PAPER *Special Section on Recent Advances in Machine Learning for Spoken Language Processing*

A Statistical Sample-Based Approach to GMM-Based Voice Conversion Using Tied-Covariance Acoustic Models

Shinnosuke TAKAMICHI[†a)], *Nonmember*, **Tomoki TODA**[††b)], *Member*, **Graham NEUBIG**[†c)], *Nonmember*,
Sakriani SAKTI[†d)], *and* **Satoshi NAKAMURA**[†e)], *Members*

SUMMARY This paper presents a novel statistical sample-based approach for Gaussian Mixture Model (GMM)-based Voice Conversion (VC). Although GMM-based VC has the promising flexibility of model adaptation, quality in converted speech is significantly worse than that of natural speech. This paper addresses the problem of inaccurate modeling, which is one of the main reasons causing the quality degradation. Recently, we have proposed statistical sample-based speech synthesis using rich context models for high-quality and flexible Hidden Markov Model (HMM)-based Text-To-Speech (TTS) synthesis. This method makes it possible not only to produce high-quality speech by introducing ideas from unit selection synthesis, but also to preserve flexibility of the original HMM-based TTS. In this paper, we apply this idea to GMM-based VC. The rich context models are first trained for individual joint speech feature vectors, and then we gather them mixture by mixture to form a Rich context-GMM (R-GMM). In conversion, an iterative generation algorithm using R-GMMs is used to convert speech parameters, after initialization using over-trained probability distributions. Because the proposed method utilizes individual speech features, and its formulation is the same as that of conventional GMM-based VC, it makes it possible to produce high-quality speech while keeping flexibility of the original GMM-based VC. The experimental results demonstrate that the proposed method yields significant improvements in term of speech quality and speaker individuality in converted speech.

key words: *GMM-based voice conversion, sample-based speech synthesis, speech parameter conversion, rich context model*

1. Introduction

Statistical Voice Conversion (VC) is an effective technique for modifying speech parameters to convert non-/para-linguistic information while keeping linguistic information unchanged, making it possible to enhance various speech-based systems [1]–[4]. While a variety of methods exist for VC [5]–[7], Gaussian Mixture Model (GMM)-based VC [8], [9] is still popular thanks to its stability and flexibility such as model adaptation [10][*].

In GMM-based VC, a GMM is used to jointly model the source and target speech features in order to convert the speech parameters. One of the biggest issues in GMM-based VC is poor quality in converted speech. The reasons causing the quality degradation are classified into analysis-synthesis errors [14], the over-smoothing effect [15], [16], and inaccurate modeling, the latter of which we handle in this paper. For example, as previous work has noted [5], [9], within each mixture component, the conversion function is linear which is a poor match for VC.

Inaccurate modeling is a common issue among statistical approaches for speech synthesis. One promising approach that has been proposed for Text-To-Speech (TTS) is to introduce ideas of unit selection synthesis (sample-based speech synthesis) [17], [18] which has excellent speech quality compared to statistical approaches. For Hidden Markov Model (HMM)-based TTS [19], Maximum Likelihood (ML)-based unit selection [20] has been proposed to guide speech parameter segments to maximize HMM likelihoods. Although the use of individual speech segments significantly improves quality in synthetic speech, it loses the flexibility of the original HMM-based TTS. Recently, we have proposed a statistical sample-based approach using tied-covariance acoustic models called *rich context models* for TTS that is not only high-quality but also flexible [21], as shown in Fig. 1. In conventional HMM-based TTS, because parameters of an acoustic model are estimated using some speech segments, this approach loses information of individual speech segments. On the other hand, our previous work in TTS [21] has modeled the individual speech segments with rich context models [22]. We have built the mixture model using the rich context models, and synthetic speech parameters are generated from the selected rich context models of the mixture model. By reformulating the statistical models in the same form as in HMM-based TTS, this method makes it possible to preserve the original flexibility.

In this paper, we apply this idea to GMM-based VC reviewed in Sect. 2. In Sect. 3, we construct rich context models[**] corresponding to individual joint speech features, then, construct a mixture model called a *Rich context-GMM (R-GMM)* using the trained rich context models belonging to the same mixture component of the conventional GMM. An

Manuscript received February 5, 2016.
Manuscript revised May 25, 2016.
Manuscript publicized July 19, 2016.

[†] The authors are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
[††] The author is with Information Technology Center, Nagoya University, Nagoya-shi, 464–8601 Japan.
a) E-mail: shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
b) E-mail: tomoki@icts.nagoya-u.ac.jp
c) E-mail: neubig@is.naist.jp
d) E-mail: ssakti@is.naist.jp
e) E-mail: s-nakamura@is.naist.jp

DOI: 10.1587/transinf.2016SLP0020

[*] Also, GMM-based frameworks can be combined with deep neural nets-based speech synthesis [11]–[13].

[**] The defined name “rich context model” is not strictly accurate for GMM-based VC because no context labels are used. However, we use this name to maintain consistency with our previous work [21].

