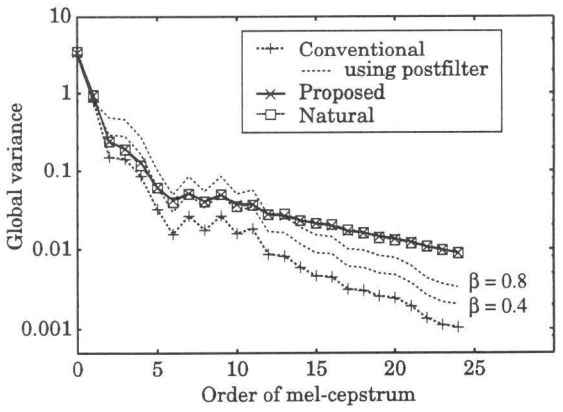


Fig. 4 GVs of several mel-cepstrum sequences. These values show GV means over all test sentences and all speakers.

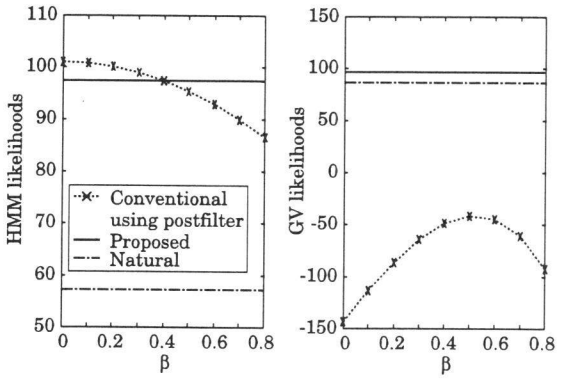


Fig. 5 Log-scaled HMM and GV likelihoods on mel-cepstrum sequences as a function of postfilter coefficient β. The HMM likelihoods are normalized by the number of frames.

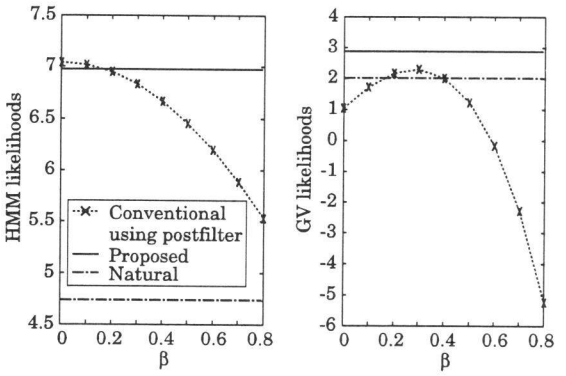


Fig. 6 Log-scaled HMM and GV likelihoods on F_0 sequences as a function of postfilter coefficient β. The HMM likelihoods are normalized by the number of frames.

on mel-cepstrum sequences. It is reasonable that the largest HMM likelihood is caused by the conventional algorithm ($\beta = 0.0$) and it decreases when applying the postfilter or considering the GV likelihood in the generation process. An interesting point is that the HMM likelihood for the natural trajectory is smaller than those for the generated trajectories. This implies that we don't necessary to generate the trajectory that maximizing only the HMM likelihood, though it seems reasonable to keep the likelihood larger at least than that for the natural trajectory.

The GV likelihood is quite small when using the conventional algorithm because of the GV reduction shown in Fig 4. Although it is recovered by postfiltering, the resulting likelihoods are still much smaller than that for the natural trajectory. On the other hand, the proposed algorithm generates the trajectory for which the GV likelihood is enough large. Consequently, both HMM and GV likelihoods for the proposed trajectory are larger than those two for the natural one.[†]

Figure 6 shows results in the F_0 sequences. We can see the same tendencies as shown in the mel-cepstrum sequences, though there are a few differences such as postfiltering for the conventional F_0 trajectory[††]also makes the GV likelihood larger than that for the natural one.

These results demonstrate that the proposed algorithm generates more similar trajectories to those of the natural speech than the conventional algorithm in view of satisfying more various characteristics.

Table 1 Synthetic voices used for an opinion test.

