SUMMARY We have developed a one-to-many eigenvoice conversion (EVC) system that allows us to convert a single source speaker’s voice into an arbitrary target speaker’s voice using an eigenvoice Gaussian mixture model (EV-GMM). This system is capable of effectively building a conversion model for an arbitrary target speaker by adapting the EV-GMM using only a small amount of speech data uttered by the target speaker in a text-independent manner. However, the conversion performance is still insufficient for the following reasons: 1) the excitation signal is not precisely modeled; 2) the oversmoothing of the converted spectrum causes muffled sounds in converted speech; and 3) the conversion model is affected by redundant acoustic variations among a lot of pre-stored target speakers used for building the EV-GMM. In order to address these problems, we apply the following promising techniques to one-to-many EVC: 1) mixed excitation; 2) a conversion algorithm considering global variance; and 3) adaptive training of the EV-GMM. The experimental results demonstrate that the conversion performance of one-to-many EVC is significantly improved by integrating all of these techniques into the one-to-many EVC system.

key words: speech synthesis, eigenvoice conversion, STRAIGHT mixed excitation, global variance, adaptive training

1. Introduction

Voice conversion (VC) [1] is a technique for converting a source speaker’s voice into another speaker’s voice without changing linguistic information. In the VC framework, a statistical method based on the Gaussian mixture model (GMM), which is employed as a conversion model, is used widely [2]. In this method, a GMM of the joint probability density of source and target acoustic features is trained using parallel data consisting of dozens of utterance-pairs of the source and target speakers. The trained GMM is capable of converting the source features into the target features based on minimum mean square error [2] or the maximum likelihood (ML) criterion [3]. Although this VC framework works reasonably well, its applications are limited due to lower flexibility of the training process, which requires a certain amount of parallel data.

In order to make the training process more flexible, several effective approaches have been proposed that use different speakers’ voices as prior knowledge for building the conversion model for different speakers, e.g., [4] and [5]. As one of these approaches, we have proposed eigenvoice conversion (EVC) [6]. In order to use more informative prior knowledge extracted from many other speakers, the eigenvoice technique [7] that was originally proposed as a speaker adaptation method in speech recognition has been applied to the GMM-based VC. We have developed new VC frameworks based on EVC for making VC applications more practical [8], e.g., one-to-many EVC. One-to-many EVC is a technique for converting a single source speaker’s voice into an arbitrary target speaker’s voice. Eigenvoice GMM (EV-GMM) is trained in advance with multiple parallel data sets consisting of the specified source speaker and many pre-stored target speakers. The trained EV-GMM is flexibly adapted to a new target speaker by estimating a small number of parameters (i.e., weights for eigenvectors) using a small amount of speech data of the target speaker in a text-independent manner. Moreover, the EV-GMM allows us to manually control the converted voice quality by manipulating a few of those parameters.

Our developed one-to-many EVC system outperforms the standard VC system in view of both converted speech quality and conversion accuracy for speaker individuality when using a small amount of speech data of the target speaker [8]. Toward the practical use of VC applications, it is essential to further improve its conversion performance. The conventional one-to-many EVC system often causes muffled and buzzy sounds in the converted speech. This is because the conventional system employs the following techniques:

- STRAIGHT simple excitation (STSE) based on switching a phase-manipulated pulse train and white noise [9], which is too simple to model the excitation signal appropriately
- The spectral conversion algorithm not considering global variance (GV) [3], which often causes oversmoothed spectral parameters
- The EV-GMM based on the target-speaker-independent GMM (TI-GMM) [6], which usually causes the conversion model to improperly capture acoustic variations among many pre-stored target speakers.

We call this VC framework “One-to-many VC” because a single adaptive conversion model is trained for representing mappings from a single source speaker’s voice into multiple target speakers’ voices.
Therefore, there remains room to improve the conventional one-to-many EVC system.

In this paper, to improve the one-to-many EVC system, we apply the following promising techniques to one-to-many EVC:

- The spectral conversion algorithm considering the GV [3]

It has been reported that each of these techniques causes significant improvements of the converted speech quality or the conversion accuracy for speaker individuality. Therefore, it is worthwhile to investigate the effectiveness of integrating all of these methods into the one-to-many EVC system. The results of experimental evaluations demonstrate that a combination of these techniques yields significant improvements in the converted speech quality and the conversion accuracy for speaker individuality.