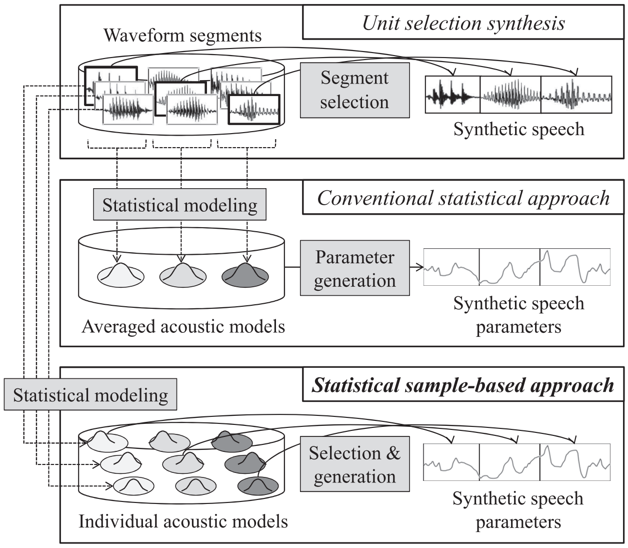

Fig. 1 Comparison of approaches, unit selection synthesis, conventional statistical approaches (e.g., HMM-based TTS and GMM-based VC), and the proposed statistical sample-based approach. Whereas an acoustic model corresponds to some speech segments in the conventional statistical approaches, it corresponds to just one speech segment in the statistical sample-based approach. Note that the individual acoustic models are calculated using individual speech segments, but their covariance matrices are the same to those of averaged acoustic models.

iterative algorithm is used to convert speech parameters using the rich context models selected from the R-GMMs. For initialization of the iteration, we further build over-trained acoustic models to generate a less-averaged initial parameter sequence. Discontinuous transitions are observed in the initial parameters, but it can be alleviated by the iterative parameter conversion. The experimental results in Sect. 4 demonstrate that the proposed method yields significant improvements in speech quality and speaker individuality in converted speech.

2. Conventional GMM-Based Voice Conversion

2.1 Acoustic Modeling

A joint probability density function of speech parameters of the source and target speakers is modeled with a GMM using parallel data as follows:

$$P(\boldsymbol{Z}_t|\boldsymbol{\lambda}) = \sum_{q=1}^{Q} w_q^{(Z)} P(\boldsymbol{Z}_t|q,\boldsymbol{\lambda}), \tag{1}$$

$$P(\boldsymbol{Z}_t|q,\boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{Z}_t;\boldsymbol{\mu}_q^{(Z)},\boldsymbol{\Sigma}_q^{(Z)}\right), \tag{2}$$

$$\boldsymbol{\mu}_q^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_q^{(X)} \\ \boldsymbol{\mu}_q^{(Y)} \end{bmatrix}, \; \boldsymbol{\Sigma}_q^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_q^{(XX)} & \boldsymbol{\Sigma}_q^{(XY)} \\ \boldsymbol{\Sigma}_q^{(YX)} & \boldsymbol{\Sigma}_q^{(YY)} \end{bmatrix}, \tag{3}$$

where, $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ is the joint vector of the input spectral features $\boldsymbol{X}_t$ and the output spectral features $\boldsymbol{Y}_t$ at frame t, and $\boldsymbol{Y}_t$ is given by $2D$-dimensional joint static and dynamic feature vectors, $[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$, where $\boldsymbol{y}_t$ is represented as a D-dimensional static feature vector. The source feature vector $\boldsymbol{X}_t$ is also given by the same form in this paper. A GMM parameter set $\boldsymbol{\lambda}$ consists of the Q mixture components, and each component has the mixture weight $w_q^{(Z)}$, the mean vector $\boldsymbol{\mu}_q^{(Z)}$, and the covariance matrix $\boldsymbol{\Sigma}_q^{(Z)}$. $\boldsymbol{\mu}_q^{(Z)}$ consists of the input and output mean vectors, $\boldsymbol{\mu}_q^{(X)}$ and $\boldsymbol{\mu}_q^{(Y)}$. $\boldsymbol{\Sigma}_q^{(Z)}$ consists of the source and target covariance matrices, $\boldsymbol{\Sigma}_q^{(XX)}$ and $\boldsymbol{\Sigma}_q^{(YY)}$ and cross-covariance matrices, $\boldsymbol{\Sigma}_q^{(YX)}$ and $\boldsymbol{\Sigma}_q^{(XY)}$.

2.2 Speech Parameter Conversion

Given the T-frame source speech feature sequence $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$, we determine an optimal GMM mixture sequence $\hat{\boldsymbol{q}} = [\hat{q}_1, \cdots, \hat{q}_t, \cdots, \hat{q}_T]$ as follows:

$$\hat{q}_t = \operatorname*{argmax}_{q} P(q|\boldsymbol{X}_t, \boldsymbol{\lambda}), \tag{4}$$

where $\hat{q}_t$ is the optimal GMM mixture component at frame t. The speech parameters are converted by maximizing an objective function using a GMM likelihood as follows:

$$\begin{aligned} \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} &= \operatorname*{argmax}_{\boldsymbol{y}} P(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}) & (5) \\ &= \left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{E}_{\hat{\boldsymbol{q}}}, & (6) \end{aligned}$$

where $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ is the converted speech parameter sequence, and $\hat{\boldsymbol{y}}_t$ is the D-dimensional converted speech parameters at frame t. The components of the $2DT$-dimensional mean vector, $\boldsymbol{E}_{\hat{\boldsymbol{q}}} = \left[\boldsymbol{\mu}_{\hat{q}_1,1}^\top, \cdots, \boldsymbol{\mu}_{\hat{q}_t,t}^\top, \cdots, \boldsymbol{\mu}_{\hat{q}_T,T}^\top\right]^\top$, and $2DT$-by-$2DT$ covariance matrix, $\boldsymbol{D}_{\hat{\boldsymbol{q}}} = \mathrm{diag}_{2D}\left[\boldsymbol{\Sigma}_{\hat{q}_1}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_t}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_T}\right]$, are given as:

$$\boldsymbol{\mu}_{\hat{q},t} = \boldsymbol{A}_{\hat{q}} \boldsymbol{X}_t + \boldsymbol{b}_{\hat{q}}, \tag{7}$$

$$\boldsymbol{\Sigma}_{\hat{q}} = \boldsymbol{\Sigma}_{\hat{q}}^{(YY)} - \boldsymbol{A}_{\hat{q}} \boldsymbol{\Sigma}_{\hat{q}}^{(XX)} \boldsymbol{A}_{\hat{q}}^\top, \tag{8}$$

$$\boldsymbol{A}_{\hat{q}} = \boldsymbol{\Sigma}_{\hat{q}}^{(YX)} \boldsymbol{\Sigma}_{\hat{q}}^{(XX)-1}, \tag{9}$$

$$\boldsymbol{b}_{\hat{q}} = \boldsymbol{\mu}_{\hat{q}}^{(Y)} - \boldsymbol{A}_{\hat{q}} \boldsymbol{\mu}_{\hat{q}}^{(X)}, \tag{10}$$

where the notation diag_{2D} denotes the construction of a block diagonal matrix that has the $2D$-by-$2D$ diagonal elements. $\boldsymbol{W}$ is the weighting matrix to calculate the dynamic features [23]. As shown in Eq. (7), the conventional GMM performs linear conversion within one mixture component.

3. Statistical Sample-Based Voice Conversion Using Rich Context Models

3.1 Acoustic Modeling

After conventional training, rich context models are trained for individual joint speech features, $\boldsymbol{Z}_t$, by updating the mean vector of the GMM mixture components while tying its covariance matrix. The m-th rich context model of the q-th GMM mixture component is

$$P(\boldsymbol{Z}_t|q,m,\boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{Z}_t;\boldsymbol{\mu}_{q,m}^{(Z)},\boldsymbol{\Sigma}_q^{(Z)}\right), \tag{11}$$

$$\boldsymbol{\mu}^{(Z)}_{q,m} = \begin{bmatrix} \boldsymbol{\mu}^{(X)}_{q,m} \\ \boldsymbol{\mu}^{(Y)}_{q,m} \end{bmatrix}, \tag{12}$$

where the mean vector $\boldsymbol{\mu}^{(Z)}_{q,m}$ consists of the individual input and output mean vectors, $\boldsymbol{\mu}^{(X)}_{q,m}$ and $\boldsymbol{\mu}^{(Y)}_{q,m}$. The individual mean vectors are estimated based on the ML criterion, and each of them is equal to one joint feature vector. The mixture component that $\boldsymbol{Z}_t$ belongs to is determined as follows:

$$\hat{q}_t = \operatorname*{argmax}_{q} P\left(q|\boldsymbol{Z}_t, \boldsymbol{\lambda}\right). \tag{13}$$

In this paper, we perform discriminative GMM training [24] between conventional training and rich context model training in order to alleviate mismatch between Eq. (13) and Eq. (4)[†]. As described in Sect. 3.2, the rich context models are selected from the mixture component determined with Eq. (4) in speech parameter conversion. Therefore, we expect that the discriminative training is effective to select the better rich context models.

After training rich context models, the output probability density for each GMM mixture component is given as a R-GMM constructed with all rich context models belonging to the same mixture component. The R-GMM of the q-th mixture component is

$$P\left(\boldsymbol{Z}_t|q, \boldsymbol{\lambda}\right) = \sum_{m=1}^{M_q} w^{(Z)}_{q,m} P\left(\boldsymbol{Z}_t|q, m, \boldsymbol{\lambda}\right) \tag{14}$$

$$= \sum_{m=1}^{M_q} w^{(Z)}_{q,m} \mathcal{N}\left(\boldsymbol{Z}_t; \boldsymbol{\mu}^{(Z)}_{q,m}, \boldsymbol{\Sigma}^{(Z)}_{q}\right), \tag{15}$$

where $w^{(Z)}_{q,m}$ is the weight of the m-th rich context model of the q-th mixture component. M_q is the total number of the rich context models of the q-th mixture component, and is equal to the number of speech features belonging to the mixture component. We can calculate the ML estimate of $w^{(Z)}_{q,m}$ based on the occupancy counts[††] but we set it to an equivalent value, $w^{(Z)}_{q,m} = 1/M_q$ for each component, following [21]. These procedures are shown in Fig. 2.

3.2 Speech Parameter Conversion

3.2.1 Iterative Conversion

After determining $\hat{\boldsymbol{q}}$ in the standard manner with Eq. (4), we calculate the output probability density function $P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right)$ as follows:

$$P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right) = \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \boldsymbol{m}, \boldsymbol{X}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{m}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right), \tag{16}$$

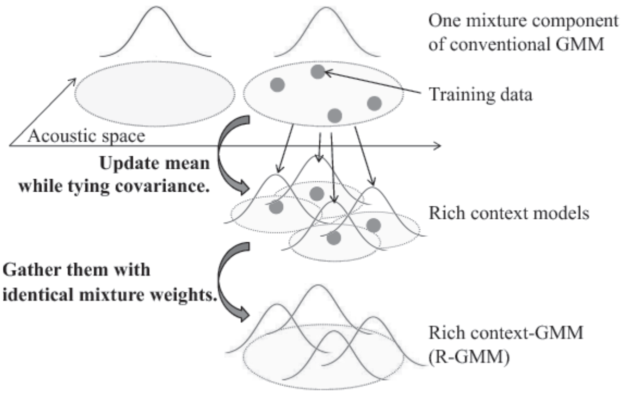

Fig. 2 Procedure to create Rich context-GMMs (R-GMMs) using the rich context models belonging to the same mixture component of the conventional GMM.