	Mel-cepstrum sequence	F_0 sequence
A	Conventional	Conventional
B	Conventional	Proposed
C	Proposed	Conventional
D	Proposed	Proposed
E	Natural	Natural

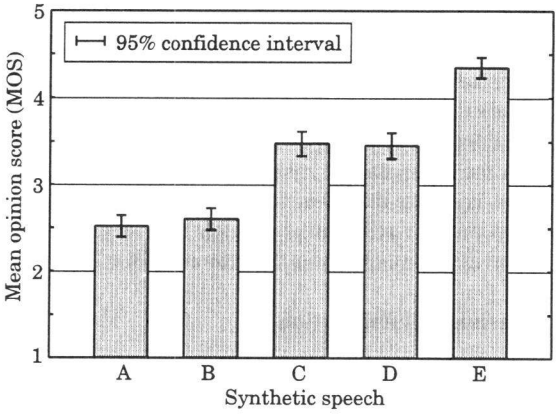


Fig. 7 Result of an opinion test.

4.3 Perceptual Evaluation

We conducted an opinion test on the naturalness of synthetic speech to demonstrate the effectiveness of the proposed method. We evaluated five kinds of voices shown in Table 1. Seven Japanese listeners participated in the test. Each listener evaluated 25 samples consisting of five sentences for each speaker. Those sentences were randomly selected for each listener from the 53 test sentences.

Figure 7 shows the result of the test. It is observed that the proposed algorithm works very well in the spectral parameter generation. It effectively reduces muffled sounds of synthetic voices caused by the over-smoothing effect. An example of spectrum sequences is shown in Fig 8. The proposed algorithm makes spectral peaks much sharper than those generated by the conventional algorithm. It is possible that a process of simply increasing the GV causes the quality degradation of synthetic speech because it doesn't always make the generated sequence close to the natural one. The proposed algorithm increases the GV considering the GV

[†]It has also been reported that considering the GV in the spectral parameter conversion between two speakers causes the better performance than postfiltering in view of not only the converted speech quality but also the conversion accuracy for speaker individuality [32].

[††]We designed a postfilter for an F_0 sequence as follows: $c'_t = (\beta + 1)(c_t(d) - \overline{c}(d)) + \overline{c}(d)$.

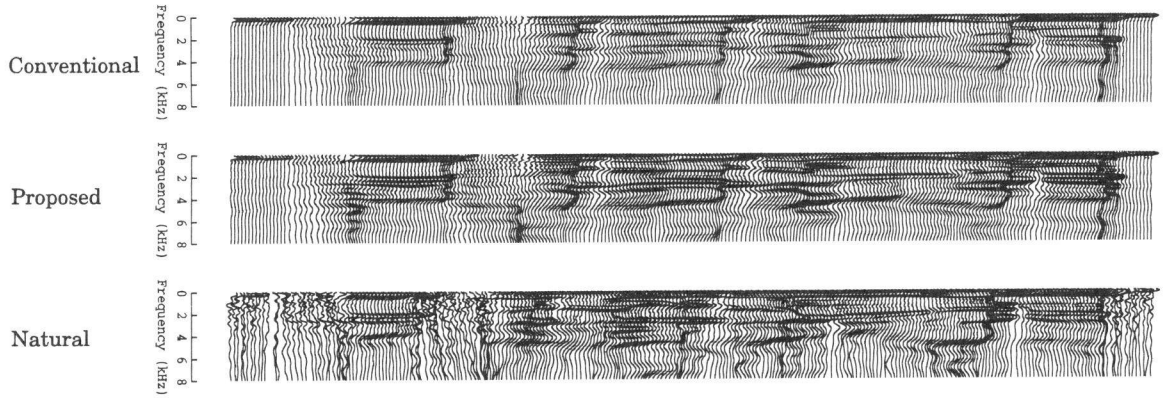


Fig. 8 An example of spectrum sequences of generated speech with conventional algorithm, generated speech with proposed algorithm, and natural speech. Note that phoneme duration of the natural sequence is different from those of the generated ones.

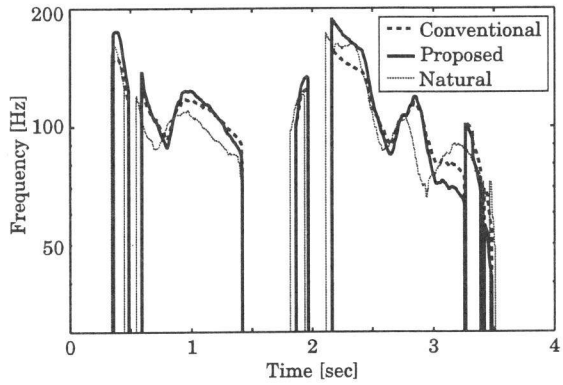


Fig. 9 An example of F_0 sequences of generated speech with conventional algorithm, generated speech with proposed algorithm, and natural speech. Standard deviation of each F_0 sequence is 21.0 Hz for conventional, 31.6 Hz for proposed, and 27.2 Hz for natural, respectively. Note that phoneme duration of the natural sequence is different from those of the generated ones.

likelihood while considering the HMM likelihood as well for alleviating the quality degradation due to increasing the GV too much.

