



Speaker Adaptive Training for One-to-Many Eigenvoice Conversion Based on Gaussian Mixture Model

Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science,
Nara Institute of Science and Technology, Japan

{yamato-o, tomoki, sawatari, shikano}@is.naist.jp

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.

Index Terms: Speech synthesis, Voice conversion, Eigenvoice, Gaussian mixture model, Speaker adaptive training

1. Introduction

Voice conversion (VC) is a technique for modifying non-linguistic information such as voice characteristics while not changing linguistic information. One typical application of voice conversion is speaker conversion [1], and this application can be extended to cross-language speaker conversion [2][3], which is a technique that makes it possible for us to speak any language with our own voice.

A conversion method based on a Gaussian mixture model (GMM) [4] is used widely. This method trains a GMM with a parallel data set consisting of utterance pairs of source and target speakers. In practice, this training framework causes many limitations to VC applications. In order to make the training framework more flexible, Eigenvoice conversion (EVC) [5] has been proposed by applying eigenvoices [6] to the GMM-based conversion method.

There are two main frameworks in EVC, one-to-many EVC and many-to-one EVC [7]. One-to-many EVC allows the conversion of a specific source speaker into arbitrary target speakers. Eigenvoice GMM (EV-GMM) is trained in advance with multiple parallel data sets consisting of the source speaker and many pre-stored target speakers. The resulting EV-GMM enables us to control voice quality of the converted speech by manipulating a few free parameters, i.e. weights for eigenvectors. Moreover, we can estimate appropriate values of those parameters for given target speaker's voices without any linguistic information.

The conventional EV-GMM is based on the target speaker independent GMM (TSI-GMM). Although it works well, it would not be appropriate as an initial model for speaker adaptation, because the speaker independent model includes acoustic

variations among speakers. Therefore, as the initial model, a normalized speaker model is preferable to the speaker independent model. To train such a canonical model, adaptive training techniques such as speaker adaptive training (SAT) [8] and cluster adaptive training (CAT) [9] have been proposed in speech recognition area. This paper proposes SAT for EV-GMM in one-to-many EVC. Effectiveness of the proposed method is demonstrated through objective and subjective evaluations.

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, $\mathbf{X}_t = [\mathbf{x}_t^\top, \Delta\mathbf{x}_t^\top]^\top$ (source speaker's) and $\mathbf{Y}_t^{(s)} = [\mathbf{y}_t^{(s)\top}, \Delta\mathbf{y}_t^{(s)\top}]^\top$ (the s -th target speaker's), consisting of D -dimensional static and dynamic features, where \top denotes transposition of the vector. Joint probability density of $\mathbf{Z}_t^{(s)} = [\mathbf{X}_t^\top, \mathbf{Y}_t^{(s)\top}]^\top$ consisting of time-aligned source and target features determined by DTW is modeled with EV-GMM as follows:

$$P\left(\mathbf{Z}_t^{(s)} | \lambda^{(EV)}\right) = \sum_{i=1}^M \alpha_i N\left(\mathbf{Z}_t^{(s)}; \boldsymbol{\mu}_i^{(Z)}, \boldsymbol{\Sigma}_i^{(ZZ)}\right), \quad (1)$$

$$\boldsymbol{\mu}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(X)} \\ \mathbf{B}_i \mathbf{w}_s + \mathbf{b}_i^{(0)} \end{bmatrix}, \quad (2)$$

$$\boldsymbol{\Sigma}_i^{(ZZ)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix}, \quad (3)$$

where $N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes Gaussian distribution with mean vector $\boldsymbol{\mu}$ and diagonal covariance matrix $\boldsymbol{\Sigma}$. In EV-GMM $\lambda^{(EV)}$, a target mean vector is modeled as linear combination with the bias vector $\mathbf{b}_i^{(0)}$, representative vectors $\mathbf{B}_i = [\mathbf{b}_i^{(1)}, \mathbf{b}_i^{(2)}, \dots, \mathbf{b}_i^{(J)}]$ and the weight vector \mathbf{w}_s . EV-GMM models arbitrary target speaker's individualities by setting \mathbf{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 $\mathbf{SV}^{(s)} = [\boldsymbol{\mu}_1^{(Y)\top}(s), \dots, \boldsymbol{\mu}_M^{(Y)\top}(s)]^\top$ is continued for each pre-stored target speaker by concatenating the resulting target mean vectors. Finally, bias vector $\mathbf{b}_i^{(0)}$ and representative vectors \mathbf{B}_i are extracted from the supervectors $\mathbf{SV}^{(s)}$ by principal component analysis:

$$\begin{aligned} \mathbf{SV}^{(s)} &\simeq [\mathbf{B}_1^\top, \dots, \mathbf{B}_M^\top]^\top \mathbf{w}_s + [\mathbf{b}_1^{(0)\top}, \dots, \mathbf{b}_M^{(0)\top}]^\top \quad (4) \\ \mathbf{b}_i^{(0)} &= \frac{1}{S} \sum_{s=1}^S \boldsymbol{\mu}_i^{(Y)}(s), \quad (5) \end{aligned}$$

where S denotes the number of pre-stored target speakers and \mathbf{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, \mathbf{w} is estimated so that a likelihood of the marginal distribution for a time sequence of the given target features $\mathbf{Y}^{(tar)}$ is maximized [6] as follows:

$$\hat{\mathbf{w}} = \arg \max_{\mathbf{w}} \int P(\mathbf{X}, \mathbf{Y}^{(tar)} | \lambda^{(EV)}) d\mathbf{X}. \quad (6)$$

One-to-many EVC is also constructed conversion models with various voice characteristics by manipulating \mathbf{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 $\mathbf{X} = [\mathbf{X}_1^\top, \dots, \mathbf{X}_T^\top]^\top$ and $\mathbf{Y} = [\mathbf{Y}_1^\top, \dots, \mathbf{Y}_T^\top]^\top$ consisting of D -dimensional static and dynamic features. Converted static feature vectors $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_1^\top, \dots, \hat{\mathbf{y}}_T^\top]^\top$ can be obtained as follows:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} \log P(\mathbf{Y} | \mathbf{X}, \hat{\mathbf{m}}, \lambda^{(EV)}), \quad (7)$$

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 $\hat{\mathbf{m}}$ shows the optimum mixture sequence for maximizing the likelihood function $P(m|\mathbf{X}, \lambda^{(EV)})$.

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:

$$\hat{\lambda}^{(EV)}(\hat{\mathbf{w}}_1^S) = \arg \max_{\lambda} \prod_{s=1}^S \prod_{t=1}^{T_s} P(\mathbf{Z}_t^{(s)} | \lambda^{(EV)}(\mathbf{w}_s)), \quad (8)$$

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

$$\begin{aligned} Q(\lambda^{(EV)}(\mathbf{w}_1^S), \hat{\lambda}^{(EV)}(\hat{\mathbf{w}}_1^S)) \\ = \sum_{s=1}^S \sum_{i=1}^M \tilde{\gamma}_i^{(s)} \log P(\mathbf{Z}^{(s)}, m_i | \hat{\lambda}^{(EV)}(\hat{\mathbf{w}}_1^S)), \quad (9) \end{aligned}$$

where

$$\tilde{\gamma}_i^{(s)} = \sum_{t=1}^{T_s} P(m_i | \mathbf{Z}_t^{(s)}, \lambda^{(EV)}(\mathbf{w}_s)).$$

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:

$$\begin{aligned} Q(\lambda^{(EV)}(\mathbf{w}_1^S), \lambda^{(EV)}(\mathbf{w}_1^S)) \\ \leq Q(\lambda^{(EV)}(\mathbf{w}_1^S), (\hat{\mathbf{w}}_1^S, \alpha_i, \mathbf{B}_i, \mathbf{b}_i^{(0)}, \boldsymbol{\mu}_i^{(X)}, \boldsymbol{\Sigma}_i^{(zz)})) \\ \leq Q(\lambda^{(EV)}(\mathbf{w}_1^S), (\hat{\mathbf{w}}_1^S, \hat{\alpha}_i, \hat{\mathbf{B}}_i, \hat{\mathbf{b}}_i^{(0)}, \hat{\boldsymbol{\mu}}_i^{(X)}, \hat{\boldsymbol{\Sigma}}_i^{(zz)})) \\ \leq Q(\lambda^{(EV)}(\mathbf{w}_1^S), (\hat{\mathbf{w}}_1^S, \hat{\alpha}_i, \hat{\mathbf{B}}_i, \hat{\mathbf{b}}_i^{(0)}, \hat{\boldsymbol{\mu}}_i^{(X)}, \hat{\boldsymbol{\Sigma}}_i^{(zz)})). \end{aligned}$$

