Eigenvoice Conversion Based on Gaussian Mixture Model

Tomoki Toda, Yamato Ohtani, Kiyohiro Shikano

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

tomoki@is.naist.jp

Abstract

This paper describes a novel framework of voice conversion (VC). We call it eigenvoice conversion (EVC). We apply EVC to the conversion from a source speaker's voice to arbitrary target speakers' voices. Using multiple parallel data sets consisting of utterance-pairs of the source and multiple pre-stored target speakers, a canonical eigenvoice GMM (EV-GMM) is trained in advance. That conversion model enables us to flexibly control the speaker individuality of the converted speech by manually setting weight parameters. In addition, the optimum weight set for a specific target speaker is estimated using only speech data of the target speaker without any linguistic restrictions. We evaluate the performance of EVC by a spectral distortion measure. Experimental results demonstrate that EVC works very well even if we use only a few utterances of the target speaker for the weight estimation.

Index Terms: speech synthesis, voice conversion, GMM, eigenvoice, unsupervised training.

1. Introduction

Voice conversion (VC) is a remarkable technique for flexibly modifying voice characteristics. There are many applications of VC such as a post-process of Text-to-Speech (TTS) for flexibly synthesizing speech of various speakers, an enhancement of speech quality for telecommunications, and a multi-lingual speech synthesizer.

Many statistical approaches to VC have been studied since the late 1980's [1]. Abe et al. [2] proposed a codebook mapping method based on hard clustering and discrete mapping. In order to directly model the correlation between source and target features, Valbret et al. [3] proposed a conversion method using linear multivariate regression (LMR), i.e., continuous mapping based on the hard clustering. As the most popular conversion method, Stylianou et al. [4] proposed a conversion method with a Gaussian mixture model (GMM). That method realizes the continuous mapping based on the soft clustering. Recently, Toda et al. [5] significantly improved the performance of the GMM-based conversion method by introducing maximum likelihood estimation (MLE) considering dynamic features and global variance (GV). That method shifts a conversion form from the conventional frame-based process to the trajectory-based one. It is indispensable to continue to make progress in the conversion method for making VC capable of practical applications.

Several approaches for improving a training method of the conversion function have been studied as well. As for the GMM-based conversion, Stylianou et al. [4] proposed a training method based on least mean square error (LMSE). In order to improve the robustness against a small amount of training data, Kain and Macon [6] proposed a training method based on joint density estimation (JDE). Those methods use a parallel data set consisting of utterance-pairs of source and target speakers. Such a training framework causes many limitations of VC applications. In order to address this problem, Mouchtaris et al. [7] proposed a non-parallel training method based on maximum likelihood constrained adaptation. The GMM trained with an existing parallel data set of a certain source and target speakers is adapted for the desired source and target speakers separately. This adaptation is supported by the fact that the feature correlation between a speaker-pair is useful as a prior knowledge for VC between another speaker-pair.

This paper describes a novel framework for the GMM-based conversion using the information extracted from a lot of pre-stored speakers as the prior knowledge. We call it eigenvoice conversion (EVC). The eigenvoice is a popular technique in the speech recognition area [9]. It realizes the hidden Markov model (HMM) speaker adaptation using a quite small amount of adaptation data by reducing the number of free parameters for controlling speaker dependencies of HMMs. It also realizes an HMM-based TTS having a voice quality controller [10] or a speaking style controller [11]. We apply that technique to the GMM-based conversion method for realizing flexible VC from the source speaker, e.g., an user into arbitrary speakers. EVC is similar to the speaker interpolation proposed by Iwahashi and Sagisaka [8] in terms of using the information of multiple pre-stored speakers. EVC outperforms it in view of the ability to convert any sample of the source to that of the target while the speaker interpolation can convert only feature segments included in the pre-stored database. Experimental results demonstrate that EVC works very well in the non-parallel training for arbitrary target speakers.

The paper is organized as follows. In **Section 2**, a framework of conventional VC is described. In **Section 3**, a framework of EVC is described. In **Section 4**, an experimental evaluation is described. Finally, we summarize this paper in **Section 5**.