where $\boldsymbol{m} = [m_1, \cdots, m_t, \cdots, m_T]$ is a rich context model sequence, and m_t is the rich context model at frame t. $P\left(\boldsymbol{m}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right)$ is given as:

$$P\left(\boldsymbol{m}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right) = \prod_{t=1}^{T} P\left(m|\hat{q}_t, \boldsymbol{X}_t, \boldsymbol{\lambda}\right), \tag{17}$$

$$P\left(m|\hat{q}_t, \boldsymbol{X}_t, \boldsymbol{\lambda}\right) \equiv \frac{1}{M_{\hat{q}_t}}. \tag{18}$$

Traditionally, the posterior probability $P\left(m|\hat{q}_t, \boldsymbol{X}_t, \boldsymbol{\lambda}\right)$ is calculated as the similar as Eq. (4), but we set it constant among rich context models belonging to the same mixture component, following [21]. In practice, there are enormous numbers of candidates for rich context models in speech parameter conversion[†††]. Therefore, we calculate $P\left(m|\hat{q}_t, \boldsymbol{X}_t, \boldsymbol{\lambda}\right)$ in a similar fashion to Eq. (4), then, we set $P\left(m|\hat{q}_t, \boldsymbol{X}_t, \boldsymbol{\lambda}\right) = 1/M_{\hat{q}_t,t}$ for the rich context models having the $M_{\hat{q}_t,t}$-best posterior probability, and $P\left(m|\hat{q}_t, \boldsymbol{X}_t, \boldsymbol{\lambda}\right) = 0$ otherwise, where $M_{\hat{q}_t,t}$ $(1 \leq M_{\hat{q}_t,t} \leq M_{\hat{q}_t})$ is the number of candidates at frame t.

The output probability density function is approximated with the single rich context model sequence $\hat{\boldsymbol{m}}$ as follows:

$$P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right) \simeq P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \hat{\boldsymbol{m}}, \boldsymbol{X}, \boldsymbol{\lambda}\right) P\left(\hat{\boldsymbol{m}}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right). \tag{19}$$

After determining the initial speech parameter sequence $\boldsymbol{y}^{(0)}_{\hat{\boldsymbol{q}},\hat{\boldsymbol{m}}}$, the converted speech parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}},\hat{\boldsymbol{m}}}$ is determined by iteratively maximizing the likelihood as follows:

$$\hat{\boldsymbol{m}}^{(i+1)} = \operatorname*{argmax}_{\boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{W}\hat{\boldsymbol{y}}^{(i)}_{\hat{\boldsymbol{q}},\hat{\boldsymbol{m}}}, \hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right), \tag{20}$$

$$\hat{\boldsymbol{y}}^{(i+1)}_{\hat{\boldsymbol{q}},\hat{\boldsymbol{m}}} = \operatorname*{argmax}_{\boldsymbol{y}} P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \hat{\boldsymbol{m}}^{(i+1)}, \boldsymbol{X}, \boldsymbol{\lambda}\right), \tag{21}$$

where i is the iteration index. $P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \hat{\boldsymbol{m}}, \boldsymbol{X}, \boldsymbol{\lambda}\right)$ is given as:

[†] Whereas $P\left(q|\boldsymbol{Z}_t, \boldsymbol{\lambda}\right)$ is used in the training stage, $P\left(q|\boldsymbol{X}_t, \boldsymbol{\lambda}\right)$ is used in the conversion stage. The discriminative training algorithm [24] trains the GMM parameters to alleviate this inconsistency.

[††] There are basically no duplicated joint speech features, but such features are included in the training data because we employ Dynamic Time Warping (DTW) to make joint feature vectors.

[†††] Even if the training data size of the proposed method is the same to that of [21], the number of the candidates becomes bigger. This is because one rich context model corresponds to one speech feature vector in this study whereas it corresponds to one speech segment in [21].

$$P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}},\hat{\boldsymbol{m}},\boldsymbol{X},\lambda\right)=\mathcal{N}\left(\boldsymbol{W}\boldsymbol{y};\boldsymbol{E}_{\hat{\boldsymbol{q}},\hat{\boldsymbol{m}}},\boldsymbol{D}_{\hat{\boldsymbol{q}}}\right),\tag{22}$$

$$\boldsymbol{E}_{\hat{\boldsymbol{q}},\hat{\boldsymbol{m}}}=\left[\boldsymbol{\mu}_{\hat{q}_1,\hat{m}_1,1}^{\top},\cdots,\boldsymbol{\mu}_{\hat{q}_T,\hat{m}_T,T}^{\top}\right],\tag{23}$$

$$\boldsymbol{\mu}_{\hat{q},\hat{m},t}=\boldsymbol{A}_{\hat{q}}\boldsymbol{X}_t+\boldsymbol{b}_{\hat{q},\hat{m}},\tag{24}$$

$$\boldsymbol{b}_{\hat{q},\hat{m}}=\boldsymbol{\mu}_{\hat{q},\hat{m}}^{(Y)}-\boldsymbol{A}_{\hat{q}}\boldsymbol{\mu}_{\hat{q},\hat{m}}^{(X)}.\tag{25}$$

Comparing Eq. (7) and Eq. (24), while the bias component is constant in the conventional GMM, it varies depending on the selected rich context models in the proposed method.