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

$$\begin{aligned} \hat{\mathbf{w}}_s = & \left(\sum_{i=1}^M \tilde{\gamma}_i^{(s)} \mathbf{B}_i^\top \mathbf{P}_i^{(YY)} \mathbf{B}_i \right)^{-1} \\ & \times \left[\sum_{i=1}^M \left\{ \mathbf{B}_i^\top \mathbf{P}_i^{(YX)} \left(\bar{\mathbf{X}}_i^{(s)} - \tilde{\gamma}_i^{(s)} \boldsymbol{\mu}_i^{(X)} \right) \right. \right. \\ & \left. \left. + \mathbf{B}_i^\top \mathbf{P}_i^{(YY)} \left(\bar{\mathbf{Y}}_i^{(s)} - \tilde{\gamma}_i^{(s)} \mathbf{b}_i^{(0)} \right) \right\} \right], \quad (10) \end{aligned}$$

where

$$\bar{\mathbf{Z}}_i^{(s)} = \begin{bmatrix} \bar{\mathbf{X}}_i^{(s)} \\ \bar{\mathbf{Y}}_i^{(s)} \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T_s} p(m_i | \mathbf{Z}_t^{(s)}, \lambda^{(EV)}(\mathbf{w}_s)) \mathbf{X}_t^{(s)} \\ \sum_{t=1}^{T_s} p(m_i | \mathbf{Z}_t^{(s)}, \lambda^{(EV)}(\mathbf{w}_s)) \mathbf{Y}_t^{(s)} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_i^{(ZZ)^{-1}} = \begin{bmatrix} \mathbf{P}_i^{(XX)} & \mathbf{P}_i^{(XY)} \\ \mathbf{P}_i^{(YX)} & \mathbf{P}_i^{(YY)} \end{bmatrix}.$$

ML estimates of the tied parameters are written as

$$\hat{\alpha}_i = \frac{\sum_{s=1}^S \tilde{\gamma}_i^{(s)}}{\sum_{i=1}^M \sum_{s=1}^S \tilde{\gamma}_i^{(s)}}, \quad (11)$$

$$\hat{\mathbf{v}}_i = \left(\sum_{s=1}^S \tilde{\gamma}_i^{(s)} \hat{\mathbf{W}}_s^\top \Sigma_i^{(ZZ)^{-1}} \hat{\mathbf{W}}_s \right)^{-1} \left(\sum_{s=1}^S \hat{\mathbf{W}}_s^\top \Sigma_i^{(ZZ)^{-1}} \bar{\mathbf{z}}_i^{(s)} \right), \quad (12)$$

$$\hat{\Sigma}_i^{(ZZ)} = \frac{1}{\sum_{s=1}^S \tilde{\gamma}_i^{(s)}} \sum_{s=1}^S \left\{ \bar{\mathbf{v}}_i^{(s)} + \tilde{\gamma}_i^{(s)} \hat{\boldsymbol{\mu}}_i^{(s)} \hat{\boldsymbol{\mu}}_i^{(s)\top} - \left(\hat{\boldsymbol{\mu}}_i^{(s)} \bar{\mathbf{z}}_i^{(s)\top} + \bar{\mathbf{z}}_i^{(s)} \hat{\boldsymbol{\mu}}_i^{(s)\top} \right) \right\}, \quad (13)$$

