Abstract

One-to-many eigenvoice conversion (EVC) allows the conversion of a specific source speaker into arbitrary target speakers. Eigenvoice Gaussian mixture model (EV-GMM) is trained in advance with multiple parallel data sets consisting of the source speaker and many pre-stored target speakers. The EV-GMM is adapted for arbitrary target speakers using only a few utterances by estimating a small number of free parameters. Therefore, the initial EV-GMM directly affects the conversion performance of the adapted EV-GMM. In order to prepare a better initial model, this paper proposes Speaker Adaptive Training (SAT) of a canonical EV-GMM in one-to-many EVC. Results of objective and subjective evaluations demonstrate that SAT causes significant improvements in the performance of EVC.

The paper is organized as follows: In Section 2, we describe EVC. In Section 3, SAT for EV-GMM is described. In Section 4, we describe experimental evaluations. Finally, we summarize this paper in Section 5.

2. Eigenvoice Conversion

2.1. Eigenvoice Gaussian Mixture Model (EV-GMM)

We use 2D-dimensional acoustic features, $X_i = [x_i^T, Δx_i^T]^T$ (source speaker’s) and $Y_i^{(s)} = [y_i^{(s)^T}, Δy_i^{(s)^T}]^T$ (the $s$-th target speaker’s), consisting of $D$-dimensional static and dynamic features, where $^T$ denotes transposition of the vector. Joint probability density of $Z_i^{(s)}$ is modeled with EV-GMM consisting of time-aligned source and target features determined by DTW is modeled with EV-GMM as follows:

$$
P(Z_i^{(s)}) = \sum_{i=1}^{M} \alpha_i N(Z_i^{(s)}, \mu_i^{(Z)}, \Sigma_i^{(ZZ)}).$$

$$\mu_i^{(Z)} = \begin{bmatrix} \mu_i^{(X)} \\ B_i \cdot w_s + b_i^{(0)} \end{bmatrix}$$

$$\Sigma_i^{(ZZ)} = \begin{bmatrix} \Sigma_i^{(XX)} & \Sigma_i^{(XY)} \\ \Sigma_i^{(YX)} & \Sigma_i^{(YY)} \end{bmatrix}$$

where $N(x; \mu, \Sigma)$ denotes Gaussian distribution with mean vector $\mu$ and diagonal covariance matrix $\Sigma$. In EV-GMM $\lambda^{(EV)}$, a target mean vector is modeled as linear combination with the bias vector $b_i^{(0)}$, representative vectors $B_i = [b_i^{(1)}, b_i^{(2)}, \ldots, b_i^{(3)}]$ and the weight vector $w_s$. EV-GMM models arbitrary target speaker’s individualities by setting $w_s$ to appropriate values. The other parameters such as mixture-weights, source mean vectors, bias vectors, representative vectors and covariance matrices are tied for every target speaker.
2.2. Training of EV-GMM Based on Principal Component Analysis

Firstly, TSI-GMM $\lambda^{(0)}$ is trained with multiple parallel data sets consisting of utterance-pairs of the source speaker and multiple pre-stored target speakers. And then, using each parallel data set, the $s$-th target dependent GMM $\lambda^{(s)}$ is estimated by only updating target mean vectors of $\lambda^{(0)}$. A $2DM$-dimensional supervector $SV^{(s)} = (\mu_Y^{(s)})^T \cdot \cdots \cdot (\mu_M^{(s)})^T$ is constructed for each pre-stored target speaker by concatenating the resulting target mean vectors. Finally, bias vector $b_i^{(0)}$ and representative vectors $B_i$ are extracted from the supervectors $SV^{(s)}$ by principal component analysis:

$$SV^{(s)} = [B_1^T, \cdots, B_M^T]^T w_s + [b_1^{(0)}^T, \cdots, b_M^{(0)}]^T \text{ (4)}$$

$$b_i^{(0)} = \frac{1}{S} \sum_{s=1}^{S} h_i^{(Y)}(s), \quad \text{ (5)}$$

where $S$ denotes the number of pre-stored target speakers and $w_s$ is $J(<S \ll 2DM)$ principle components for the $s$-th target speaker.