2. Framework of Conventional Voice Conversion (VC)

2.1. Parallel Training

We use $2D$-dimensional acoustic features $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ (source speaker's) and $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ (target speaker's) consisting of D-dimensional static and dynamic features, where $\top$ denotes transposition of the vector. As described in [6], using parallel training data set consisting of time-aligned source and target features $[\boldsymbol{X}_1^\top, \boldsymbol{Y}_1^\top], \cdots, [\boldsymbol{X}_T^\top, \boldsymbol{Y}_T^\top]$ determined by Dynamic Time Warping (DTW), a GMM on joint probability density $p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\lambda})$ is trained in advance as follows:

$$\boldsymbol{\lambda} = \arg\max \textstyle\prod_{t=1}^{T} p(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}), \tag{1}$$

where $\boldsymbol{\lambda}$ denotes model parameters. The joint probability density is written as

$$p(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}) = \textstyle\sum_{i=1}^{M} \alpha_i N(\boldsymbol{X}_t, \boldsymbol{Y}_t; \boldsymbol{\mu}_i^{(X,Y)}, \boldsymbol{\Sigma}_i^{(X,Y)}),$$
$$\boldsymbol{\mu}_i^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{\mu}_i^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_i^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix}, \tag{2}$$

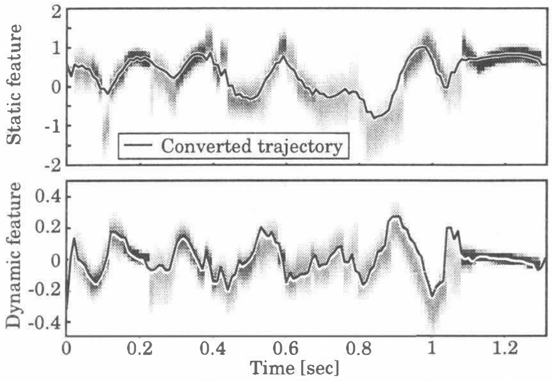

Figure 1: *An example of converted trajectory. Black parts show high conditional probability density areas.*

where $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ shows the normal distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The i^{th} mixture weight is α_i. The total number of mixtures is M. The model parameters can be estimated with the EM algorithm.

2.2. VC Based on MLE [5]

Let $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ be a time sequence of the source feature vectors, and let $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ be that of the target feature vectors. We perform the spectral conversion based on the maximization of the following likelihood function,

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) = \textstyle\sum_{\{\text{all } \boldsymbol{m}\}} p(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}) p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}), \tag{3}$$

where $\boldsymbol{m} = \{m_{i1}, m_{i2}, \cdots, m_{iT}\}$ is a mixture sequence. At frame t, $p(m_i|\boldsymbol{X}_t, \boldsymbol{\lambda})$ and $p(\boldsymbol{Y}_t|\boldsymbol{X}_t, m_i, \boldsymbol{\lambda})$ are given by

$$p(m_i|\boldsymbol{X}_t, \boldsymbol{\lambda}) = \frac{w_i N(\boldsymbol{X}_t; \boldsymbol{\mu}_i^{(X)}, \boldsymbol{\Sigma}_i^{(XX)})}{\sum_{j=1}^{M} w_j N(\boldsymbol{X}_t; \boldsymbol{\mu}_j^{(X)}, \boldsymbol{\Sigma}_j^{(XX)})}, \tag{4}$$

$$p(\boldsymbol{Y}_t|\boldsymbol{X}_t, m_i, \boldsymbol{\lambda}) = N(\boldsymbol{Y}_t; \boldsymbol{E}_t(m_i), \boldsymbol{D}(m_i)), \tag{5}$$

where

$$\boldsymbol{E}_t(m_i) = \boldsymbol{\mu}_i^{(Y)} + \boldsymbol{\Sigma}_i^{(YX)} \boldsymbol{\Sigma}_i^{(XX)^{-1}} (\boldsymbol{X}_t - \boldsymbol{\mu}_i^{(X)}), \tag{6}$$

$$\boldsymbol{D}(m_i) = \boldsymbol{\Sigma}_i^{(YY)} - \boldsymbol{\Sigma}_i^{(YX)} \boldsymbol{\Sigma}_i^{(XX)^{-1}} \boldsymbol{\Sigma}_i^{(XY)}. \tag{7}$$

