Acoust. Sci. & Tech. **34**, 2 (2013)

PAPER

A generation error function considering dynamic properties of speech parameters for minimum generation error training for hidden Markov model-based speech synthesis

Duy Khanh Ninh*, Masanori Morise† and Yoichi Yamashita‡

Graduate School of Science and Engineering, Ritsumeikan University, 1–1–1 Noji-higashi, Kusatsu, 525–8577 Japan

(*Received 28 May 2012, Accepted for publication 7 November 2012*)

Abstract: A minimum generation error (MGE) criterion has been proposed for model training in hidden Markov model (HMM)-based speech synthesis to minimize the error between generated and original static parameter sequences of speech. However, dynamic properties of speech parameters are ignored in the generation error definition. In this study, we incorporate these dynamic properties into MGE training by introducing the error component of dynamic features (i.e., delta and delta-delta parameters) into the generation error function. We propose two methods for setting the weight associated with the additional error component. In the fixed weighting approach, this weight is kept constant over the course of speech. In the adaptive weighting approach, it is adjusted according to the degree of dynamicity of speech segments. An objective evaluation shows that the newly derived MGE criterion with the adaptive weighting method results in comparable performance for the static feature and better performance for the delta feature compared with the baseline MGE criterion. Subjective listening tests exhibit a small but statistically significant improvement in the quality of speech synthesized by the proposed technique. The newly derived criterion improves the capability of HMMs in capturing dynamic properties of speech without increasing the computational complexity of the training process compared with the baseline criterion.

Keywords: Statistical parametric speech synthesis, Hidden Markov model, Minimum generation error training, Generation error function, Dynamic features

PACS number: 43.72.Ne [doi:10.1250/ast.34.123]

1. INTRODUCTION

Hidden Markov model (HMM)-based statistical parametric speech synthesis (SPSS) was first proposed nearly two decades ago [1,2]. In this method, spectral and prosodic features of speech are modeled and generated in a unified statistical framework using HMMs [2,3]. It has become increasingly popular in speech synthesis research and application owing to the high flexibility in transforming voice characteristics and speaking styles, the small footprint, and the stable and high synthetic speech quality of state-of-the-art systems [4].

In HMM-based SPSS, the use of dynamic features (e.g., delta and delta-delta cepstral coefficients) [5] is crucial for generating smoothly varying parameter trajectories [2]. In the training phase, dynamic features are modeled independently with static feature. In the conventional system [3], HMM parameters are trained under the maximum likelihood (ML) criterion. In the synthesis phase, the most probable parameter sequence is generated given the distributions of static and dynamic features using the speech parameter generation algorithm [6]. Here, the constraints between static and dynamic features are taken into account to generate smooth, realistic feature trajectories. Although dynamic features of both spectral and F_0 parameters are used in HMM-based SPSS, we restrict ourselves to the use of dynamic spectral features in later discussions.

Considering the ignorance of the constraints between static and dynamic features in the training phase and the mismatch between the ML-based training criterion and the objective of speech synthesis, the minimum generation error (MGE) criterion [7] has been proposed for training HMMs in SPSS. By incorporating the parameter generation

*e-mail: ninhkd@slp.is.ritsumei.ac.jp
†e-mail: morise@fc.ritsumei.ac.jp
‡e-mail: yama@media.ritsumei.ac.jp

process into the HMM training, the error between the original and generated data for a training sentence can be calculated as a function of HMM parameters, which is called the generation error function (GEF). HMM parameters are then re-estimated to minimize the total generation error for all training sentences. As a result, the above two issues of the ML-training-based conventional system can be solved effectively, and improved synthetic speech quality has been reported [7].

A key point in MGE training is the definition of the GEF. In the baseline MGE criterion [7], the GEF is defined as the Euclidean distance between the natural and generated static feature vector sequences. Under the viewpoint of parameter trajectory modeling, this has a drawback, which is the ignorance of dynamic properties of parameter trajectories in the generation error definition. These dynamic properties are captured, to a certain extent, by the dynamic features of speech parameters since the dynamic features of a speech frame are generally calculated as regression coefficients from the static features of neighboring frames [5]. Moreover, dynamic features convey spectral transition information, which is believed to be an important acoustic cue in speech perception [8]. Therefore, we expect that the introduction of the error component of dynamic features into the GEF could be beneficial.

In this paper, we define the generation error of dynamic features and incorporate this new error component into the GEF. Then the newly derived GEF is minimized under the MGE criterion. It is worth pointing out that the objective of our research is different from those of recent improvements of the baseline MGE criterion [9–11]. In [9], the error component of the global/local variance of feature trajectories was introduced into the GEF to obtain an over-smoothing alleviation effect similar to the parameter generation algorithm considering global variance (GV) [12] without introducing any extra computational cost during synthesis. In [10] and [11], two perceptually motivated distance metrics on line spectral pairs (LSPs), log spectral distortion and weighted Euclidean distance, respectively, were proposed to enhance the correlation between the objective GEF and the subjective perception of spectral distortion. In contrast, our research aims to investigate the effect of more accurately modeling the dynamic properties of speech parameters under the MGE training framework.

The rest of the paper is organized as follows. Section 2 reviews the baseline MGE criterion. Our proposed MGE criterion considering dynamic properties of speech parameters is described in Sect. 3. Section 4 presents experimental results. Section 5 comprises several comments. Finally, conclusions are given in Sect. 6.

2. MINIMUM GENERATION ERROR CRITERION

This section gives a review of the baseline MGE criterion [7]. In HMM-based SPSS, the speech parameter generation algorithm [6] is used to generate the most probable feature vector sequence. Then the HMM parameters are optimized to minimize the total generation error of all training data under the MGE criterion. The following subsections follow the notations and formulations in [6] and [9] for the sake of coherence and compactness.