where

$$\begin{aligned} \bar{\mathbf{v}}_i^{(s)} &= \sum_{t=1}^{T_s} p(m_i | \mathbf{z}_t^{(s)}, \lambda^{(EV)}(\mathbf{w}_s)) \mathbf{z}_t^{(s)} \mathbf{z}_t^{(s)\top}, \\ \hat{\boldsymbol{\mu}}_i^{(s)} &= \hat{\mathbf{W}}_s \hat{\mathbf{v}}_i = \begin{bmatrix} \hat{\boldsymbol{\mu}}_i^{(X)} \\ \hat{\mathbf{B}}_i \hat{\mathbf{w}}_s + \hat{\mathbf{b}}_i^{(0)} \end{bmatrix}, \\ \hat{\mathbf{v}}_i &= \begin{bmatrix} \hat{\boldsymbol{\mu}}_i^{(X)\top}, \hat{\mathbf{b}}_i^{(0)\top}, \hat{\mathbf{b}}_i^{(1)\top}, \dots, \hat{\mathbf{b}}_i^{(J)\top} \end{bmatrix}^\top, \\ \hat{\mathbf{W}}_s &= \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \hat{w}_1^{(s)} \mathbf{I} & \hat{w}_2^{(s)} \mathbf{I} & \dots & \hat{w}_J^{(s)} \mathbf{I} \end{bmatrix}, \end{aligned}$$

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

In equation (12), we need to calculate $\{D \cdot (J+2)\} \times \{D \cdot (J+2)\}$ -sized inverse matrix of $\sum_{s=1}^S \tilde{\gamma}_i^{(s)} \hat{\mathbf{W}}_s^\top \Sigma_i^{(ZZ)^{-1}} \hat{\mathbf{W}}_s$. In the case of using diagonal covariance matrices $\Sigma^{(XX)}$, $\Sigma^{(XY)}$ and $\Sigma^{(YY)}$, $\sum_{s=1}^S \tilde{\gamma}_i^{(s)} \hat{\mathbf{W}}_s^\top \Sigma_i^{(ZZ)^{-1}} \hat{\mathbf{W}}_s$ is the block diagonal matrix. Computational cost can be reduced significantly by calculating $(J+2) \times (J+2)$ -sized inverse matrices for individual dimensions separately.

4. Experimental Evaluation

We objectively and subjectively evaluate the conversion performance of SAT-based EV-GMM compared with that of the PCA-based EV-GMM in one-to-many EVC.

4.1. Experimental Conditions

We used parallel data sets of one source male speaker and 160 pre-stored target speakers consisting of 80 male and 80 female speakers for training the EV-GMM. Each speaker uttered 50 phoneme-balanced sentences (details are in [5]). These speakers had been included in the Japanese Newspaper Article Sentences (JNAS) database [11]. We used PCA-based EV-GMM based on PCA as the initial model for SAT.

In evaluations, we used 10 target speakers consisting of five male and five female speakers who had not been included among the pre-stored target speakers. We used 1 to 32 utterances for the adaptation, and 21 utterances for evaluations.

We used 24-dimensional mel-cepstrum as spectral features analyzed by STRAIGHT [12]. The number of representative vectors was 159. We converted source fundamental frequency

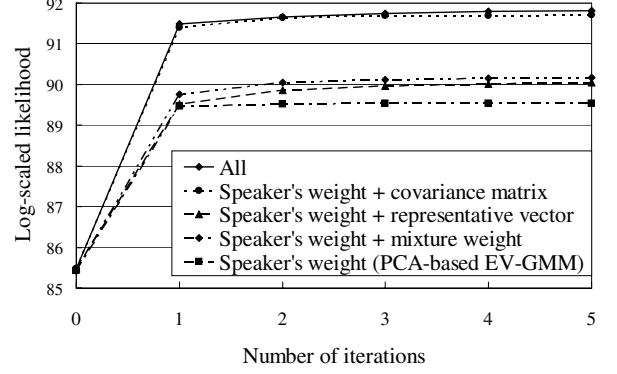


Figure 1: Log-scaled likelihood as a function of the number of iterations

to target frequency as follows:

$$\log \tilde{F}_0 = \frac{\sigma_y}{\sigma_x} (\log F_0 - \mu_{(x)}) + \mu_{(y)}, \quad (14)$$

where $\mu_{(x)}$ and σ_x denote source mean and standard deviation, and $\mu_{(y)}$ and σ_y denote target mean and deviation respectively. The number of mixtures was set to 128.