A time sequence of the converted static features $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ is determined as follows:

$$\hat{\boldsymbol{y}} = \arg\max p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \qquad \text{subject to} \quad \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \tag{8}$$

where $\boldsymbol{W}$ is a transformation matrix from static features into static and dynamic features. The converted features can be estimated with the EM algorithm. We may significantly reduce computation time by approximating the likelihood function as $p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda})$ with the optimum mixture sequence $\hat{\boldsymbol{m}}$ that maximizes the posterior probability $p(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda})$.

Figure 1 shows an example of the converted trajectory on a time sequence of the conditional probability density functions, i.e., the approximated likelihood function. Note that the trajectory on the dynamic feature is derived from that on the static feature. This conversion method estimates the converted static features with appropriate both static and dynamic characteristics. We can considerably improve the naturalness of converted speech by further considering GV of the converted static features [5].

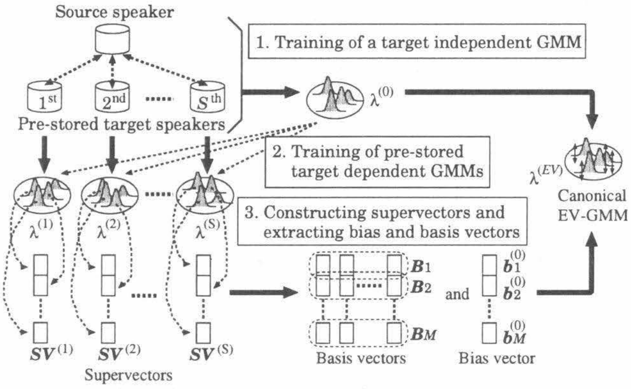
Source speaker
1. Training of a target independent GMM
1st
2nd
Sth
Pre-stored target speakers
λ(0)
2. Training of pre-stored target dependent GMMs
λ(EV)
Canonical EV-GMM
λ(1)
λ(2)
λ(S)
3. Constructing supervectors and extracting bias and basis vectors
B1
B2
BM
and
b1(0)
b2(0)
bM(0)
SV(1)
SV(2)
SV(S)
Supervectors
Basis vectors
Bias vector

Figure 2: *Training process of canonical EV-GMM.*

3. Framework of Eigenvoice Conversion (EVC)

In this paper, we describe only an example of EVC from one source speaker to arbitrary target speakers, i.e., one-to-many VC. We may apply EVC to other cases such as many-to-one VC or many-to-many VC.

3.1. Eigenvoice GMM (EV-GMM)

EV-GMM represents the joint probability density in the same manner as the conventional GMM shown in Eq. (2) except for a definition of the target mean vector written as

$$\boldsymbol{\mu}_i^{(Y)} = \boldsymbol{B}_i \boldsymbol{w} + \boldsymbol{b}_i^{(0)}, \tag{9}$$

where $\boldsymbol{b}_i^{(0)}$ is a bias vector for the i^{th} mixture. The matrix $\boldsymbol{B}_i = [\boldsymbol{b}_i(1), \cdots, \boldsymbol{b}_i(J)]$ consists of basis vectors $\boldsymbol{b}_i(j)$ for the i^{th} mixture. The number of basis vectors is J. The target speaker individuality is controlled with the J-dimensional weight vector $\boldsymbol{w} = [w(1), \cdots, w(J)]^\top$. Consequently, the EV-GMM has a parameter set $\boldsymbol{\lambda}^{(EV)}$ consisting of the single weight vector and parameters for individual mixtures such as the mixture weights, the source mean vectors, the bias and basis vectors, and the covariance matrices.

3.2. Training of Canonical EV-GMM

In order to train a canonical EV-GMM, we use multiple parallel data sets. Each of them consists of utterance-pairs of the source speaker and one of the multiple pre-stored target speakers. A training process of the canonical EV-GMM is shown in **Figure 2**.