2.1. Speech Parameter Generation Algorithm

For a given HMM λ and a state sequence $\boldsymbol{q}$, the speech parameter generation algorithm aims to generate the parameter vector sequence $\boldsymbol{o} = [\boldsymbol{o}_1^{\mathrm{T}}, \boldsymbol{o}_2^{\mathrm{T}}, \ldots, \boldsymbol{o}_T^{\mathrm{T}}]^{\mathrm{T}}$ by maximizing $P(\boldsymbol{o}|\lambda, \boldsymbol{q})$ with respect to $\boldsymbol{o}$ [6] (Case 1), where T is the number of frames in $\boldsymbol{o}$ and T denotes the matrix transpose operation. The tth frame's parameter vector $\boldsymbol{o}_t$ includes the M-dimensional static feature vector $\boldsymbol{c}_t$ and dynamic feature vectors $\Delta^{(1)}\boldsymbol{c}_t$ and $\Delta^{(2)}\boldsymbol{c}_t$ (i.e., delta and delta-delta coefficients, respectively), and can be written as $\boldsymbol{o}_t = [\boldsymbol{c}_t^{\mathrm{T}}, \Delta^{(1)}\boldsymbol{c}_t^{\mathrm{T}}, \Delta^{(2)}\boldsymbol{c}_t^{\mathrm{T}}]^{\mathrm{T}}$, where the dth-order dynamic feature vector is calculated as

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

Here, $w^{(d)}(\cdot)$ are window coefficients; $L_-^{(d)}$ and $L_+^{(d)}$ are the numbers of frames preceding and succeeding frame t involved in the calculation of $\Delta^{(d)}\boldsymbol{c}_t$ ($d = 1, 2$), respectively.

From Eq. (1), the constraints between static and dynamic features can be expressed as $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$, where $\boldsymbol{c} = [\boldsymbol{c}_1^{\mathrm{T}}, \boldsymbol{c}_2^{\mathrm{T}}, \ldots, \boldsymbol{c}_T^{\mathrm{T}}]^{\mathrm{T}}$ and $\boldsymbol{W}$ is a $3MT \times MT$ window matrix defined as

$$\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_T]^{\mathrm{T}} \otimes \boldsymbol{I}_{M\times M}, \tag{2}$$

$$\boldsymbol{W}_t = [\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}], \tag{3}$$

$$\boldsymbol{w}_t^{(0)} = [\underbrace{0, \ldots, 0}_{t-1}, 1, \underbrace{0, \ldots, 0}_{T-t}]^{\mathrm{T}}, \tag{4}$$

$$\boldsymbol{w}_t^{(d)} = [\underbrace{0, \ldots, 0}_{t-L_-^{(d)}-1}, w^{(d)}(-L_-^{(d)}), \ldots, w^{(d)}(0), \ldots, w^{(d)}(L_+^{(d)}), \underbrace{0, \ldots, 0}_{T-(t+L_+^{(d)})}]^{\mathrm{T}}, \tag{5}$$

where $\otimes$ denotes the Kronecker product operation and $\boldsymbol{I}$ is the identity matrix.

Under these constraints, finding $\boldsymbol{o}$ that maximizes $P(\boldsymbol{o}|\lambda, \boldsymbol{q})$ is equivalent to finding $\boldsymbol{c}$ that maximizes $P(\boldsymbol{o}|\lambda, \boldsymbol{q})$. By solving $\partial P(\boldsymbol{o}|\lambda, \boldsymbol{q})/\partial \boldsymbol{c} = 0$, the generated static feature vector sequence is obtained as

$$\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{R}_{\boldsymbol{q}}^{-1}\boldsymbol{r}_{\boldsymbol{q}}, \tag{6}$$

where

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}, \tag{7}$$

$$\boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{\mu}_{\boldsymbol{q}}, \tag{8}$$

and $\boldsymbol{\mu}_{\boldsymbol{q}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{q}}$ are the mean vector and covariance matrix related to $\boldsymbol{q}$, respectively [6].

2.2. Minimum Generation Error Criterion

In the baseline MGE criterion, the Euclidean distance is used to measure the error between the original and generated static feature vector sequences, which is

$$D(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\boldsymbol{q}}) = \|\boldsymbol{c} - \bar{\boldsymbol{c}}_{\boldsymbol{q}}\|^2. \tag{9}$$

Theoretically, all possible state sequences underlying the original parameter vector sequence $\boldsymbol{o}$ could be involved in the calculation of the generation error, where $P(\boldsymbol{q}|\lambda, \boldsymbol{o})$ can be used to weight the error corresponding to the state sequence $\boldsymbol{q}$. However, this is computationally expensive. In practice, only the most probable state sequence $\hat{\boldsymbol{q}}$ for $\boldsymbol{o}$ is used and the GEF is defined as

$$e(\boldsymbol{c}, \lambda) = D(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\hat{\boldsymbol{q}}}). \tag{10}$$

For notational convenience, we use $\boldsymbol{q}$ to denote $\hat{\boldsymbol{q}}$ in the rest of the paper.

The MGE criterion aims to minimize the total generation error for all training sentences,

$$\hat{\lambda} = \arg\min_{\lambda} \sum_{n} e(\boldsymbol{c}_n, \lambda), \tag{11}$$

with respect to

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1^{\mathrm{T}}, \boldsymbol{\mu}_2^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_K^{\mathrm{T}}]^{\mathrm{T}}, \tag{12}$$

$$\boldsymbol{U} = [\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1}, \ldots, \boldsymbol{\Sigma}_K^{-1}]^{\mathrm{T}}, \tag{13}$$

where $\boldsymbol{c}_n$ is the static feature vector sequence of the nth training sentence, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the kth unique Gaussian component, respectively, and K is the total number of Gaussian components in the model set λ.

For each training data $\boldsymbol{c}_n$, the model parameters are updated using the probabilistic descent (PD) method [13] as

$$\lambda(n+1) = \lambda(n) - \varepsilon_n \nabla e(\boldsymbol{c}_n, \lambda)|_{\lambda=\lambda(n)}, \tag{14}$$

where ε_n is the learning rate, which decreases when the sentence index n increases.