3.2.2 Initialization

For initialization of the speech parameter conversion, one reasonable way is to use the parameter sequence generated from conventional GMMs. However, as noted in [21], [25], it is inappropriate to initialize using the conventional GMM. Also, we have reported in [21] that the speech parameter sequence generated from the over-trained acoustic models provides better initialization. The over-trained acoustic models efficiently avoid an averaging effect in the initial speech parameters. Though the over-trained models cause the discontinuous transitions in the initial parameters, the discontinuity can be alleviated by the iterative conversion process considering delta features. Also, we have reported that the number of the over-trained models can be determined to maximize a Global Variance (GV) likelihood [9] of the finally generated speech parameters.

In this paper, we train the over-trained acoustic models for each sub-region as shown in Fig. 3. The acoustic space is divided into Q sub-regions by Eq. (13) first, then an acoustic models are trained to fit the training data of the each sub-region. This over-trained acoustic model is given as a GMM for each sub-region, and is trained in the standard manner. The total number of the over-trained models is the sum of the number of mixture components of the GMMs. The Minimum Description Length (MDL) criterion [26] can be utilized to determine the number of over-trained models, but we determine it by the Linde-Buzo-Gray (LBG) algorithm [27]. After determining $\hat{\boldsymbol{q}}$, the over-trained models are selected as the similar as Eq. (4), and the initial parameter sequence is generated in the standard manner using the over-trained models.

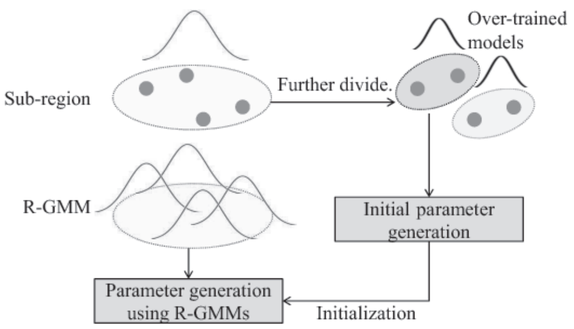

Fig. 3 Proposed initialization for iterative speech parameter conversion. The over-trained models are trained for each sub-region divided by Eq. (13).

3.3 Discussion

Because one rich context model corresponds to one joint feature vector, the proposed processes are related to sample-based voice conversion [28]. The target cost and concatenation cost of the sample-based approach are regarded as the likelihoods for the static and dynamic parameters [29], [30].

From the perspective of utilizing information of individual speech features, the conversion process of the proposed method is the similar to that of kernel-based speech synthesis [31], [32], and exemplar-based voice conversion [33], [34]. One of the comparable advantages is that the individual acoustic models can be re-selected by the iterative conversion process.

From the perspective of the conversion function, whereas the conventional GMM performs linear conversion within one mixture component, the proposed method can perform piece-wise linear conversion as shown in Fig. 4.

As described in Sect. 3.2.2, the iterative conversion process is done to refine the discontinuity of the initial speech parameters. Figure 5 shows the objective function given by Eq. (19) in each iteration. Because the value is almost converge at the 1st iteration, only a few iterations are required.

Compared with conventional GMM-based VC, the proposed framework increases the number of model parameters because we need (1) a conventional GMM for mixture selection, (2) over-trained models for initialization, and (3) rich context models for conversion. As shown in Table 1, the number of the mean vectors notably increases as increasing the training data size. Similarly, the computation cost in conversion increases by the proposed method as shown in Table 2, and is proportional to the training data size. On the other hand, compared with the conven-

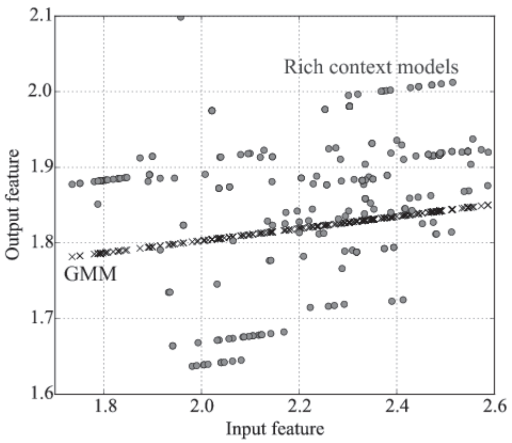

Fig. 4 An example of the conversion function within one GMM mixture component. Whereas a conversion function of the conventional GMM-based VC is given by a linear function $\boldsymbol{A}_q\boldsymbol{X}_t+\boldsymbol{b}_q$, that by the rich context models is given by a piece-wise linear function $\boldsymbol{A}_q\boldsymbol{X}_t+\boldsymbol{b}_{q,m}$, where a bias term varies depending on individual rich context models.

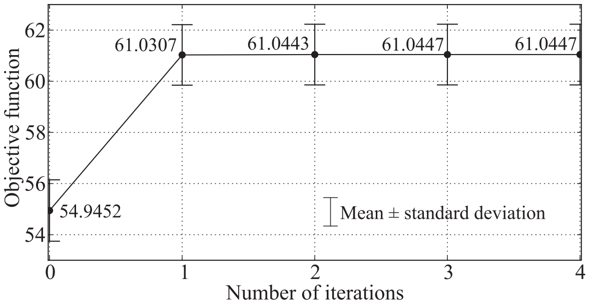

Fig. 5 Proposed objective functions in each iteration step. “0” of the x-axis indicates the objective function just after the initialization. We can see that the value is significantly increased at the 1st iteration, and is converged at the 4th iteration.

Table 1 The numbers of model parameters for the conventional GMM-based VC and proposed statistical sample-based VC. $M_q^{(\mathrm{init})}$ is the number of over-trained models belonging to the q-th GMM mixture component. In our evaluation in Sect. 4, we set Q, $\sum_{q=1}^{Q} M_q^{(\mathrm{init})}$, and $\sum_{q=1}^{Q} M_q$ to 128, 3616, and 590,745, respectively.