This paper is organized as follows. In Sect. 2, we describe the conventional one-to-many EVC system. In Sect. 3, the proposed one-to-many EVC system is described. In Sect. 4, we describe experimental evaluations. Finally, we summarize this paper in Sect. 5.

2. Conventional One-to-Many EVC System

Figure 1 shows an overview of the conventional one-to-many EVC system that includes training, adaptation and conversion processes. In this system, we train the EV-GMM only for spectral features and employ the adapted EV-GMM for the spectral conversion.

2.1 Eigenvoice Gaussian Mixture Model (EV-GMM)

We employ 2D-dimensional acoustic features, \( \mathbf{X}_t = [\mathbf{x}_t, \Delta \mathbf{x}_t^\top] \) (the source speaker’s feature) and \( \mathbf{Y}_t^{(s)} = [\mathbf{y}_t^{(s)}, \Delta \mathbf{y}_t^{(s)}] \top \) (the \( s \)th target speaker’s feature), including \( D \)-dimensional static and dynamic features, where \( \tau \) denotes transposition of the vector. Joint probability density of time-aligned source and target features \( \mathbf{Z}_t^{(s)} = [\mathbf{X}_t^\top, \mathbf{Y}_t^{(s)}] \top \) which are determined by DTW is modeled with an EV-GMM as follows:

\[
P(\mathbf{Z}_t^{(s)}|\lambda^{(EV)}, \mathbf{w}^{(s)}) = \sum_{m=1}^{M} \alpha_m N\left( \mathbf{Z}_t^{(s)}, \mathbf{\mu}_m^{(Z)}(\mathbf{w}^{(s)}), \mathbf{\Sigma}_m^{(Z)} \right),
\]

\[
\alpha_m = \frac{1}{\mathbf{b}_m^\top \mathbf{b}_m + \mathbf{b}_m^0 \mathbf{b}_m^0},
\]

\[
\mathbf{\mu}_m^{(Z)}(\mathbf{w}^{(s)}) = \left[ \mathbf{B}_m \mathbf{w}^{(s)} + \mathbf{b}_m^0 \right],
\]

\[
\mathbf{\Sigma}_m^{(Z)} = \mathbf{B}_m \mathbf{\Sigma}_m \mathbf{B}_m^\top + \left( \mathbf{b}_m^0 \mathbf{b}_m^\top \right),
\]

where \( \alpha_m \) is the \( m \)th mixture component weight and \( N(\cdot; \mu, \Sigma) \) denotes Gaussian distribution with mean vector \( \mu \) and diagonal covariance matrix \( \Sigma \). A target mean vector of the EV-GMM is modeled by linear combination with the bias vector \( \mathbf{b}_m^0 \), representative vectors \( \mathbf{B}_m = [\mathbf{b}_m^{(1)}, \mathbf{b}_m^{(2)}, \ldots, \mathbf{b}_m^{(D)}] \) and the weight vector \( \mathbf{w}^{(s)} \). Voice characteristics of various target speakers are effectively modeled by setting \( \mathbf{w}^{(s)} \) to appropriate values. The other parameters \( \lambda^{(EV)} \) such as mixture component weights, source mean vectors, bias vectors, representative vectors and covariance matrices are tied for every target speaker.

2.2 Training Process

First, a TI-GMM is trained with multiple parallel data sets consisting of utterance-pairs of the source speaker and multiple pre-stored target speakers. Then, using only the parallel data set for source and the \( s \)th pre-stored target speakers, the \( s \)th target-speaker-dependent GMM (TD-GMM) is trained only by updating the target mean vectors of the TI-GMM. After training the TD-GMMs for all pre-stored target speakers, a 2DM-dimensional super-vector \( \mathbf{SV}^{(s)} = [\mathbf{\mu}^{(1)}(s), \mathbf{\mu}^{(2)}(s), \ldots, \mathbf{\mu}^{(M)}(s)] \top \) is constructed by concatenating the updated target mean vectors \( [\mathbf{\mu}^{(1)}(s), \mathbf{\mu}^{(2)}(s), \ldots, \mathbf{\mu}^{(M)}(s)] \) of the TD-GMM for each pre-stored target speaker. Finally, bias vector \( \mathbf{b}_m^0 \) and representative vectors \( \mathbf{B}_m \) are extracted by performing principal component analysis (PCA) for the super vectors for all pre-stored target speakers \( \{\mathbf{SV}^{(1)}, \mathbf{SV}^{(2)}, \ldots, \mathbf{SV}^{(S)}\} \), where \( S \) denotes the number of pre-stored target speakers.
2.3 Adaptation Process