The derivatives of the GEF with respect to the model set's mean and variance parameters can be derived as

$$\frac{\partial e(\boldsymbol{c}_n, \lambda)}{\partial \boldsymbol{\mu}} = 2\boldsymbol{S}_{\boldsymbol{q}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1}\boldsymbol{W}\boldsymbol{R}_{\boldsymbol{q}}^{-1}\boldsymbol{\varsigma}, \tag{15}$$

$$\frac{\partial e(\boldsymbol{c}_n, \lambda)}{\partial \boldsymbol{U}} = 2\boldsymbol{S}_{\boldsymbol{q}}^{\mathrm{T}}\,\mathrm{diag}^{-1}(\boldsymbol{W}\boldsymbol{R}_{\boldsymbol{q}}^{-1}\boldsymbol{\varsigma}(\boldsymbol{\mu}_{\boldsymbol{q}} - \boldsymbol{W}\bar{\boldsymbol{c}}_{\boldsymbol{q}})^{\mathrm{T}}), \tag{16}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{q}}^{-1} = \mathrm{diag}(\boldsymbol{S}_{\boldsymbol{q}}\boldsymbol{U}), \tag{17}$$

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \boldsymbol{S}_{\boldsymbol{q}}\boldsymbol{\mu}, \tag{18}$$

$$\boldsymbol{\varsigma} = \bar{\boldsymbol{c}}_{\boldsymbol{q}} - \boldsymbol{c}_n. \tag{19}$$

In the above equations, $\boldsymbol{S}_{\boldsymbol{q}}$ is a $3MT \times 3MK$ matrix whose elements are 0 or 1, determined according to the optimal state sequence $\boldsymbol{q}$ for $\boldsymbol{c}_n$, diag(.) is the operation to convert a $3MT \times 3M$ matrix to a $3MT \times 3MT$ block-diagonal matrix with a block size of $3M$, and $\mathrm{diag}^{-1}(.)$ is the inverse operation of diag(.).

It should be noted that the highest computational cost of MGE training is related to the calculation of $\boldsymbol{R}_{\boldsymbol{q}}^{-1}$ in Eqs. (15) and (16). This cost can be markedly reduced by using an approximation of $\boldsymbol{R}_{\boldsymbol{q}}^{-1}$ taking into account the special structure of the matrix $\boldsymbol{R}_{\boldsymbol{q}}$ [7].

3. MGE CRITERION WITH DYNAMIC FEATURES

The baseline MGE criterion described in the previous section has a drawback, which is the ignorance of dynamic properties of speech parameters in the generation error definition. In this section, we incorporate these dynamic properties into MGE training by defining the generation error of dynamic features and introducing this new error component into the GEF. Controlling the weight associated with the newly added error component is of essential importance in balancing the performance of the two error components comprising the newly derived GEF. We propose two methods for setting this weight: fixed and adaptive weighting. The effects of these two weighting methods are discussed in Sect. 4.

The generation error of the dth-order dynamic feature is defined as the Euclidean distance between the dth-order dynamic feature vector sequences derived from the corresponding original and generated static feature vector sequences, that is,

$$D(\Delta^{(d)}\boldsymbol{c}, \Delta^{(d)}\bar{\boldsymbol{c}}_{\boldsymbol{q}}) = \|\Delta^{(d)}\boldsymbol{c} - \Delta^{(d)}\bar{\boldsymbol{c}}_{\boldsymbol{q}}\|^2, \tag{20}$$

where

$$\Delta^{(d)}\boldsymbol{c} = \boldsymbol{A}_d\boldsymbol{c}, \tag{21}$$

$$\Delta^{(d)}\bar{\boldsymbol{c}}_{\boldsymbol{q}} = \boldsymbol{A}_d\bar{\boldsymbol{c}}_{\boldsymbol{q}}. \tag{22}$$

Here, $\boldsymbol{A}_d$ is an $MT \times MT$ window matrix defined as

$$\boldsymbol{A}_d = [\boldsymbol{w}_1^{(d)}, \boldsymbol{w}_2^{(d)}, \ldots, \boldsymbol{w}_T^{(d)}]^{\mathrm{T}} \otimes \boldsymbol{I}_{M\times M}, \tag{23}$$

where $\boldsymbol{w}_t^{(d)}$ is given by Eq. (5).

The new GEF incorporating the original error component of the static feature and that of the delta feature (i.e., the first-order dynamic feature) is defined as

$$e'(\boldsymbol{c}, \lambda) = D(\boldsymbol{c}, \bar{\boldsymbol{c}}_{\boldsymbol{q}}) + aD(\Delta^{(1)}\boldsymbol{c}, \Delta^{(1)}\bar{\boldsymbol{c}}_{\boldsymbol{q}}), \tag{24}$$

where a is the weight associated with the error component of the delta feature used to control the balance between the two error components. The effects of different values of a are discussed in the next section.

It should be noted that the GEF given by Eq. (24) can be extended to higher-order dynamic feature(s) in a straightforward way. We refer to the MGE criterion with the new GEF incorporating dynamic feature(s) as *MGE-dynamics* for brevity.

3.1. Fixed Weighting Approach to MGE-dynamics

In this weighting approach, the delta weight a is kept unchanged over all training data. By substituting Eqs. (21) and (22) into Eq. (20), the derivatives of the new GEF with respect to the mean and variance parameters can be obtained as

$$\frac{\partial e'(\boldsymbol{c}_n, \lambda)}{\partial \boldsymbol{\mu}} = 2\boldsymbol{S}_q^{\mathrm{T}} \boldsymbol{\Sigma}_q^{-1} \boldsymbol{W} \boldsymbol{R}_q^{-1} \boldsymbol{P} \boldsymbol{\varsigma}, \tag{25}$$

$$\frac{\partial e'(\boldsymbol{c}_n, \lambda)}{\partial \boldsymbol{U}} = 2\boldsymbol{S}_q^{\mathrm{T}} \operatorname{diag}^{-1}(\boldsymbol{W} \boldsymbol{R}_q^{-1} \boldsymbol{P} \boldsymbol{\varsigma} (\boldsymbol{\mu}_q - \boldsymbol{W} \bar{\boldsymbol{c}}_q)^{\mathrm{T}}), \tag{26}$$

where

$$\boldsymbol{P} = \boldsymbol{I} + a\boldsymbol{A}_1^{\mathrm{T}} \boldsymbol{A}_1. \tag{27}$$

Here, $\boldsymbol{A}_1$ is given by Eq. (23) when d is equal to one.

Comparing the newly formulated updating rules (Eqs. (25) and (26)) with the original ones (Eqs. (15) and (16)), we see that a matrix factor $\boldsymbol{P}$ is added as a consequence of the introduction of the delta feature error component into the GEF. It can be seen that $\boldsymbol{P}$ is a constant matrix for a given number of frames T of a training sentence and delta weight a. Hence, the MGE-dynamics criterion with the fixed weighting approach gives rise to no additional computational complexity compared with the baseline MGE criterion.