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 $\Sigma_i^{(YY)}$ 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

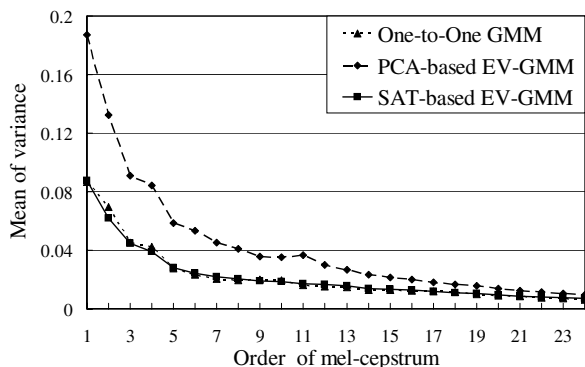


Figure 2: Mean of variances for target features of individual mixtures

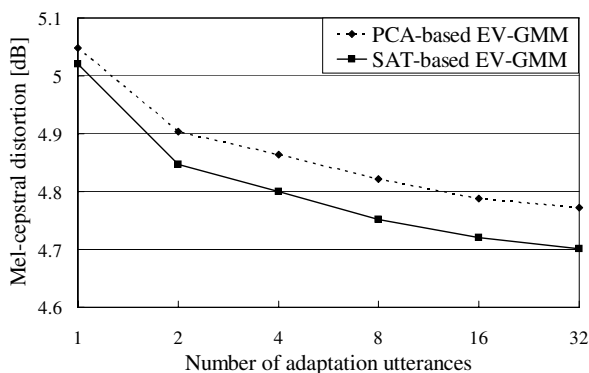


Figure 3: Mel-cepstral distortion as a function of the number of adaptation utterances

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

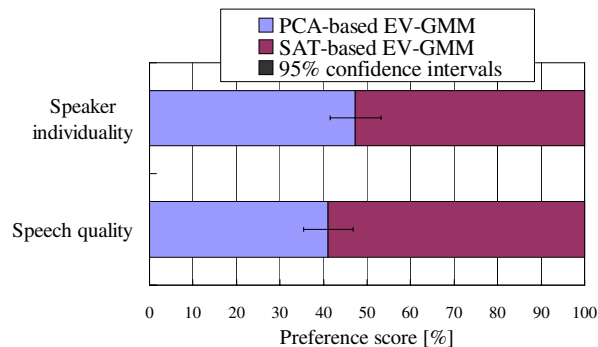


Figure 4: Results of subjective evaluation

project.

7. References

- [1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, Vol. 11, No. 2, pp. 71–76, 1990.
- [2] M. Abe, K. Shikano, and H. Kuwabara, "Statistical analysis of bilingual speaker's speech for cross-language voice conversion," *J. Acoust. Soc. Am*, Vol. 90, No. 1, pp. 76–82, 1991.
- [3] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," *IPSJ Journal*, Vol. 43, No. 7, pp. 2177–2185, 2002.
- [4] 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, 1998.
- [5] T. Toda, Y. Ohtani, K. Shikano, "Eigenvoice Conversion Based on Gaussian Mixture Model", *Proc. ICSLP*, pp. 2446–2449, September, 2006.
- [6] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space" *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 6, pp. 695–707, 2000.
- [7] T. Toda, Y. Ohtani, K. Shikano, "One-to-Many and Many-to-One Voice Conversion Based on Eigenvoices", *Proc. ICASSP*, April, 2007.
- [8] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, "A Compact Model for Speaker-Adaptive Training" *Proc. ICSLP*, vol.2, pp. 1137–1140, 1996.
- [9] M. J. F. Gales, "Cluster adaptive training for hidden Markov models," *IEEE Trans. Speech and Audio Process.* Vol. 8, No. 4, pp.417–428, 2000.
- [10] T. Toda, A.W. Black and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," *Proc. ICASSP*, pp. 9–12, March 2005.
- [11] JNAS: Japanese Newspaper Article Sentences. <http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html>
- [12] H. Kawahara, I. Masuda-Katsuse and A.de Cheveigné, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds, *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.
- [13] Y. Ohtani, T. Toda, H. Saruwatari, K. Shikano, "Maximum Likelihood Voice Conversion Based on GMM with STRAIGHT Mixed Excitation", *Proc. ICSLP*, pp. 2266–2269, September, 2006.