1. Introduction
Voice conversion (VC) [1] is a technique that allows us to convert voice characteristics of a source speaker into those of a target speaker without changing linguistic information. As one of the statistical approaches to VC, a conversion method based on the Gaussian mixture model (GMM) has been proposed [2]. In this method, a GMM of joint probability density of source and target acoustic features is previously trained with a parallel data set consisting of utterance pairs of source and target voices. The trained GMM allows the conversion from the source into the target based on minimum mean square error [2] or maximum likelihood criterion [3]. Although this method works reasonably well, this training framework is less flexible and causes many limitations of VC applications.

There have been proposed several approaches of effectively using voices of other speakers as prior knowledge for flexibly developing the GMM for a desired speaker pair. Moucharitsis et al. [4] proposed an unsupervised training method based on maximum likelihood constrained adaptation of the GMM trained with an existing parallel data set of a different speaker pair. Lee et al. [5] proposed another unsupervised training method based on maximum a posteriori (MAP) adaptation.

In order to use more informative prior knowledge extracted from many other speakers, we proposed eigenvoice conversion (EVC) [6]. This method has brought new VC frameworks, i.e., one-to-many EVC and many-to-one EVC [7]. The one-to-many EVC framework allows the conversion from a specific source speaker's voice into an arbitrary target speaker's voice and many-to-one EVC allows the conversion in reverse. Eigenvoice GMM (EV-GMM) is trained in advance with multiple parallel data sets consisting of a single pre-defined speaker, called a reference speaker in this paper, and many pre-stored speakers. The GMM between the reference speaker and an arbitrary speaker is flexibly developed by estimating a small amount of free parameters of the EV-GMM, i.e., weights for eigenvectors, using only a few utterances of the adapted speaker in an unsupervised manner. Although one-to-many EVC and many-to-one EVC are very flexible compared with the conventional VC, it is still hard to flexibly perform the conversion between arbitrary speaker-pairs.

In this paper, we propose many-to-many VC as a much more flexible VC framework by extending the conventional EVC frameworks. A basic idea of the proposed framework is to sequentially perform many-to-one EVC and one-to-many EVC through the reference speaker. In this paper, two many-to-many EVC algorithms are investigated; one is the sequential conversion based on multiple voice conversion [8], and the other is the sequential conversion sharing mixture components between many-to-one EVC and one-to-many EVC.

This paper is organized as follows. In Section 2, we describe the basic EVC framework. In Section 3, the proposed many-to-many EVC algorithms are described. In Section 4, we describe experimental evaluations. Finally, we summarize this paper in Section 5.

2. Eigenvoice conversion (EVC)
In this section, one-to-many EVC is described as a typical EVC framework. It is straightforward to apply EVC to many-to-one VC by switching the source and the target as described in [7].

2.1. Eigenvoice Gaussian mixture model (EV-GMM)
We employ 2D-dimensional acoustic features, $X_t = [x_t^\text{source}, \Delta x_t^\text{source}]^T$ (source speaker's) and $Y_t^{(s)} = [y_t^{(s)}(s), \Delta y_t^{(s)}(s)]^T$ (the $s$th target speaker's), respectively consisting of $D$-dimensional static and dynamic features, where $T$ denotes transposition of the vector. Joint probability density of time-aligned source and target features determined by DTW is modeled with EV-GMM as follows:

$$P(X_t, Y_t^{(s)}) = \sum_{m=1}^{M} \alpha_m N \left( \begin{bmatrix} X_t \ Y_t^{(s)} \end{bmatrix}^T ; \mu_m^{(X,Y)}(w^{(s)}), \Sigma_m^{(X,Y)} \right)$$

$$\mu_m^{(X,Y)}(w^{(s)}) = \frac{\mu_m^{(X)} + \mu_m^{(Y)}(w^{(s)})}{B_m + \mu_m^{(0)}(w^{(s)})} \Sigma_m^{(X,Y)}$$

where $N(x; \mu, \Sigma)$ denotes Gaussian distribution with mean vector $\mu$ and covariance matrix $\Sigma$. In EV-GMM, a target mean vector is modeled by the bias vector $b_m^{(0)}$, representative vectors $B_m = [b_m^{(1)}, b_m^{(2)}, \ldots, b_m^{(D)}]$ and the weight vector $w^{(s)}$. Acoustic features of an arbitrary target speaker are modeled by setting only $w^{(s)}$ to appropriate values. The other parameters $\lambda^{(EV)}$ including mixture component weights, source mean vectors, bias vectors, representative vectors and covariance matrices are tied for every target speaker.