3.2. Adaptive Weighting Approach to MGE-dynamics

The above fixed weighting approach has the effect of assigning an equal delta weight to every time sample of an utterance. However, speech consists of stationary and transitional parts, and transitional parts possess higher dynamicity than stationary ones. This suggests that the error component of the delta feature corresponding to transitional parts should be given more emphasis than that corresponding to stationary ones. Therefore, in this subsection, we propose an adaptive weighting approach where the delta weight is adjusted according to the degree of dynamicity of speech segments. The effects of emphasizing transitional or stationary parts of speech have also been reported in speech recognition [14].

To characterize the degree of dynamicity of speech segments, we propose to divide speech signals into portions corresponding to the state boundaries of phoneme HMMs found by Viterbi decoding [15]. The reason for this is twofold. Firstly, the states of phoneme HMMs provide natural subphonetic boundaries and may contain dynamic information of speech segments. Secondly, the updating rules of MGE training (i.e., Eqs. (15) and (16) as well as Eqs. (25) and (26)) are basically performed on a statewise basis, which means that the original and generated feature vectors of frames belonging to an HMM state are used to re-estimate the model parameters of that state HMM.

We use the formulation proposed in [14] to estimate the degree of dynamicity of a frame t, that is,

$$DF_t = \sum_{p=1}^{M-1} |\Delta^{(1)} c_t(p)|, \tag{28}$$

where $\Delta^{(1)} c_t(p)$ is the pth component of the M-dimensional delta feature vector of frame t. Note that the zeroth component of this vector, which is related to the log-energy of the speech signal, is excluded from the sum.

We propose to calculate the degree of dynamicity of an HMM state s as the average of the degree of dynamicity of all frames belonging to that state, i.e.,

$$DS_s = \frac{1}{T_s} \sum_{t=t_s}^{t_s+T_s-1} DF_t, \tag{29}$$

where T_s is the number of consecutive frames, starting from frame t_s, belonging to state s.

Equation (29) gives a rough estimate of the degree of dynamicity of an HMM state considered as a speech segment. Figure 1 shows an example where the degree of dynamicity of frames and that of HMM states for an utterance are illustrated on the same plot. Here, a three-state left-to-right no-skip HMM structure was used. This example indicates that the formulation in Eqs. (28) and (29) can capture, to a certain extent, the degree of dynamicity of speech segments, even for portions where a sudden change in spectral dynamics occurs (e.g., the middle part of the stop consonant /k/).

Then, the delta weight for all frames belonging to an HMM state s is set according to the degree of dynamicity of that state as

$$a_s = a_{\max} \frac{DS_s}{DS_{\max}}, \tag{30}$$

where $a_{\max}$ is the maximum delta weight, which is assigned to the state possessing the maximum degree of dynamicity $DS_{\max}$ of all HMM states in the training data.

Finally, the same updating rules as those in the fixed weighting approach can be reused, although matrix $\boldsymbol{P}$ in Eq. (27) should be reformulated appropriately since the delta weight is adjusted state-by-state according to Eq. (30). Specifically, we assume that the nth training sentence $\boldsymbol{c}_n$ has T frames belonging to N HMM states, where state i has state duration T_i, i.e., $T = \sum_{i=1}^{N} T_i$. The

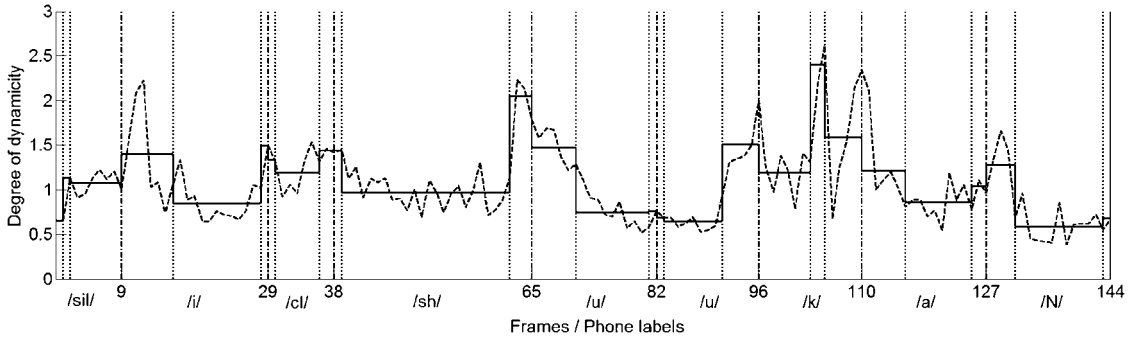


Fig. 1 Degree of dynamicity of frames (ragged dashed line) and that of HMM states (stepwise solid line) for the Japanese utterance /sil-i-cl-sh-u-u-k-a-N/ ("one week" in English). Vertical dotted lines show HMM state boundaries and vertical dash-dotted lines associated with frame numbers show HMM phoneme boundaries.

$MT \times MT$ banded matrix $\boldsymbol{A}_1$ given in Eq. (23) can be approximated as a block-diagonal matrix having block sizes that change according to the durations of HMM states, that is,

$$\boldsymbol{A}_1 \simeq \boldsymbol{A}_{1,1} \oplus \boldsymbol{A}_{1,2} \oplus \cdots \oplus \boldsymbol{A}_{1,N}, \tag{31}$$

where $\oplus$ denotes the direct sum operation, and $\boldsymbol{A}_{1,i}$ is an $MT_i \times MT_i$ matrix having a similar composition to $\boldsymbol{A}_1$.

Similarly, $\boldsymbol{A}_1^{\mathrm{T}}$ can be approximated as

$$\boldsymbol{A}_1^{\mathrm{T}} \simeq \boldsymbol{A}_{1,1}^{\mathrm{T}} \oplus \boldsymbol{A}_{1,2}^{\mathrm{T}} \oplus \cdots \oplus \boldsymbol{A}_{1,N}^{\mathrm{T}}. \tag{32}$$

The delta weight a in Eq. (27) is now adjusted state-by-state in the adaptive weighting approach. Thus, the matrix $a\boldsymbol{I}_{MT\times MT}$ appearing implicitly in Eq. (27) can be rewritten as

$$\begin{aligned} a\boldsymbol{I}_{MT\times MT} = {} & a_1\boldsymbol{I}_{MT_1\times MT_1} \oplus a_2\boldsymbol{I}_{MT_2\times MT_2} \\ & \oplus \cdots \oplus a_N\boldsymbol{I}_{MT_N\times MT_N}, \end{aligned} \tag{33}$$

where a_i is the delta weight for HMM state i, which is determined by Eq. (30).