	Conventional	Proposed
Mixture weights	Q	$2Q + \sum_{q=1}^{Q} M_q^{(\mathrm{init})}$
Mean vectors	Q	$Q + \sum_{q=1}^{Q} \left(M_q^{(\mathrm{init})} + M_q\right)$
Covariance matrices	Q	$Q + \sum_{q=1}^{Q} M_q^{(\mathrm{init})}$

Table 2 The numbers of model combinations for the conventional GMM-based VC and proposed statistical sample-based VC. The computation cost during the model selection is proportional to these values, but that during the speech parameter conversion is the same between the conventional GMM-based VC, proposed initialization, and proposed conversion (for one iteration). For simplicity, the $M_{\hat{q}_t,t}$-best approximation method used in Sect. 3.2.1 is not applied here.

Conventional	Proposed		
	Mix. selection	Initialization	Conversion
QT	QT	$\sum_{t=1}^{T} M_{q_t}^{(\mathrm{init})}$	$\sum_{t=1}^{T} M_{q_t}$

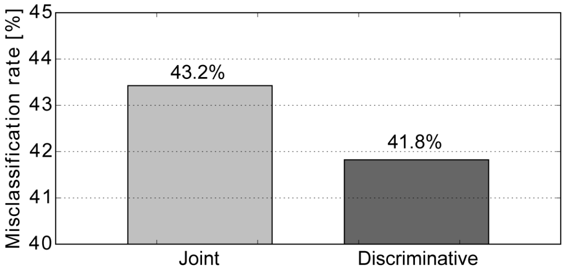

Fig. 6 Misclassification rate for the training data to confirm the effect of the discriminative GMM training.

tional GMM-based VC, the model complexity in the proposed method linearly increases according to an increase of the training data size as in unit selection synthesis. We expect that the speech quality in the proposed method tends to be significantly improved by increasing the training data size as in unit selection synthesis [35].

4. Experimental Evaluation

4.1 Experimental Condition

We selected the 450 parallel sentences of subsets A-through-I from the 503 phonetically balanced sentences included in the ATR Japanese speech database [36] for training, and the 53 sentences of subset J for evaluation. We trained female-to-male GMMs. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 0th-through-24th mel-cepstral coefficients were extracted as spectral parameters and log-scaled F_0 and 5-band aperiodicity [37], [38] were extracted as excitation parameters. The STRAIGHT analysis-synthesis system [39] was employed for parameter extraction and waveform generation. The feature vector consisted of spectral and excitation parameters and their delta and features. We built a 128-mixture GMM for spectral parameter conversion and a 16-mixture GMM for band-aperiodicity conversion. The proposed method was applied to spectral parameters. The log-scaled F_0 was linearly converted. The band-aperiodicity was converted using the conventional GMM. The total number of rich context models were 590, 745. In the parameter conversion, we selected the 128-best candidates for each frame. GV [9] and modulation spectra [40] were not considered in speech parameter conversion.

We compared the following approaches:

Conventional: conventional GMM-based VC[†]
Proposed: proposed approach using rich context models
Target: rich context models selected by reference data

In initialization for “Target,” the best rich context models were selected using target reference speech parameters. We first calculated misclassification rate for the training data to confirm the effect of the discriminative training [24]. Then, after determining the number of over-trained models, subjective evaluations were conducted to confirm effectiveness of the proposed method.

4.2 Effect of Discriminative Training

We evaluated the effect of the discriminative training done after the conventional joint density model training. The misclassification error rates were calculated for these training algorithms. The error rate was calculated as the number of the misclassified training data divided by the number of the training data. Here “misclassified data” indicates the joint speech feature that the mixture component determined with Eq. (13) is different from that determined with Eq. (4).

The error rates are shown in Fig. 6. Because we can see the 1.4% reduction of the error rates, it is expected that the discriminative training [41] makes it possible to select better rich context models.

[†] The discriminative training [24] was performed.

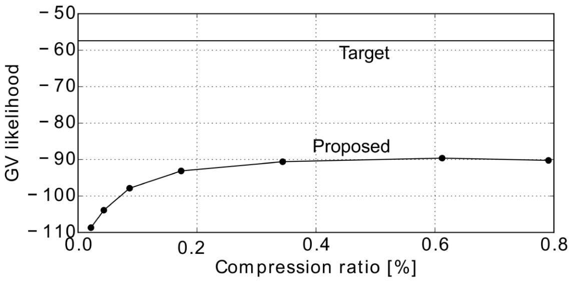

Fig. 7 GV likelihoods for the finally converted speech parameter sequence. A compression ratio (x-axis) is the number of over-trained models divided by that of the training data.

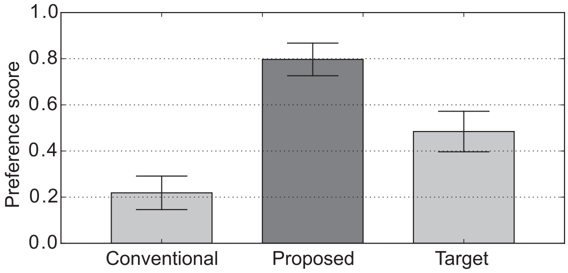

Fig. 8 Preference scores on speech quality in converted speech with 95% confidence intervals.

4.3 The Number of the Over-Trained Models

We calculated GV likelihoods for the finally converted speech parameters in order to determine the number of the over-trained models. In each sub-region, We increased the number with the LBG algorithm until we cannot estimate the model parameters. Although we can change the number sub-region by sub-region, the number was the same among the sub-regions[†].