Considering the GV in the F_0 parameter generation doesn't cause significant improvements of the naturalness of synthetic speech. An example of F_0 contours is shown in Fig 9. The proposed algorithm generates an F_0 contour with the larger GV compared with the conventional algorithm but it is not close to the natural one. The GV model seems too simple to capture natural variance characteristics of an F_0 contour. It would be useful to model the GV over not an utterance but a part of an utterance, e.g., a prosodic phrase or an accentual phrase. Moreover, it is possible that the GVs used in the training were affected by errors of the automatic F_0 extraction, especially halving and doubling, which were often observed on the extracted F_0s. Those errors would make the GV too large.

Although there still remain several problems to be improved, the proposed method certainly causes dramatic quality improvements in the HMM-based speech synthesis.[†]

5. Conclusions

We proposed a parameter generation algorithm considering global variance (GV) of the generated parameters for the HMM-based speech synthesis. The proposed algorithm generated a time sequence of static features that maximizes a likelihood based on not only an HMM likelihood for a parameter sequence of the static and dynamic features but also a likelihood for the GV under the constraint that the dynamic features and the GV were calculated from the static features. We applied this algorithm to both spectral and F_0 parameter generation processes. As a result of the perceptual evaluation, it was shown that the proposed algorithm dramatically improves the naturalness of synthetic speech.

The improved quality is still worse than the analysis-synthesized speech quality. This quality difference is caused by not only the generated spectral and F_0 parameters but also the generated duration. Further improvements of the acoustic modeling are indispensable for achieving higher-quality synthetic speech. Moreover, it is worthwhile to deal with such problems as applying context-dependent GV models, training all of model parameters including the likelihood weight based on maximizing the proposed likelihood, and applying adaptation techniques to the proposed framework.

Acknowledgment

This research was supported in part by MEXT e-Society leading project. The authors are grateful to Prof. Hideki Kawahara of Wakayama University in Japan for permission to use the STRAIGHT analysis-synthesis method and Prof. Alan W Black of Carnegie Mellon University in USA, Dr. Yoshihiko Nankaku and Dr. Heiga Zen of Nagoya Institute of Technology in Japan for helpful discussions.

[†]Several samples are available from http://spalab.naist.jp/~tomoki/IEICE/HTS+GV/index.html

References

[1] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," Proc. ICASSP, pp.679–682, New York, USA, April 1988.

[2] T. Hirokawa, "Speech synthesis using a waveform dictionary," Proc. EUROSPEECH, pp.140–143, Paris, France, Sept. 1989.

[3] D.H. Klatt, "Review of text-to-speech conversion for English," J. Acoust. Soc. Am., vol.82, no.3, pp.737–793, 1987.

[4] 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, Nov. 1993.

[5] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proc. ICASSP, pp.373–376, Atlanta, USA, May 1996.

[6] M. Isogai and H. Mizuno, "A new F_0 contour control method based on vector representation of F_0 contour," Proc. EUROSPEECH, pp.727–730, Budapest, Hungary, Sept. 1999.

[7] A. Raux and A.W. Black, "A unit selection approach to F_0 modeling and its application to emphasis," Proc. ASRU, pp.700–705, St. Thomas, USA, Dec. 2003.

[8] T. Saito, "Generating F_0 contours by statistical manipulation of natural F_0 shapes," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.1100–1106, March 2006.

[9] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," Speech Commun., vol.27, no.3-4, pp.187–207, 1999.

[10] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," IEEE Trans. Speech Audio Process., vol.9, no.1, pp.21–29, 2001.

[11] T. Toda, H. Kawai, and M. Tsuzaki, "Effectiveness of prosodic modification in concatenative Text-to-Speech synthesis," Proc. Autumn Meeting of ASJ, 1-8-10, pp.201–202, Sept. 2003.

[12] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis," Speech Commun., vol.48, no.1, pp.45–56, Jan. 2006.