From Eqs. (31)–(33), the matrix $\boldsymbol{P}$ in Eq. (27) can be reformulated as

$$\begin{aligned} \boldsymbol{P} \simeq {} & \boldsymbol{I} + (a_1\boldsymbol{A}_{1,1}^{\mathrm{T}}\boldsymbol{A}_{1,1} \oplus a_2\boldsymbol{A}_{1,2}^{\mathrm{T}}\boldsymbol{A}_{1,2} \\ & \oplus \cdots \oplus a_N\boldsymbol{A}_{1,N}^{\mathrm{T}}\boldsymbol{A}_{1,N}). \end{aligned} \tag{34}$$

Equation (34) allows us to compute $\boldsymbol{P}$ for a given training sentence in a straightforward manner if the delta weight for each HMM state has already been obtained following Eq. (30). It can be concluded that the MGE-dynamics criterion with adaptive weighting has similar computational complexity to that with fixed weighting and the baseline MGE criterion, since the highest computational cost of the training process is still related to the calculation of $\boldsymbol{R}_{\boldsymbol{q}}^{-1}$.

4. EXPERIMENTS

To evaluate the effectiveness of our two proposed methods, i.e., the MGE-dynamics criterion with fixed weighting (MGE-dynamics-FW) and the MGE-dynamics criterion with adaptive weighting (MGE-dynamics-AW), we carried out evaluation experiments to compare the performance of HMMs trained by the proposed techniques with that of HMMs trained by the baseline MGE technique.

4.1. Experimental Conditions

We used 503 phonetically balanced sentences uttered by the male speaker MHT from the ATR Japanese speech database (B-set) [16] in the experiments. The first 450 sentences were used for training, and the remaining 53 sentences were used for testing. Speech signals were sampled at 16 kHz and windowed by a 25 ms Hamming window with a 5 ms shift. The feature vector consists of static feature, including the 0th through 24th mel-cepstral coefficients obtained by a mel-cepstral analysis technique [17] and the logarithm of F_0, delta and delta-delta features. A three-state left-to-right no-skip HMM structure was used. Each state output distribution was composed of spectrum and F_0 streams. The spectrum stream was modeled by single multivariate Gaussian distributions with diagonal covariance matrices. The F_0 stream was modeled by multispace probability distributions [18]. In the synthesis part, the mel log spectrum approximation (MLSA) filter [19] was used to synthesize the speech waveform from the generated mel-cepstral coefficients and F_0 values. Our experiments were based on the HTS toolkit [20], where the dynamic feature vectors are calculated as

$$\Delta^{(1)}\boldsymbol{c}_t = 0.5(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1}), \tag{35}$$

$$\Delta^{(2)}\boldsymbol{c}_t = \boldsymbol{c}_{t+1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t-1}. \tag{36}$$

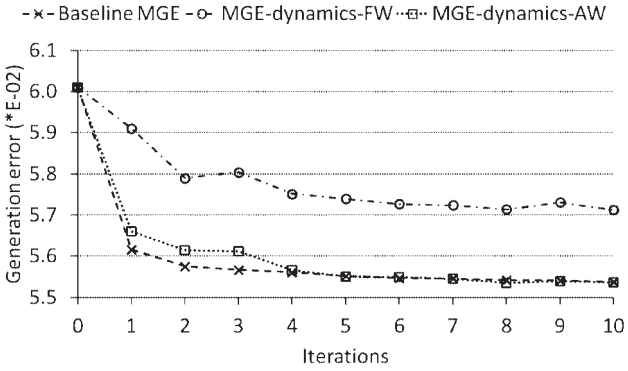
The HMM training procedure was conducted as follows. Firstly, HMMs were trained based on the conventional ML criterion [3]. Then, the resulting HMMs were used as the initial models for MGE-based training techniques. These ML-trained HMMs were also utilized to obtain the optimal state sequences for all training sentences with the Viterbi algorithm. Finally, each MGE training technique was

performed iteratively until the total generation error of static and delta features for all training data converged. The online PD updating strategy on a sentence-by-sentence basis, as described in Sect. 2.2, was adopted for its insensitivity to the learning rate and ease of implementation [21]. The learning rate ε_n in Eq. (14) was empirically set as

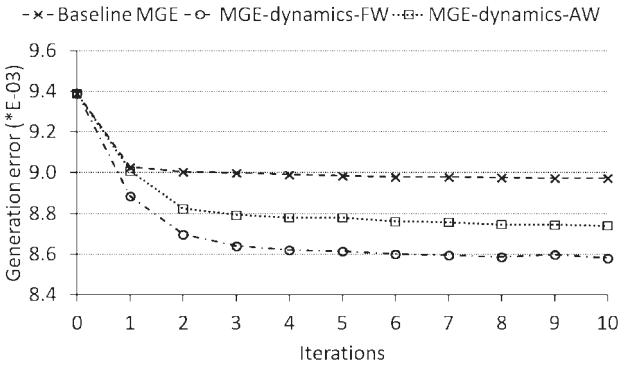
$$\varepsilon_n = \frac{1}{100 + n}. \tag{37}$$

For each training sentence, HMM parameters related to the optimal state sequence were re-estimated state-by-state using the updating rules. In the experiments, only spectral model parameters were updated as the effect of dynamic spectral features is of interest in this paper.

We first objectively and subjectively evaluated the three MGE training techniques with the original parameter generation algorithm without postprocessing [6], which was used in the MGE training framework. Then we conducted an additional subjective evaluation with the parameter generation algorithm considering GV [12] to show their effectiveness under this highly effective synthesis scheme. No-silence GV modeling in which the GV weight was set to 1.0 was used. In all evaluation experiments, the optimal state sequences obtained by the forced alignment of the original speech features with the ML-trained HMMs were used for synthesis.



(a) Generation error of static feature.