We adapt the EV-GMM to an arbitrary target speaker by estimating the optimum weight vector for given speech samples of the target speaker without any linguistic information. We apply maximum likelihood eigen-decomposition (MLED) [7] to the weight vector estimation of the EV-GMM. The weight vector $\mathbf{w}$ is estimated so that a likelihood of the marginal distribution for a time sequence of the given target features $\{Y_1^{(tar)}, Y_2^{(tar)}, \ldots, Y_T^{(tar)}\}$ is maximized [6] as follows:

\[
\hat{\mathbf{w}} = \arg \max_{\mathbf{w}} \prod_{t=1}^{T} P \left( \mathbf{X}_t, Y_t^{(tar)} \mid \lambda^{(EV)}, \mathbf{w} \right) d\mathbf{X}_t
\]

\[
= \arg \max_{\mathbf{w}} \prod_{t=1}^{T} P \left( Y_t^{(tar)} \mid \lambda^{(EV)}, \mathbf{w} \right).
\]

Note that this process is completely unsupervised adaptation using only arbitrary utterances of the target speaker. Moreover, we need to use only a small amount of adaptation data because there are few parameters to be adapted.

2.4 Conversion Process

In the conversion process for spectral features, we use the conversion method based on maximum likelihood estimation (MLE) considering dynamic features [3]. Let a time sequence of the source features and that of the target features be $\mathbf{X} = [X_1^T, X_2^T, \ldots, X_T^T]$ and $\mathbf{Y} = [Y_1^T, Y_2^T, \ldots, Y_T^T]^T$, respectively. Converted static feature vectors $\hat{\mathbf{y}} = [\hat{y}_1^T, \hat{y}_2^T, \ldots, \hat{y}_T^T]^T$ are obtained as follows:

\[
\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} \sum_{m} P \left( \mathbf{m} \mid \mathbf{X}, \lambda^{(EV)} \right) P \left( \mathbf{Y} \mid \mathbf{X}, \mathbf{m}, \lambda^{(EV)}, \hat{\mathbf{w}} \right),
\]

subject to $\mathbf{Y} = \mathbf{W}\mathbf{y}$,

where $\mathbf{W}$ denotes the matrix to extend the static feature sequence to the static and dynamic feature sequence, and $P \left( \mathbf{Y} \mid \mathbf{X}, \mathbf{m}, \lambda^{(EV)}, \hat{\mathbf{w}} \right)$ is the $m$th conditional probability density of $\mathbf{Y}$ given $\mathbf{X}$ derived from the adapted EV-GMM. In this paper, we approximate the likelihood function in Eq. (5) with the suboptimum mixture component sequence as described in [3]. Note that this approximation doesn’t cause significant degradation in the conversion performance.

In the conversion process for fundamental frequency, we convert source fundamental frequency $F_0$ to target one as follows:

\[
\log \tilde{F}_0 = \frac{\sigma^{(s)}}{\sigma^{(s)}} \left( \log F_0 - \mu^{(s)} \right) + \mu^{(s)},
\]

where $\mu^{(s)}$ and $\sigma^{(s)}$ denote mean and standard deviation of log-scaled source $F_0$, and $\mu^{(s)}$ and $\sigma^{(s)}$ denote those of log-scaled target $F_0$. In this system, these statistics for the source speaker are calculated from the training data and those for the target speaker are calculated from the adaptation data.

In synthesizing converted speech, the excitation signal is generated with STSE. Figure 2 shows the generation process of STSE. In order to alleviate buzz sounds caused by using a pulse train for generating a voiced excitation signal, phase components in high-frequency bands (e.g., over 3 kHz) are dispersed with an all-pass filter [14]. A one-pitch waveform is generated by selecting the phase-manipulated pulse train based on the converted $F_0$ for voiced segments or white noise for unvoiced segments. Then, an excitation signal is generated by PSOLA (Pitch Synchronous OverLap Add) technique [15]. The converted speech is synthesized by filtering the generated excitation with the converted spectral sequence.

3. Improvement of One-to-Many EVC System

In order to improve the conversion performance of the one-to-many EVC system, we apply the excitation conversion based on STME, MLE-based spectral conversion considering GV, and the adaptive training for the EV-GMM to one-to-many EVC.