[13] H. Kawai and M. Tsuzaki, "A study on time-dependent voice quality variation in a large-scale single speaker speech corpus used for speech synthesis," Proc. IEEE 2002 Workshop on Speech Synthesis, Santa Monica, U.S.A., Sept. 2002.

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

[15] M. Chu, H. Peng, H. Yang, and E. Chang, "Selecting non-uniform units from a very large corpus for concatenative speech synthesizer," Proc. ICASSP, pp.785–788, Salt Lake City, U.S.A., May 2001.

[16] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," Proc. 5th ISCA Speech Synthesis Workshop (SSW5), pp.179–184, Pittsburgh, USA, June 2004.

[17] S. Nakajima and H. Hamada, "Automatic generation of synthesis units based on context oriented clustering," Proc. ICASSP, pp.659–662, New York, USA, April 1988.

[18] T. Kagoshima and M. Akamine, "An F_0 contour control model for totally speaker driven text to speech system," Proc. ICSLP, pp.1975–1978, Sydney, Australia, Dec. 1998.

[19] M. Akamine and T. Kagoshima, "Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS Drive TTS)," Proc. ICSLP, pp.1927–1930, Sydney, Australia, Dec. 1998.

[20] T. Mizutani and T. Kagoshima, "Concatenative speech synthesis based on the plural unit selection and fusion method," IEICE Trans. Inf. & Syst., vol.E88-D, no.11, pp.2565–2572, Nov. 2005.

[21] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," Proc. IEEE 2002 Workshop on Speech Synthesis, Santa Monica, USA, Sept. 2002.

[22] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. EUROSPEECH, pp.2347–2350, Budapest, Hungary, Sept. 1999.

[23] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," Proc. ICASSP, pp.660–663, Detroit, USA, May 1995.

[24] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," Proc. EUROSPEECH, pp.757–760, Madrid, Spain, Sept. 1995.

[25] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. ICASSP, pp.1315–1318, Istanbul, Turkey, June 2000.

[26] H. Zen and T. Toda, "An overview of nitech HMM-based speech synthesis system for Blizzard Challenge 2005," Proc. Interspeech, pp.93–96, Lisbon, Portugal, Sept. 2005.

[27] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuwabara, "A large-scale Japanese speech database," ICSLP90, pp.1089–1092, Kobe, Japan, Nov. 1990.

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

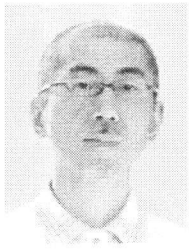
[29] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), vol.21, no.2, pp.79–86, 2000.

[30] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," Proc. ICASSP, pp.93–96, Boston, USA, April 1983.

[31] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.87-D-II, no.8, pp.1565–1571, Aug. 2004.

[32] T. Toda, A.W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," Proc. ICASSP, vol.1, pp.9–12, Philadelphia, USA, March 2005.

Tomoki Toda received the B.E. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 1999 and the M.E. and Ph.D. degrees in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan, in 2001 and 2003, respectively. During 2001–2003, he was an Intern Researcher at ATR Spoken Language Translation Research Laboratories, Kyoto, Japan. He was a Research Fellow of the Japan Society for the Promotion of Science in Graduate School of Engineering, Nagoya Institute of Technology during 2003–2005. He was a Visiting Researcher at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA, from October 2003 to September 2004. He is currently an Assistant Professor of the Graduate School of Information Science, NAIST and a Visiting Researcher at ATR Spoken Language Communication Research Laboratories. He received the TELECOM System Technology Award for Student from the Telecommunications Advancement Foundation in 2003. He is a member of the Speech and Language Technical Committee of the IEEE Signal Processing Society from January 2007. He is a member of IEEE, ISCA, and ASJ. His research interests include speech synthesis, voice conversion, speech analysis, and speech recognition.



Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr. Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004 he was an Associate Professor at the Department of Computer Science, Nagoya Institute of Technology, and now he is a Professor at the same institute. He is also an Invited Researcher at ATR Spoken Language Communication Research Laboratories (ATR-SLC), Kyoto, Japan, and was a Visiting Researcher at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA, from 2001 to 2002. He is a co-recipient of the Paper Award and the Inose Award both from the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) in 2001, and the TELECOM System Technology Award from the Telecommunications Advancement Foundation, Japan, in 2001. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. He is a member of IEEE, ISCA, IPSJ, ASJ, and JSAI. His research interests include speech coding, speech synthesis and recognition, and statistical machine learning.