(b) Generation error of delta feature.

Fig. 2 Example of the evolution of generation error of the 2nd mel-cepstral coefficient.

4.2. Experimental Results

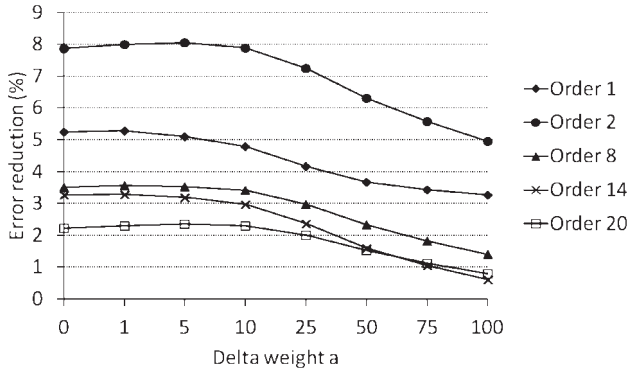
4.2.1. Evaluation with the original parameter generation algorithm without postprocessing

4.2.1.1. Objective evaluation

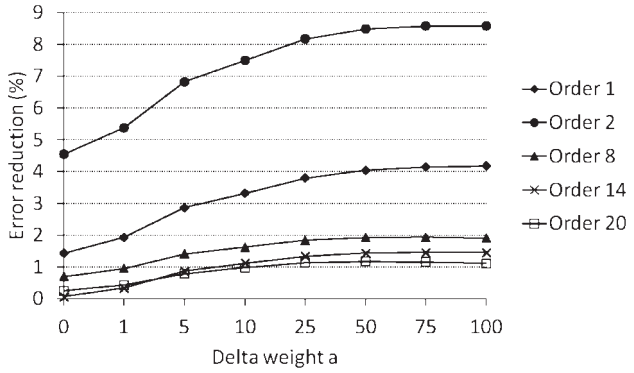
Since multivariate Gaussian distributions with diagonal covariance matrices were used for spectral parameter modeling, HMM parameters were optimized and the generation error was calculated independently for each dimension of mel-cepstral coefficients. With the above setting of the learning rate, it was observed that all MGE training techniques under investigation converged within 10 iterations. Figure 2 shows plots of the evolution of the generation error of the 2nd mel-cepstral coefficient on the test data as an example, where the delta weight for MGE-dynamics-FW and the maximum delta weight for MGE-dynamics-AW were both set to 100. Similar evolutions were also observed for other dimensions, on the training data, and for other settings related to the delta weight.

Figure 3 shows the relative error reduction (the performance of ML training was used as the reference) of static and delta features on the test data for several representative mel-cepstrum orders after MGE-dynamics-FW training with various settings of the delta weight. When the delta weight was set to zero, we obtained the result of baseline MGE training. It can be seen that baseline MGE training reduces the generation error of the delta feature as a side effect, although its GEF does not incorporate the delta feature. When the delta weight is increased, the relative error reduction of the static feature exhibits a steady downward trend while that of the delta feature increases and saturates as the delta weight approaches 100. This trade-off between the performances of error components included in the GEF was also observed in MGE-dynamics-FW training when the delta-delta feature was incorporated into its GEF in a similar manner to that described in Sect. 3. Considering the lower significance of the delta-delta feature compared with its static and delta counterparts, it is sufficient to investigate the effect of introducing the delta feature, without considering higher-order dynamic features, into MGE training in this paper.

To compare the performance of MGE-dynamics-AW with those of MGE-dynamics-FW and baseline MGE, we performed training experiments in which the delta weight for MGE-dynamics-FW and the maximum delta weight for MGE-dynamics-AW were both set to 100. Figure 4 shows the relative error reduction of the static and delta features on the test data for these MGE training techniques. Compared with baseline MGE, MGE-dynamics-AW has a comparable relative error reduction of the static feature

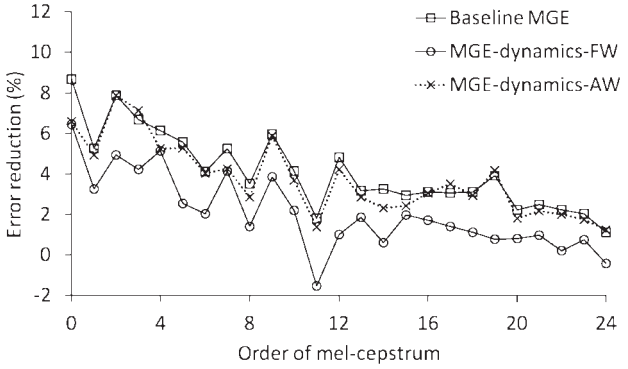


(a) Relative error reduction of static feature.

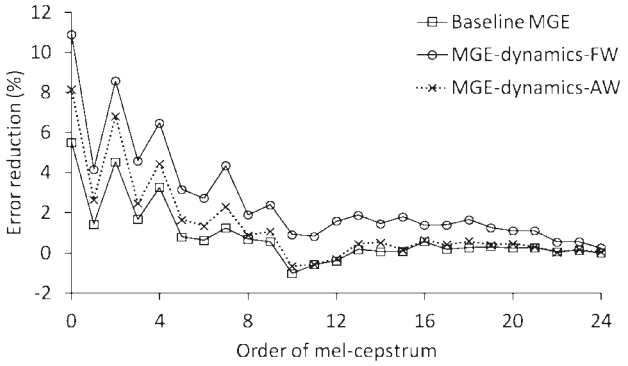


(b) Relative error reduction of delta feature.

Fig. 3 Performance of MGE-dynamics-FW with different delta weights on test data for several mel-cepstrum orders. Other orders showed similar trends but are not plotted here for readability.



(a) Relative error reduction of static feature.



(b) Relative error reduction of delta feature.

Fig. 4 Performances of three MGE training techniques on test data. The delta weight for MGE-dynamics-FW and the maximum delta weight for MGE-dynamics-AW were both set to 100.

and a larger relative error reduction of the delta feature. Compared with MGE-dynamics-FW, MGE-dynamics-AW has a larger relative error reduction of the static feature but a smaller relative error reduction of the delta feature. It can be seen that MGE-dynamics-AW alleviates the trade-off effect observed in MGE-dynamics-FW. Since both MGE-dynamics-AW and MGE-dynamics-FW obtain better performance on the delta feature than baseline MGE, we can conclude that the MGE-dynamics criterion improves the capability of HMMs in capturing dynamic properties of speech over the baseline MGE criterion.