2.3. Speaker Adaptation for EV-GMM and Conversion

The EV-GMM is adapted for arbitrary target speakers by estimating the optimum weight vector for their given speech samples without any linguistic information. In one-to-many EVC, $\hat{w}$ is estimated so that a likelihood of the marginal distribution for a time sequence of the given target features $Y^{(tar)}$ is maximized [6] as follows:

$$\hat{w} = \arg \max \limits_{\hat{w}} \int P(X, Y^{(tar)} | \lambda^{(EV)}) \, dX. \quad \text{ (6)}$$

One-to-many EVC is also constructed conversion models with various voice characteristics by manipulating $w$ manually.

In the conversion process, we use the conversion method based on maximum likelihood estimation (MLE) considering dynamic features [10]. We use $2D$-dimensional source and target time sequences $X = [X_1^T, \cdots, X_T^T]$ and $Y = [Y_1^T, \cdots, Y_T^T]$ consisting of $D$-dimensional static and dynamic features. Converted static feature vectors $\bar{y} = [\bar{y}_1^T, \cdots, \bar{y}_T^T]$ can be obtained as follows:

$$\bar{y} = \arg \max \limits_{\bar{y}} \log P(Y \mid X, \bar{m}, \lambda^{(EV)}), \quad \text{ (7)}$$

Subject to $Y = W \bar{y}$, where $W$ denotes the matrix to extend the static feature sequence to the static and dynamic feature sequence, and $\bar{m}$ shows the optimum mixture sequence for maximizing the likelihood function $P(m \mid X, \lambda^{(EV)})$.

2.4. Problem of PCA-based EV-GMM

The tied parameters of the PCA-based EV-GMM are from the TSI-GMM. They are affected by acoustic variations of pre-stored target speakers. Target covariance values are, especially, much larger than those of the speaker dependent GMM. They would cause performance degradation of the adapted EV-GMM.

3. Speaker Adaptive Training for EV-GMM

In order to train an appropriate canonical EV-GMM, we apply speaker adaptive training (SAT) to the EV-GMM training.

The canonical EV-GMM is trained by maximizing likelihood of the adapted models for individual pre-stored target speakers as follows:

$$\hat{\lambda}^{(EV)}(\bar{w}^{S}) = \arg \max \lambda \prod_{s=1}^{S} \prod_{t=1}^{T} P(Z^{(s)} \mid \lambda^{(EV)}(w_s)), \quad \text{ (8)}$$

where $\lambda^{(EV)}(w_s)$ denotes the adapted model for the $s$-th pre-stored target speaker with the weight vector $w_s$. SAT estimates both canonical EV-GMM parameters $\hat{\lambda}^{(EV)}$ and a set of pre-stored target weight vectors $\bar{w}^{S}$ = $(\bar{w}_1, \cdots, \bar{w}_S)$. The estimation is performed with EM algorithm by maximizing the following auxiliary function:

$$Q(\hat{\lambda}^{(EV)}(\bar{w}^{S}), \lambda^{(EV)}(\bar{w}^{S})) = \sum_{s=1}^{S} \sum_{t=1}^{T} \tilde{z}^{(s)} \log P(Z^{(s)} \mid m_s, \lambda^{(EV)}(\bar{w}^{S})) \), \quad \text{ (9)}$$

where

$$\tilde{z}^{(s)} = \sum_{t=1}^{T} P(m_s | Z^{(s)} \mid \lambda^{(EV)}(\bar{w}^{S})).$$