3.1 Excitation Conversion with STRAIGHT Mixed Excitation (STME)

Figure 3 shows the generation process of STME. For voiced segments, we generate both the phase-manipulated pulse train based on the converted $F_0$ and white noise. Then, a one-pitch waveform is generated by the frequency-dependent weighted sum of the phase-manipulated pulse train and white noise. Note that the frequency-dependent weight value varies frame by frame. This time-varying weight is determined based on an aperiodic component, which represents the degree of the noise component (lower value: periodic and upper value: noisy) in each frequency bin [11], at each time frame. In order to reduce the dimensionality of the parameter to be statistically modeled, the aperiodic components are averaged on five frequency sub-bands, i.e., 0–1, 1–2, 2–4, 4–6, and 6–8 kHz, in the same manner as described in [10]. For unvoiced segments, we generate white noise. Finally, an excitation signal is generated by PSOLA technique [15]. Figure 4 shows an example of a residual signal, STSE, STME and each of the excitation parameters, i.e., an $F_0$ contour and aperiodic component se-
sequences in individual frequency sub-bands. STSE is quite different from the residual signal especially at voiced segments with less periodicity, e.g., around 1.7–2.0 [sec]. Because a voiced excitation signal is generated using only the phase-manipulated pulse train, the strength of periodicity depends on the pre-defined all-pass filter for phase dispersion and it doesn’t vary frame by frame. On the other hand, STME is more similar to the residual signal than STSE because STME is capable of modeling the strength of periodicity at voiced frames. We can observe that aperiodic components in lower frequency sub-bands are inversely correlated with the strength of periodicity of the residual signal. These properties of aperiodic components are very helpful for generating an excitation signal with more characteristics similar to the residual signal.

We have proposed the conversion of aperiodic components based on a GMM in order to apply STME to VC and have demonstrated its effectiveness in the conventional VC framework [10]. In this paper, the joint probability density of aperiodic components between the source and target speakers is modeled by an EV-GMM rather than a GMM for applying STME to the one-to-many EVC system.

### 3.2 MLE-Based Conversion Considering GV

The GV is calculated as variances at individual dimensions of the feature vectors over a time sequence as follows:

\[
v_y = \left[ v_y(1), v_y(2), \ldots, v_y(D) \right]^T, \tag{8}
\]

\[
v_y(d) = \frac{1}{T} \sum_{t=1}^{T} \left( y_t(d) - \frac{1}{T} \sum_{r=1}^{T} y_r(d) \right)^2, \tag{9}
\]

where \( y_t(d) \) is the \( d \)-th dimensional element of the target vector at the \( t \)-th frame. In this paper, the GV is calculated utterance by utterance.

Figure 5 shows a time sequence of the 7th mel-cepstral coefficient extracted from the target speech and that of the converted coefficient by the conventional EVC system, respectively. We can observe that the GV of the converted mel-cepstral sequence is smaller than that of the target one.

![Fig. 5](image_url)
This is because the oversmoothing is caused through a statistical modeling process. It has been reported that both the converted speech quality and conversion accuracy for speaker individuality are dramatically improved by considering the GV of the converted parameters in the conversion process [3].

In the proposed system, the probability density of the GV is modeled by an eigenvoice-based single Gaussian distribution (EV-SG) as follows:

$$P(v(s)|\lambda^{(EV)}, w(s)) = \mathcal{N}(v(s); \mu(w(s)), \Sigma(v)),$$

$$\mu(w(s)) = B_v w(s) + b_v(0),$$

where $v(s)$ denotes the GV vector of the $s$th target speaker. A tied-parameter set $\lambda_v^{(EV)}$ includes representative vectors $B_v$, a bias vector $b_v(0)$ and a covariance matrix $\Sigma(v)$. The weight vector of the $s$th target speaker is $w(s)$.

In the spectral conversion process, the converted spectral features are determined by maximizing the objective function considering both the likelihood for the converted spectral features and that for those GV with respect to $y$ as follows:

$$\hat{y} = \arg \max_y \sum_{all \ m} P(m|X, \lambda^{(EV)}) P(Y|X, m, \lambda^{(EV)}, \hat{w}) \times P(v_s|\lambda^{(EV)}, \hat{w}),$$

where $\omega$ is a weight parameter for controlling the balance between those two likelihoods, and it is set to the ratio of the number of dimensions between spectral features and the GV vector, i.e., $2T$, in this paper. Note that we again approximate the objective function with the suboptimum mixture component sequence in this paper as in the conventional one-to-many EVC system.