Figure 5 illustrates an example of the trajectory of the 2nd mel-cepstral coefficient of natural speech included in the training data and those generated from HMMs trained by the baseline and proposed MGE training techniques. It can be seen that the generated trajectories from our proposed techniques almost always have more similar dynamics to the natural trajectory than that from the baseline MGE (the clearest improvements in the dynamics can be observed around the 45th and 305th frames in the figure).

4.2.1.2. Subjective evaluation

We also carried out subjective listening tests to evaluate the effectiveness of the two proposed techniques. Two preference tests were conducted. In the first test, MGE-dynamics-FW was compared with baseline MGE. In the second test, MGE-dynamics-AW was compared with baseline MGE. The settings related to the delta weight for MGE-dynamics-FW and MGE-dynamics-AW were the same as those in the previous experiment. Twelve Japanese listeners participated in the tests. They were presented with pairs of synthesized speech in random order, and asked to choose which one sounded better or to give an answer of "No preference" if the stimuli sounded the same. For each listener, 20 test sentences were randomly selected from the evaluation set consisting of 53 sentences.

Table 1 shows the results of the two preference tests. It can be seen that the difference in preference between MGE-dynamics-FW and baseline MGE is insignificant (28.3% vs 29.6%), whereas MGE-dynamics-AW has a higher preference score than baseline MGE (32.9% vs 27.1%). Furthermore, the result of a paired one-tailed *t*-test [22] indicates that the mean preference score of MGE-dynamics-AW is statistically significantly greater than that of baseline MGE at a 5% significance level (p-value = 0.031). For the second listening test, informal feedback from the test subjects suggested that the utterances

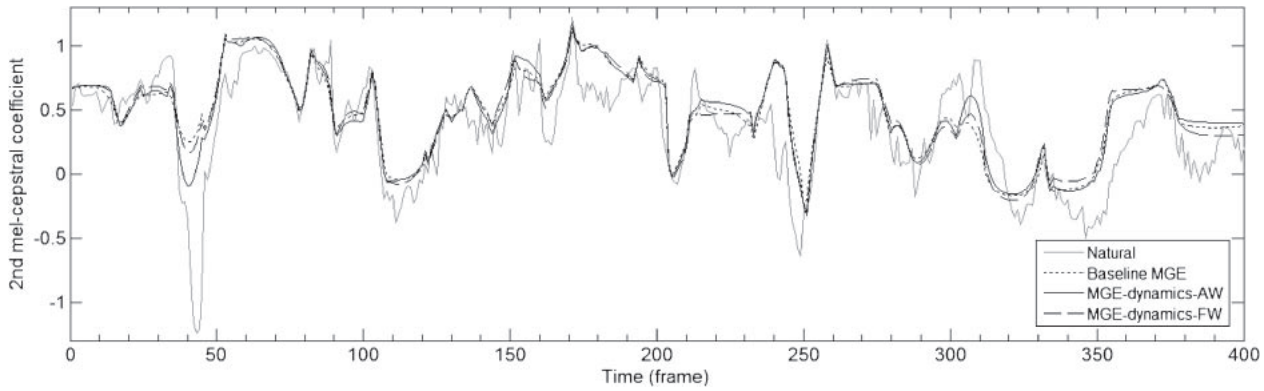


Fig. 5 Natural and generated trajectories of the 2nd mel-cepstral coefficient for an utterance included in the training data.

Table 1 Mean preference score (with 95% confidence interval) in evaluation with the original parameter generation algorithm.

	No preference (%)	Baseline (%)	Proposed (%)
Baseline vs dynamics-FW	42.1 ± 16.4	29.6 ± 9.7	28.3 ± 10.3
Baseline vs dynamics-AW	40.0 ± 12.6	27.1 ± 6.1	32.9 ± 7.8

Table 2 Mean preference score (with 95% confidence interval) in evaluation with the GV technique.

	No preference (%)	Baseline (%)	Proposed (%)
Baseline vs dynamics-FW	58.7 ± 15.7	17.1 ± 8.0	24.2 ± 10.9
Baseline vs dynamics-AW	55.0 ± 14.1	17.9 ± 6.1	27.1 ± 9.6

synthesized using the models trained by MGE-dynamics-AW and baseline MGE had almost the same naturalness; however, the clearness of some parts of utterances belonging to the MGE-dynamics-AW category was improved. The higher preference score of MGE-dynamics-AW than baseline MGE could be interpreted as the consequence of improved objective performance of the delta feature while maintaining an objective performance of the static feature comparable with that of baseline MGE.

4.2.2. Evaluation with the parameter generation algorithm considering GV

We additionally conducted two preference tests similar to those described in the preceding subsection. The only difference from the previous tests is that GV was considered in the synthesis process. Table 2 shows the results of the preference tests with the GV technique. Although the "No preference" rates increase by around 15% compared with those in the previous tests, both of the proposed techniques have a higher preference score than the baseline MGE. The results of paired one-tailed t-tests indicate that while the preference of MGE-dynamics-AW over baseline MGE is again significant at a 5% significance level (p-value $= 0.012$), the preference of MGE-dynamics-FW over baseline MGE is much less significant (p-value $= 0.090$). The visual inspection of several samples revealed that while enhancing the dynamic range of the generated parameter trajectory, the GV technique seems to keep the trajectory dynamics similar to that in the case without considering GV.

5. DISCUSSION

Although a preference test between MGE-dynamics-AW and MGE-dynamics-FW was not conducted, it is necessary to point out the merit of the former over the latter. MGE-dynamics-AW provides a data-driven technique of determining the delta weight for each portion of speech, provided that the maximum delta weight is set appropriately. In contrast, one must manually tune the delta weight to obtain the best performance for MGE-dynamics-FW. Moreover, considering the fact that HMM parameters are optimized independently for each dimension owing to the use of diagonal covariance matrices in spectral parameter modeling, MGE-dynamics-AW is likely to result in improved dynamics occurring synchronously among the dimensions since the delta weight is adjusted segment-by-segment over the course of an utterance. This synchronously improved dynamic property is less likely to occur in MGE-dynamics-FW because the delta weight is kept constant over the entire speech. Whether synchronously adaptive control among the dimensions is effective for spectral dynamics representation is still unclear. More work is needed to confirm this.