It is difficult to update all parameters simultaneously because some of them depend on each other. Therefore, each parameter of EV-GMM is updated as follows:

$$Q(\hat{\lambda}^{(EV)}(\bar{w}^{S}), \lambda^{(EV)}(\bar{w}^{S})) \leq Q(\lambda^{(EV)}(w_1^{S}), (w_1^{S}, \alpha_i, B_i, b_i^{(0)}, \mu_i(X), \Sigma_i)) \leq Q(\lambda^{(EV)}(w_1^{S}), (w_1^{S}, \alpha_i, B_i, b_i^{(0)}, \mu_i(X), \Sigma_i)) \leq Q(\lambda^{(EV)}(w_1^{S}), (w_1^{S}, \alpha_i, B_i, b_i^{(0)}, \mu_i(X), \Sigma_i)).$$

ML estimates of the weight vector for the $s$-th pre-stored target speaker is written as

$$\bar{w}_s = \left( \sum_{i=1}^{M} \tilde{z}_i^{(s)} B_i P^{(Y \mid X)} \right)^{-1} \times \left[ \sum_{i=1}^{M} \left( B_i P^{(Y \mid X)} (\bar{X}_i^{(s)} - \tilde{z}_i^{(s)} \mu_i(X)) \right) + B_i P^{(Y \mid X)} (\bar{Y}_i^{(s)} - \tilde{z}_i^{(s)} b_i^{(0)}) \right].$$

where

$$\tilde{z}_i^{(s)} = \left[ \sum_{t=1}^{T} P(m_s | Z^{(s)} \mid \lambda^{(EV)}(w_s)) X_i^{(s)} \right], \quad \text{ (10)}$$

and

$$\Sigma_i^{(ZZ)}^{-1} = \left[ P^{(XX)} P^{(XY)} P^{(YY)} \right].$$
ML estimates of the tied parameters are written as:

$$\hat{\alpha}_t = \frac{\sum_{s=1}^{S} \hat{\alpha}_t^{(s)}}{\sum_{s=1}^{S} \bar{\gamma}_i(s)};$$

(11)

$$\hat{v}_t = \left( \sum_{s=1}^{S} \hat{z}_i^{(s)} W_s^T \Sigma_i^{(ZZ)}^{-1} \bar{W}_s \right) \left( \sum_{s=1}^{S} \bar{W}_s^T \Sigma_i^{(ZZ)}^{-1} \bar{Z}_i^{(s)} \right)^{-1},$$

(12)

$$\Sigma_i^{(ZZ)} = \frac{1}{\sum_{s=1}^{S} \hat{z}_i^{(s)}} \sum_{s=1}^{S} \left\{ \bar{v}_i^{(s)} + \bar{z}_i^{(s)} (\bar{\mu}_i^{(s)} - \mu_i^{(s)})^T - (\bar{\mu}_i^{(s)} \bar{Z}_i^{(s)} + \bar{Z}_i^{(s)} \bar{\mu}_i^{(s)})^T \right\},$$

(13)

where

$$V_i^{(s)}(x) = \sum_{t=1}^{T} p(m_i|Z_i^{(s)}, \lambda^{(EV)}(w_i)) Z_i^{(s)} Z_i^{(s)^T},$$

$$\mu_i^{(s)} = W_s \hat{v}_t = \left[ \begin{array}{c} \mu_i^{(X)} \\ \bar{b}_i^{(0)} + \bar{b}_i^{(0)} \\ \bar{b}_i^{(1)} \\ \vdots \\ \bar{b}_i^{(J)} \end{array} \right]^T,$$

$$\hat{v}_t = \left[ \begin{array}{c} \mu_i^{(x)} \\ \mu_i^{(y)} \end{array} \right]^T,$$

$$\hat{W}_s = \left[ \begin{array}{cccc} I & 0 & 0 & 0 \\ 0 & I & \hat{w}_i^{(1)} & \hat{w}_i^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \hat{w}_i^{(J)} & I \end{array} \right],$$

and the matrix $I$ is a $D \times D$ unit matrix.