### 3.3 Adaptive Training for EVC

In order to improve the adaptation performance of the EV-GMM, we implement an adaptive training method [12] for the EV-GMM for the proposed system.

The canonical EV-GMM is trained by maximizing the total likelihood of the adapted EV-GMMs for individual pre-stored target speakers with respect to both the tied-parameters of the EV-GMM $\lambda^{(EV)}$ and the weight vector $w(s)$ for each pre-stored target speaker as follows:

$$\{\lambda^{(EV)}, \hat{w}_s\} = \arg \max_{\lambda^{(EV)}, w_s} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(Z^{(t)}|\lambda^{(EV)}, w(s)),$$

where $w_s$ is a set of weight vectors for individual pre-stored target speakers $\{w^{(1)}, w^{(2)}, \ldots, w^{(S)}\}$. The EM algorithm is used for updating those parameters. It is difficult to update all parameters simultaneously because some of them depend on each other. Therefore, we first update weight vectors $\hat{w}_s$ for all of the pre-stored target speakers while fixing the other parameters. After that, while fixing these updated weight vectors, we update some tied-parameters related to mean vectors of the EV-GMM (i.e., $\mu_m, B_m$ and $\Sigma_m^{(EV)}$), and then, we update the other tied-parameters (i.e., $\sigma_m$ and $\Sigma_m^{(EV)}$). In each M-step, these parameter update procedures are iteratively performed for improving parameter estimation accuracy.

In the proposed system, we also apply the adaptive training to the EV-SG for the GV. The canonical EV-SG parameters are estimated by maximizing the total likelihood of the adapted EV-SGs for individual pre-stored target speakers’ GVs as follows:

$$\{\lambda^{(EV)}, \hat{w}_v\} = \arg \max_{\lambda_v^{(EV)}, w_v} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(v_s^{(t)}|\lambda^{(EV)}, w_v^{(s)}),$$

where $v^{(s)}$ denotes the GV vector extracted from the $n$th utterance of the $s$th pre-stored target speaker and $w_v^{(s)}$ is a set of weight vectors for EV-SG. We don’t have to perform EM algorithm because there is no hidden variable in the EV-SG. However, it is still difficult to update all parameters simultaneously for the same reason as in the EV-GMM. Therefore, individual parameters of the EV-SG are updated iteratively in the same manner mentioned above.

![Fig. 6 Overview of proposed one-to-many EVC system.](image-url)
3.4 Overview of Proposed One-to-Many EVC System

Figure 6 shows an overview of the proposed one-to-many EVC system.

In the training process, spectral features, aperiodic components, and GV vectors are extracted from speech samples of multiple parallel data sets, and joint feature vectors are constructed for spectral features and for aperiodic components. We build the EV-GMM for spectral features, the EV-GMM for aperiodic components, and the EV-SG for the GV using the PCA-based training process described in Sect. 2.2. Then, these models are independently optimized with the adaptive training. We also calculate mean and variance values of log-scaled $F_0$ of the source speaker.

In the adaptation process, spectral features, aperiodic components, GV vectors, and $F_0$ values are extracted from adaptation data of a new target speaker. Then, the weight vectors for the EV-GMMs for the spectral features and for the aperiodic components are independently estimated as shown in Eq. (4). The weight vector for the EV-SG is also estimated by maximizing the likelihood of the EV-SG in Eq. (10) for given the GV vectors. Mean and variance values of log-scaled $F_0$ of the target speaker are also calculated.

In the conversion process, spectral features are converted by MLE-based conversion considering the GV. On the other hand, aperiodic components are converted by the conventional MLE-based conversion without the GV because quality improvements yielded by considering the GV in the aperiodic conversion are not significant. The converted $F_0$ values are determined in Eq. (7). Finally, the excitation signal is generated based on STME with the converted $F_0$ values and the converted aperiodic components, and then synthetic speech is generated by filtering the excitation signal with the converted spectral features.

4. Experimental Evaluations

4.1 Experimental Conditions

We objectively and subjectively compared the performance of the proposed one-to-many EVC system with that of the conventional one. For the training process, we prepared parallel data sets of a single source male speaker and 160 pre-stored target speakers. These pre-stored target speakers were included in the Japanese Newspaper Article Sentences (JNAS) database [16] as shown in Table 1. The source male speaker was not included in JNAS and uttered all of the seven sub-sets and an additional sub-set consisting of 53 sentences used for evaluation.