2.2. Training of EV-GMM
We train the EV-GMM with speaker adaptive training (SAT) [9] as follows:
\[ \hat{\lambda}^{(E)}(X), \hat{w}^{(E)} = \arg \max_{\lambda^{(E)}, w} \prod_{t=1}^{T} P\left(X_t, Y^{(E)}(t) | \lambda^{(E)}, m^{(E)}(t) \right) \]

SAT estimates both canonical EV-GMM parameters \( \hat{\lambda}^{(E)} \) and a set of weight vectors \( \hat{w}^{(E)} = (w_{(1)}, \ldots, w_{(T)}) \) for individual pre-trained target speakers. We employ PCA-based EV-GMM [6] as the initial model for SAT.

2.3. Unsupervised adaptation of EV-GMM

The EV-GMM is adapted to a desired target speaker by estimating \( w \) in the sense of maximum likelihood as shown in [10]. In order to perform the unsupervised adaptation using only the target speaker’s speech data, the optimal weight vector \( \hat{w}^{(tar)} \) is determined by maximizing the likelihood of the marginal distribution as follows [6]:
\[ \hat{w}^{(tar)} = \arg \max_{w} \int P(X, Y^{(tar)} | \lambda^{(E)}, w) dX, \]
where \( Y^{(tar)} \) is a time sequence of the given target features.

2.4. Conversion with adapted EV-GMM

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 \( X = [X_1, \ldots, X_T] \) and \( Y = [Y_1, \ldots, Y_T] \), respectively. Converted static feature vectors \( \hat{y} = [\hat{y}_1, \ldots, \hat{y}_T]^T \) are determined by:
\[ \hat{y} = \arg \max_{\forall \ m} \sum_{\forall \ m} P(m | X_1, \lambda^{(E)}) P(Y_1 | X_1, m, \lambda^{(E)}, \hat{w}), \]

where \( \hat{w} \) is a time sequence of the given target features.

3. Many-to-many EVC Algorithms

3.1. Basic idea

In order to achieve many-to-many EVC, we employ many-to-one EVC and one-to-many EVC. Figure 1 shows a schematic image of the proposed many-to-many EVC process. We use only one EV-GMM (trained as a one-to-many EV-GMM in this paper). Because the EV-GMM models joint probability density, the one-to-many EV-GMM can also be used as the many-to-one EV-GMM by just switching the source and the target features. Given a small amount of adaptation data of arbitrary source and target speakers, ML estimates of the weight vectors, \( \hat{w}^{(s)} \) and \( \hat{w}^{(t)} \), for the source \( y^{(s)} \) and the target \( y^{(t)} \) are determined by Eq. (4), respectively. Then, the arbitrary source speaker’s voice is converted into the reference voice with many-to-one EVC. After that, the converted reference speaker’s voice is further converted into the arbitrary target speaker’s voice with one-to-many EVC.

3.2. Conversion algorithm based on multiple VC [8]

In the first step, we convert the source voice into the reference voice using the EV-GMM adapted to the source speaker. The ML estimate of a static and dynamic feature sequence of the reference voice \( \hat{X} \) is determined by maximizing the following likelihood function:
\[ \hat{X} = \arg \max_{\forall \ m} \sum_{\forall \ m} P(m | Y^{(s)}, \lambda^{(E)}), \hat{w}^{(t)} \]
\[ \times P(X | Y^{(s)}, m, \lambda^{(E)}), \hat{w}^{(t)} \). \]

In the manner described in section 2.4, the suboptimum mixture component sequence \( \hat{m}^{(s)} \) is determined as follows:
\[ \hat{m}^{(s)} = \arg \max_{\forall \ m} P(m | Y^{(s)}, \lambda^{(E)}), \hat{w}^{(t)} \]. \]