The GV likelihood is shown in Fig. 7. We can find that the GV likelihood of "Proposed" is the biggest around the compression ratio of 0.6 (3616 over-trained models). Therefore, we determine the number of over-trained models to be 3616.

4.4 Evaluation in Speech Quality and Speaker Individuality

In the perceptual evaluation, a preference test (AB test) was conducted. We presented every pair of converted speech of 3 algorithms in random order, and we forced listeners to select the better-quality speech sample. Similarly, an XAB test on speaker individuality was conducted using the analysis-synthesized speech as a reference X.8 listeners participated in each evaluation.

The results of the preference tests on speech quality and speaker individuality are shown in Fig. 8 and Fig. 9, respectively. We can find that the proposed method achieves better scores in both speech quality and speaker individuality, compared to the conventional GMM-based VC. Therefore, we have demonstrated the effectiveness of the proposed method. The score of "Target" is lower than that of "Proposed" in speech quality. We found some speech samples of "Target" sounds discontinuous, and it is expected that small training data size caused this phenomenon. Whereas "Proposed" can alleviate the discontinuity by using slightly averaged initial speech parameters, "Target" uses non-averaged initial speech parameters[††]. We expect that this degradation using non-averaged parameters can be avoided by increasing the size of the training data[†††]. The alternative solution is to perform the iterative conversion after "Target" initialization, but this is not an aim of this study.

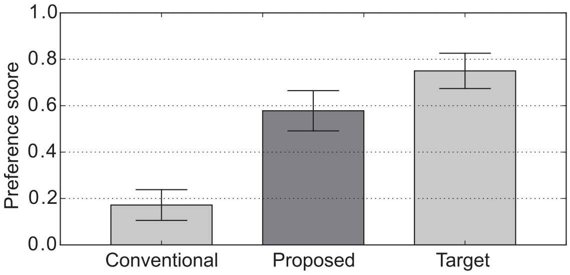

Fig. 9 Preference scores on speaker individuality in converted speech with 95% confidence intervals.

5. Conclusion

This paper has proposed a novel statistical sample-based approach using rich context models for Gaussian Mixture Model (GMM)-based voice conversion. Rich context models are trained for the individual joint speech features, and they are gathered to build a novel acoustic model called Rich context-GMM (R-GMM). After initialization using the over-trained acoustic models, speech parameters are iteratively converted using the selected rich context models. Experimental evaluation has demonstrated that the proposed method achieves better scores in both speech quality and speaker individuality. we will further integrate the GV and modulation spectra modeling techniques to the rich context models, and also investigate an adaptation method using rich context models.

Acknowledgements

Part of this work was supported by Grant-in-Aid for JSPS Fellows Grant Number 26·10354.

References

[1] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid

[†] Except we cannot estimate the GMM parameters of the sub-region.

[††] Our previous work [21] and this study use the same size of the training data, but the previous work [21] did not show such a result. We expect that this is because the rich context models used in [21] are temporally averaged, but those of this work are not.

[†††] The same solution is known in unit selection synthesis.

approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," IEICE Trans. Inf. & Syst., vol.E97-D, no.6, pp.1429–1437, June 2014.

[2] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, "Voice timbre control based on perceived age in singing voice conversion," IEICE Trans. Inf. & Syst., vol.E97-D, no.6, pp.1419–1428, June 2014.

[3] N. Hattori, T. Toda, H. Kawai, H. Saruwatari, and K. Shikano, "Speaker-adaptive speech synthesis based on Eigenvoice conversion and language-dependent prosodic conversion in speech-to-speech translation," in Proc. INTERSPEECH, pp.2769–2772, Florence, Italy, Aug. 2011.

[4] S. Aryal and R.G.-Osuna, "Can voice conversion be used to reduce non-native accents?," in Proc. ICASSP, pp.7929–7933, Florence, Italy, May 2014.

[5] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order Eigen space using deep belief nets," in Proc. INTERSPEECH, pp.369–372, Lyon, France, Aug. 2013.

[6] Z. Wu, T. Virtanen, T. Kinnunen, E.S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in Proc. SSW8, pp.201–206, Catalunya, Spain, Aug. 2013.

[7] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," IEEE Trans. Audio, Speech, Language Process., vol.20, no.3, pp.806–817, March 2012.

[8] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech and Audio Processing, vol.6, no.2, pp.131–142, March 1988.

[9] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio, Speech, Language Process., vol.15, no.8, pp.2222–2235, 2007.

[10] K. Ohta, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, "Adaptive voice-quality control based on one-to-many Eigenvoice conversion," in Proc. INTERSPEECH, pp.2158–2161, Chiba, Japan, Sept. 2010.

[11] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion," in Proc. INTERSPEECH, pp.3052–3056, Lyon, France, Sept. 2013.

[12] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in Proc. INTERSPEECH, pp.879–883, Dresden, Germany, Sept. 2015.

[13] E. Variani, E. McDermott, and G. Heigold, "A Gaussian mixture model layer jointly optimized with discriminative features within a deep neural network architecture," in Proc. ICASSP, pp.4270–4274, Brisbane, Australia, April 2015.

[14] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in Proc. INTERSPEECH, pp.2514–2518, Max Atria, Singapore, Sept. 2014.

[15] S. Takamichi, T. Toda, A.W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," IEEE Trans. Audio, Speech, Language Process., vol.24, no.4, pp.755–767, 2016.