In the evaluation, we used 10 target speakers that were composed of five male and five female speakers not included in the pre-stored target speakers. They also uttered the 53 evaluation sentences. We used 1 to 32 utterances for adapting the EV-GMM, and 21 utterances for evaluation. In the first E-step to estimate the weight vector for the EV-GMM adaptation, we used the TI-GMM as an initial model for each of the spectral features and aperiodic components.

We used 24-dimensional mel-cepstral coefficients as a spectral feature, which were extracted from smoothed spectrum analyzed by STRAIGHT [9], and aperiodic components that were averaged on five frequency bands described in Sect. 3.1. The number of representative vectors was 159 for mel-cepstrum, 64 for aperiodic components and 4 for GV, respectively. The number of mixture components was 128 for spectral features and 64 for the aperiodic features. These parameters were optimized so that the best conversion accuracy for each feature was obtained in the evaluation data.

4.2 Objective Evaluations

We evaluated the effectiveness of the proposed adaptive training method in the adaptation of the EV-GMM for aperiodic components and that of the EV-SG for the GV. Note that the effectiveness of the proposed adaptive training method in the spectral conversion has already been reported in [12]. As evaluation measures, we used RMSE on aperiodic components and the likelihood of the adapted EV-SG for GV in the evaluation data. Before the conversion, RMSE on aperiodic components between the source and target speakers was 2.70 [dB] and the log-scaled likelihood of EV-SG was 80.64. When we used STSE for synthesizing the converted speech, RMSE on aperiodic components between the converted speech and the target speech was 3.05 [dB].

Figures 7 and 8 show RMSE on aperiodic components and the log-scaled likelihood of the EV-SG for GV, respectively. We can see that the adaptation performance of both the EV-GMM for aperiodic components and the EV-SG for the GV is slightly improved by applying the proposed adaptive training. Moreover, these improvements are always ob-

<table>
<thead>
<tr>
<th>Sub-sets</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of male speakers</td>
<td>15</td>
<td>11</td>
<td>15</td>
<td>13</td>
<td>15</td>
<td>11</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>Number of female speakers</td>
<td>15</td>
<td>11</td>
<td>15</td>
<td>13</td>
<td>12</td>
<td>0</td>
<td>14</td>
<td>80</td>
</tr>
</tbody>
</table>

Fig. 7 Result of objective evaluation by RMSE on aperiodic components.
observed even if varying the amount of adaptation data. We have observed the same tendencies in the results of within-gender conversion and cross-gender conversion. These results demonstrate the effectiveness of applying the proposed adaptive training to modeling of the aperiodic components and the GV.

4.3 Subjective Evaluations

We conducted a preference test and an opinion test on speech quality and an XAB test on conversion accuracy for speaker individuality. To demonstrate the effectiveness of a combination of STME, GV and adaptive training,1 we evaluated several types of converted speech shown in Table 2. In the preference test, a pair of two different types of the converted speech was presented to listeners, and then they were asked which voice sounded better. In the opinion test, each listener evaluated speech quality of the converted voices using a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). In the XAB test, a pair of two different types of the converted speech was presented to them after presenting the target speech as a reference. Then, they were asked which voice sounded more similar to the reference. Each listener evaluated every pair-combination of all types of the converted speech.

Figure 9 shows the result of the preference test on speech quality. We can see the same tendency as observed in the preference test shown in Fig. 8. Significant improvements in the converted speech quality are yielded by applying each of the adaptive training, STME, and the conversion with the GV to the conventional system, and the best quality is yielded by the type viii system.

Figure 10 shows the result of the opinion test on speech quality. We can see the same tendency as observed in the preference test shown in Fig. 9. Significant improvements in the converted speech quality are yielded by applying each of the adaptive training, STME, and the conversion with the GV to the conventional system, and the best quality is yielded by the type viii system.

Figure 11 shows the result of the XAB test of conversion accuracy for speaker individuality. The adaptive training (type ii) does not contribute to the improvement of conversion accuracy for speaker individuality. This result is consistent with the result reported in [12]. On the other hand, STME (type iii) or the conversion with the GV (type iv) yields significant improvements of the conversion accuracy, and STME is the most effective. As observed in the result of speech quality, further improvements are yielded by combining these effective methods.