Our work also exhibits some limitations. First, the high-quality vocoder STRAIGHT [23] was not used since the magnitude of several speech samples synthesized by

the STRAIGHT filter exceeded the data range specified for a 16-bit WAV sound file. Compared with mel-cepstral vocoding, the relative error reduction of the MGE training techniques had similar trends but slightly lower magnitudes when STRAIGHT was used. Second, the more popular five-state HMM structure was not employed in our analysis because we found the use of the three-state one to be sufficient to capture the degree of dynamicity of speech segments. The use of five-state phoneme HMMs resulted in overdetailed representation of the proposed degree of dynamicity of HMM-state-sized speech segments, causing an overfitting effect when MGE-dynamics-AW training was performed with this model setting. From this viewpoint, MGE-dynamics-FW has the flexibility to work well irrespective of whether a three-state or five-state structure is used.

6. CONCLUSION

In this paper, we incorporate dynamic properties of speech parameters into the MGE criterion by defining the generation error of dynamic features and introducing this error component into the GEF, resulting in the so-called MGE-dynamics criterion. We also propose two methods for setting the weight associated with this newly added error component, which are fixed weighting (FW) and adaptive weighting (AW). An objective evaluation shows that MGE-dynamics-AW obtains comparable performance for the static feature and better performance for the delta feature compared with baseline MGE training. Subjective listening tests indicate that a small but statistically significant improvement in the quality of synthesized speech was perceived in the case of MGE-dynamics-AW training. The newly derived MGE-dynamics criterion improves the capability of HMMs in capturing dynamic properties of speech while maintaining a computational complexity similar to that of the baseline MGE criterion. Future work should target the investigation of the effect of the window length used in dynamic feature calculation on MGE-dynamics training and the effect of MGE-dynamics training on F_0 modeling.

ACKNOWLEDGEMENTS

This work is supported by a PhD Fellowship from the Ministry of Education and Training of Vietnam. We would like to thank two anonymous reviewers for their suggestions and comments, which improved the quality of our paper.

REFERENCES

[1] K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc. ICASSP*, pp. 660–663 (1995).

[2] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech synthesis using HMMs with dynamic features," *Proc. ICASSP*, pp. 389–392 (1996).

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Proc. Eurospeech*, pp. 2347–2350 (1999).

[4] H. Zen, K. Tokuda and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, **51**, 1039–1064 (2009).

[5] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust. Speech Signal. Process.*, **34**, 52–59 (1986).

[6] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. ICASSP*, pp. 1315–1318 (2000).

[7] Y.-J. Wu and R. H. Wang, "Minimum generation error training for HMM-based speech synthesis," *Proc. ICASSP*, pp. 889–892 (2006).

[8] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, **80**, 1016–1025 (1986).

[9] Y.-J. Wu, H. Zen, Y. Nankaku and K. Tokuda, "Minimum generation error criterion considering global/local variance for HMM-based speech synthesis," *Proc. ICASSP*, pp. 4621–4624 (2008).

[10] Y.-J. Wu and K. Tokuda, "Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis," *Proc. Interspeech*, pp. 577–580 (2008).

[11] M. Lei, Z.-H. Ling and L.-R. Dai, "Minimum generation error training with weighted Euclidean distance on LSP for HMM-based speech synthesis," *Proc. ICASSP*, pp. 4230–4233 (2010).

[12] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, **E90-D**, 816–824 (2007).

[13] S. Amari, "A theory of adaptive pattern classifiers," *IEEE Trans. Electron. Comput.*, **16**, 299–307 (1967).

[14] K. Elenius and M. Blomberg, "Effects of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system," *Proc. ICASSP*, pp. 535–538 (1982).

[15] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, **77**, 257–286 (1989).

[16] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Commun.*, **9**, 357–363 (1990).

[17] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. ICASSP*, pp. 137–140 (1992).

[18] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," *Proc. ICASSP*, pp. 229–232 (1999).

[19] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," *Proc. ICASSP*, pp. 93–96 (1983).

[20] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black and K. Tokuda, "The HMM-based speech synthesis system version 2.0," *Proc. ISCA SSW6*, pp. 294–299 (2007).

[21] Y.-J. Wu, H. Zen, Y. Nankaku and K. Tokuda, "Evaluation of parameter optimization methods for minimum generation error based HMM training," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 370–371 (2007).

[22] L. R. Ott and M. T. Longnecker, *An Introduction to Statistical Methods and Data Analysis* (Brooks/Cole, California, 2010), pp. 314–319.

[23] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, "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.*, **27**, 187–207 (1999).

Duy Khanh Ninh received the B.S. degree in electronics and telecommunication technology from Vietnam National University, Hanoi in 2003, and the M.S. degree in information processing and communications from Hanoi University of Science and Technology (HUST), Vietnam, in 2005. During 2003–2006, he was a researcher at International Research Institution on Multimedia Information, Communication and Applications, HUST. In 2006, he joined Danang University of Technology, Vietnam, as an Assistant Lecturer. He is currently a Ph.D. candidate at Ritsumeikan University, Japan. His research interests include speech synthesis, speech prosody, and statistical modeling. He is a student member of ASJ.

Masanori Morise received the Ph.D. degree in engineering from Wakayama University in 2008. He was a JSPS Research Fellow (DC1) in 2006–2008, and a postdoctoral researcher at Kwansei Gakuin University in 2008–2009. He is currently an Assistant Professor at Ritsumeikan University. His research interests include speech analysis/synthesis and speech perception.

Yoichi Yamashita received his B.E., M.E. and Dr.Eng. degrees from Osaka University in 1982, 1984 and 1993, respectively. He worked for the Institute of Scientific and Industrial Research of Osaka University as a Technical Official, a Research Associate, and an Assistant Professor from 1984 to 1997. In 1997, he joined Ritsumeikan University as an Associate Professor in the College of Science and Engineering. He is currently a Professor in the College of Information Science and Engineering. His research interests include speech understanding, speech synthesis, speech communication, and spoken document processing. He is a member of IEICE, IPSJ, JSAI, ISCA, and IEEE.