[16] S. Takamichi, T. Toda, A.W. Black, and S. Nakamura, "Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion," in Proc. ICASSP, pp.4859–4863, Brisbane, Australia, April 2015.

[17] 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.

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

[19] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," Proc. IEEE, vol.101, no.5, pp.1234–1252, 2013.

[20] Z. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang, Y. Jiang, Z. Zhao, J. Yang, J. Chen, and G. Hu, "The USTC and iflytek speech synthesis systems for blizzard challenge 2007," in Proc. Blizzard Challenge Workshop, pp.1–6, Bonn, Germany, Aug. 2007.

[21] S. Takamichi, T. Toda, Y. Shiga, S. Sakti, G. Neubig, and S. Nakamura, "Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis," IEEE J. Sel. Topics Signal Process., vol.8, no.2, pp.239–250, 2014.

[22] Z. Yan, Q. Yao, and S.K. Frank, "Rich context modeling for high quality HMM-based TTS," in Proc. INTERSPEECH, pp.1755–1758, Brighton, U.K., Sept. 2009.

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

[24] H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "Alleviating the over-smoothing problem in GMM-based voice conversion with discriminative training," in Proc. INTERSPEECH, pp.3062–3066, Lyon, France, Sept. 2013.

[25] T. Merritt, J. Yamagishi, Z. Wu, O. Watts, and S. King, "Deep neural network context embeddings for model selection in rich-context HMM synthesis," in Proc. INTERSPEECH, pp.2207–2211, Dresden, Germany, Sept. 2015.

[26] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn (E), vol.28, no.3, pp.140–146, 2007.

[27] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol.28, pp.84–95, 1980.

[28] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in Proc. ICASSP, pp.513–516, Hawaii, U.S.A., April 2007.

[29] S. Kataoka, N. Mizutani, K. Tokuda, and T. Kitamura, "Decision tree backing-off in HMM-based speech synthesis," in Proc. INTERSPEECH, vol.2, pp.1205–1208, Jeju, Korea, Oct. 2004.

[30] Z. Ling and R. Wang, "HMM-based unit selection using frame sized speech segments," in Proc. INTERSPEECH, pp.2034–2037, Pittsburgh U.S.A., Sept. 2006.

[31] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical parametric speech synthesis based on Gaussian process regression," IEEE J. Sel. Topics Signal Process., vol.8, no.2, pp.173–183, April 2014.

[32] N.C.V. Pilkington, H. Zen, and M.J.F. Gales, "Gaussian process experts for voice conversion," in Proc. INTERSPEECH, pp.2761–2764, Florence, Italy, July 2011.

[33] R. Aihara, T. Takiguchi, and Y. Ariki, "Activity-mapping non-negative matrix factorization for exemplar-based voice conversion," in Proc. ICASSP, pp.4899–4903, Brisbane, Australia, April 2015.

[34] Z. Wu, E.S. Chng, and H. Li, "Exemplar-based voice conversion using joint nonnegative matrix factorization," Multimedia Tools and Applications, vol.74, no.22, pp.9943–9958, 2015.

[35] 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, 2006.

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

[37] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in MAVEBA 2001, pp.1–6, Firentze, Italy, Sept. 2001.

[38] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in Proc. INTERSPEECH, pp.2266–2269, Pittsburgh, U.S.A., Sept. 2006.

[39] H. Kawahara, I. Masuda-Katsuse, and A.D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun., vol.27, no.3–4, pp.187–207, 1999.

[40] S. Takamichi, T. Toda, A.W. Black, and S. Nakamura, "Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis," in Proc. ICASSP, pp.4210–4214, Brisbane, Australia, April 2015.

[41] H. Hwang, Y. Tsao, H. Wang, Y. Wang, and S. Chen, "Incorporating global variance in the training phase of GMM-based voice conversion," in Proc. APSIPA, pp.1–6, Kaohsiung, Taiwan, Oct. 2013.

Shinnosuke Takamichi received his B.E. from Nagaoka University of Technology, Japan, in 2011 and his M.E. and D.E. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan, in 2013 and 2016, respectively. He was a short-time researcher at the NICT, Kyoto, Japan in 2013, a visiting researcher of Carnegie Mellon University (CMU) in United States, from 2014 to 2015, and Research Fellow (DC2) of Japan Society for the Promotion of Science, from 2014 to 2016. He is currently a Project Research Associate of the University of Tokyo. He received the 7th Student Presentation Award from ASJ, the 35th Awaya Prize Young Researcher Award from ASJ, the 8th Outstanding Student Paper Award from IEEE Japan Chapter SPS, the Best Paper Award from APSIPA ASC 2014, the Student Paper Award from IEEE Kansai Section, the 30th TELECOM System Technology Award from TAF, the 2014 ISS Young Researcher's Award in Speech Field from the IEICE, the NAIST Best Student Award (Ph.D course), and the Best Student Award of Graduate School of Information Science (Ph.D course). His research interests include electroacoustics, signal processing, and speech synthesis. He is a student member of ASJ and IEEE SPS, and a member of ISCA.

Tomoki Toda earned his B.E. degree from Nagoya University, Aichi, Japan, in 1999 and his M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively. He is a Professor at the Information Technology Center, Nagoya University. He was a Research Fellow of JSPS from 2003 to 2005. He was then an Assistant Professor (2005–2011) and an Associate Professor (2011–2015) at the Graduate School of Information Science, NAIST. His research interests include statistical approaches to speech processing. He received more than 10 paper/achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).

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

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003–2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005–2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003–2007), A-STAR and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. From 2011, she has been an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation". She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Satoshi Nakamura is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994–2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007–2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009–2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampoli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.