1It has been reported that a combination of STME and GV yields significant improvements in naturalness of synthetic speech in HMM-based speech synthesis [18].
These results suggest that the proposed system yields dramatic improvements in the performance of the one-to-many EVC system. Note that the same improvements have also been observed in the results of within-gender conversion and cross-gender conversion in each test.

5. Conclusion

In order to improve converted speech quality and conversion accuracy for speaker individuality of the one-to-many eigenvoice conversion (EVC) system, we have applied three promising techniques, i.e., STRAIGHT mixed excitation, the conversion algorithm considering global variance (GV) and the adaptive training method of the eigenvoice Gaussian mixture model (EV-GMM) to the conventional system. In the proposed one-to-many EVC system, we train two EV-GMMs for spectral features and for aperiodic components and an eigenvoice single Gaussian distribution for the GV separately. These models are effectively adapted to a new target speaker using a very small amount of adaptation data in a completely text-independent manner. Results of objective and subjective evaluations have demonstrated that the proposed one-to-many EVC system considerably outperforms the conventional one in view of both converted speech quality and conversion accuracy for speaker individuality.

Acknowledgment

This work was supported in part by MIC SCOPE and the authors would like to thank Prof. H. Kawahara of Wakayama University, Japan, for permission to use STRAIGHT.

References

Yamato Ohtani was born in Osaka, Japan on September 8, 1980. He received the B.E. degree in engineering from Osaka University, Osaka, Japan, in 2005 and the M.E. degree in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan, in 2007. He is now Ph.D. student in NAIST. He was an intern researcher at ATR Spoken Language Communication Research Laboratories (ATR-SLC) from 2006 to 2009. He is a student member of ISCA and ASJ.

Tomoki Toda received the B.E. degree from Nagoya University, Aichi, Japan, in 1999 and the M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively. He was a Research Fellow of the JSPS in the Graduate School of Engineering, NITECH, Aichi, Japan, from 2003 to 2005. He is currently an Assistant Professor of the Graduate School of Information Science, NAIST. He also has been a Visiting Researcher at the NICT, Kyoto, Japan, since May 2006. From March 2001 to March 2003, he was an Intern Researcher at the ATR SLC, Kyoto, Japan, and then he was a Visiting Researcher at the ATR SLC until March 2006. He was also a Visiting Researcher at the LTI, CMU, Pittsburgh, USA, from October 2003 to September 2004 and at the CUED, Cambridge, UK, from March 2008 to August 2008, respectively. His research interests include statistical approaches to speech processing such as voice transformation, speech synthesis, and speech analysis. He received the 18th TELECOM System Technology Award for Students and the 23rd TELECOM System Technology Award from the TAF, the 2007 ISS Best Paper Award from the IEICE, the 10th Ericsson Young Scientist Award from Nippon Ericsson K.K., the 4th Itakura Prize Innovative Young Researcher Award and the 26th Awaya Prize Young Researcher Award from the ASJ, and the 2009 Young Author Best Paper Award from the IEEE SPS. He is a member of IEEE, ISCA, IPSJ, and ASJ.

Hiroshi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E. and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991, 1993 and 2000, respectively. He joined Intelligent Systems Laboratory, SECOM CO., LTD., Mitaka, Tokyo, Japan, in 1993, where he engaged in the research and development on the ultrasonic array system for the acoustic imaging. He is currently an associate professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include array signal processing, blind source separation, and sound field reproduction. He received the Paper Awards from IEICE in 2001 and 2006, and from TAF in 2003. He is a member of the IEEE and the Acoustical Society of Japan.

Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a professor of Nara Institute of Science and Technology (NAIST), where he is directing speech and acoustics laboratory. From 1972 to 1993, he had been working at NTT Laboratories. During 1986-1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories. During 1984-1986, he was a visiting scientist in Carnegie Mellon University. He received the Yonezawa Prize from IEICE in 1975, the Signal Processing Society 1990 Senior Award from IEEE in 1991, the Technical Development Award from ASJ in 1994, IPSJ Yamashita SIG Research Award in 2000, and Paper Award from the Virtual Reality Society of Japan in 2001, IEICE paper award in 2005 and 2006, and Inose award in 2005. He is a fellow of the Institute of Electrical and Electronics Engineers (IEEE), and Information Processing Society of Japan, and a member of the Acoustical Society of Japan (ASJ) and International Speech Communication Society (ISCA).