In equation (12), we need to calculate the $(D \times (J+2)) \times (D \times (J+2))$-sized inverse matrix of $\sum_{s=1}^{S} \hat{z}_i^{(s)} W_s^T \Sigma_i^{(ZZ)}^{-1} W_s$. In the case of using diagonal covariance matrices $\Sigma_i^{(X)}, \Sigma_i^{(Y)}$ and $\Sigma_i^{(Y)}, \Sigma_i^{(Z)}$, $\Sigma_i^{(X)}$ and $\Sigma_i^{(Y)}$, $\Sigma_i^{(Y)}$ of SAT-based EV-GMM, PCA-based EV-GMM, and conventional one-to-one GMM. Figure 2 shows the mean of variances for target features of individual mixtures. Values of covariance components of PCA-based EV-GMM are large because PCA-based EV-GMM includes acoustic variations among pre-stored target speakers. On the other hand, those of SAT-based EV-GMM are almost equal to those of one-to-one GMM because SAT normalizes those variations.

4.2. Objective Evaluations

4.2.1. Comparison of EV-GMM

Figure 1 shows log-scaled likelihood as a function of the number of iterations. The value of the 0-th iteration is the result of TSI-GMM. Compared with PCA-based EV-GMM, log-scaled likelihood increases by updating mixture weight or representative vectors. It increases greatly by updating covariance matrices, and it is close to log-scaled likelihood when updating all parameters. Therefore, the update of covariance matrices is the most effective for improving the EV-GMM.

We compared static components of the target covariances $\Sigma_i^{(Y)}$ of SAT-based EV-GMM, PCA-based EV-GMM, and conventional one-to-one GMM. Figure 2 shows the mean of variances for target features of individual mixtures. Values of covariance components of PCA-based EV-GMM are large because PCA-based EV-GMM includes acoustic variations among pre-stored target speakers. On the other hand, those of SAT-based EV-GMM are almost equal to those of one-to-one GMM because SAT normalizes those variations.

4.2.2. Comparison of Spectral Distortion

We evaluated spectral conversion accuracy by comparing distortion between target and converted features. Figure 3 shows mel-cepstral distortion as a function of the number of adaptation utterances. SAT-based EV-GMM works better than PCA-based EV-GMM in each number of adaptation utterances because SAT-based EV-GMM models reasonable target covariances.

4.3. Subjective Evaluations

We conducted preference tests on speech quality and conversion accuracy for speaker individuality. We performed an AB test (A and B: converted voices with PCA-based and SAT-based EV-GMM, respectively) on speech quality and an XAB test (X: target speech, A and B: converted voices with PCA-based and SAT-based EV-GMM, respectively) on the conversion accuracy. In the AB test, listeners were asked which converted voice...
sounded better. In the XAB test, listeners were asked which converted voice sounded similar to the target speech. The number of listeners was five and the number of adaptation utterances was set to two in each evaluation.

Figure 4 shows the results of the preference tests. In the test of speech quality, SAT-based EV-GMM outperformed PCA-based EV-GMM. Converted voices with SAT-based EV-GMM were more intelligible than those with PCA-based EV-GMM. In the test of conversion accuracy for speaker individuality, the performance of the SAT-based EV-GMM is almost equal to that of PCA-based EV-GMM.

5. Conclusions
To improve the performance of one-to-many eigenvoice conversion (EVC), we proposed Speaker Adaptive Training (SAT) for the eigenvoice Gaussian mixture model (EV-GMM). We evaluated the effectiveness of the proposed method objectively and subjectively. Experimental results demonstrated that SAT-based EV-GMM outperforms the conventional PCA-based EV-GMM.

Although the performance of EV-GMM is improved by SAT, the quality of the converted speech is not enough. To obtain high quality speech, we have to introduce global variance [10] and STRAIGHT mixed excitation [13] in EVC framework.

6. Acknowledgements
This research was supported in part by the Japanese Ministry of Education, Culture, Sports, Science and Technology e-Society project.

7. References