Then, the converted reference voice is determined as follows:
\[ \hat{X} = \arg \max_{\forall \ m} P(X | Y^{(s)}, \hat{m}^{(s)}, \lambda^{(E)}), \hat{w}^{(t)} \). \]

In the second step, we convert the converted reference voice into the target voice using EV-GMM adapted to the target speaker. We estimate a target static feature sequence \( \hat{y}^{(o)} \) by maximizing the following likelihood function:
\[ \hat{y}^{(o)} = \arg \max_{\forall \ m} \sum_{\forall \ m} P(m | \hat{X}, \lambda^{(E)}), \hat{w}^{(t)} \]
\[ \times P(Y^{(o)} | X, m, \lambda^{(E)}), \hat{w}^{(t)} \). \]

In addition, the converted speech quality is dramatically improved by considering global variance (GV) [3], which is defined as the variance of the target static features calculated over a time sequence. A metric on the GV of the converted parameters effectively works as a penalty term in the conversion process for preventing the parameters from being over-smoothed.
We performed many-to-many EVC using a one-to-many EVC-GMM. We used one male speaker as the reference speaker and 160 pre-stored speakers including 80 male and 80 female speakers [11]. Each speaker uttered 50 phoneme-balanced sentences as shown in [6].

In the evaluations, we used eight speaker pairs (two male-to-male pairs, two female-to-female pairs, two male-to-female pairs, and two female-to-male pairs) selected from five male and five female speakers that were not included in the pre-stored speakers. We used 1 to 32 utterances for the adaptation, and the other 21 utterances for the evaluations.

We used 24-dimensional mel-cepstrum analyzed by STRAIGHT [12] as a spectral feature and aperiodic components [13] that were averaged on five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 kHz) as an excitation feature to construct STRAIGHT mixed excitation [13, 14]. Two one-to-many EVC-GMMs were trained separately for these two features. The number of representative vectors was 159 for mel-cepstrum and 64 for aperiodic components, respectively. The number of mixture components was 128 for mel-cepstrum and 64 for the aperiodic components, respectively. We converted source fundamental frequency to target one as follows:

\[
\log F_0 = \frac{\sigma_x}{\sigma_y} \left( \log F_0 - \mu_x \right) + \mu_y, 
\]  

where \(\mu_x\) and \(\sigma_x\) denote source mean and standard deviation, and \(\mu_y\) and \(\sigma_y\) denote target mean and deviation respectively.

We compared the proposed many-to-many EVC algorithms based on multiple VC “M-to-M (multiple)” and based on shared mixture components “M-to-M (shared)” with traditional VC with the parallel training “Traditional.” Note that the unsupervised adaptation was performed in the many-to-many EVC while the supervised training using parallel data was performed in the traditional VC.

### 4.2. Objective evaluations

We evaluated the conversion performance using mel-cepstral distortion for the spectral conversion and RMSE on aperiodic component [14] for the aperiodicity conversion. Figures 3 and 4 show results when varying the number of adaptation sentences (or the number of parallel training sentences in “Traditional”). The performance of both “M-to-M” methods is significantly better than “Traditional” when using a small amount of adaptation data. Moreover, we can observe that “M-to-M (shared)” outperforms “M-to-M (multiple).” When using more than 16 adaptation sentences, “Traditional” overcomes the proposed methods because a large enough amount of parallel data to train the GMM is available. Incidentally, we have never observed significant differences of the conversion performance between the within-gender conversion and the cross-gender conversion.

### 4.1. Experimental conditions

We performed many-to-many EVC using a one-to-many EV-GMM. We used one male speaker as the reference speaker and 160 pre-stored speakers including 80 male and 80 female speakers [11]. Each speaker uttered 50 phoneme-balanced sentences.
different types of the converted speech was presented to them. The proposed many-to-many EVC allows us to develop a conversion model from an arbitrary source speaker to an arbitrary target speaker by the unsupervised adaptation using a small amount of adaptation data of the arbitrary source and target speakers. Results of objective and subjective evaluations have demonstrated the effectiveness of the proposed algorithm.

6. Acknowledgements

This work was supported in part by MIC SCOPE (Strategic Information and Communications R&D Promotion Programme).

7. References