Firstly, we train a target independent GMM $\boldsymbol{\lambda}^{(0)}$ simultaneously using all of the multiple parallel data sets as follows:

$$\boldsymbol{\lambda}^{(0)} = \arg\max \textstyle\prod_{s=1}^{S} \prod_{t=1}^{T_S} p(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}), \tag{10}$$

where $\boldsymbol{Y}_t^{(s)}$ is the feature vector of the s^{th} pre-stored target speaker at frame t. The number of feature vectors for the s^{th} speaker is T_s. The number of pre-stored target speakers is S.

Secondly, we train each target dependent GMM $\boldsymbol{\lambda}^{(s)}$ by updating only target mean vectors $\boldsymbol{\mu}_i^{(Y)}$ of the target independent GMM $\boldsymbol{\lambda}^{(0)}$ using each of the multiple parallel data sets as follows:

$$\boldsymbol{\lambda}^{(s)} = \arg\max \textstyle\prod_{t=1}^{T_S} p(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}). \tag{11}$$

In this paper, we employ the EM algorithm for the training of the target independent GMM and the target dependent GMMs.

Lastly, we determine the bias vector $b_i^{(0)}$ and the basis vectors B_i. We prepare a $(2D \times M)$-dimensional supervector $SV^{(s)} = [\mu_1^{(Y)}(s)^\top, \cdots, \mu_M^{(Y)}(s)^\top]^\top$ for each pre-stored target speaker by concatenating the target mean vectors $\mu_i^{(Y)}(s)$ of the target dependent GMM $\lambda^{(s)}$. Note that the correspondence of mixtures among all of the target dependent GMMs is obviously known because those GMMs are caused from the same target independent GMM while tying several parameters. We extract the basis vectors with principal component analysis (PCA) for the supervectors. Consequently, the supervector is written as

$$\begin{aligned} SV^{(s)} &\simeq [B_i^\top, \cdots, B_M^\top]^\top w^{(s)} + [b_i^{(0)\top}, \cdots, b_M^{(0)\top}]^\top, & (12) \\ b_i^{(0)\top} &= \tfrac{1}{S} \textstyle\sum_{s=1}^{S} \mu_i^{(Y)}(s), & (13) \end{aligned}$$

where $w^{(s)}$ consists of principal components for the s^{th} pre-stored target speaker. Now, various supervectors, i.e., the target mean vectors are created by varying only $J(< S \ll 2D \times M)$ free parameters of w. We construct the canonical EV-GMM $\lambda^{(EV)}$ from the resulting bias and basis vectors and the tied parameters, i.e., the mixture weights, the source mean vectors, and the covariance matrices. We may further update all of those parameters with the EM algorithm in a Speaker Adaptive Training (SAT) paradigm [12].

3.3. EVC in One-to-Many VC

We perform EVC based on MLE in the same manner as mentioned in **Section 2.2**. The conditional mean vector of the target for the i^{th} mixture in EVC is written as

$$E_t^{(EV)}(m_i) = B_i w + b_i^{(0)} + \Sigma_i^{(YX)} \Sigma_i^{(XX)^{-1}} (X_t - \mu_i^{(X)}). \quad (14)$$

We can see that varying the weight vector w causes shifts of the conditional mean vectors.

3.3.1. Non-Parallel (Unsupervised) Training

The EV-GMM for the conversion from the source speaker's voice to any target speaker's voice is created by estimating the optimum weight vector for the target speaker. Because only the target data is used for the estimation, we don't have to use the parallel data of the source and the target. Furthermore, we don't have to know sentences uttered by the target speaker. Namely, the non-parallel or unsupervised training for any target speaker is available.

We apply the maximum likelihood eigen-decomposition (MLED) [9] to the weight vector estimation in EVC as follows:

$$\begin{aligned} \hat{w} &= \arg\max \textstyle\int p(X, Y^{(tar)} | \lambda^{(EV)}) dX \\ &= \arg\max \textstyle\int p(Y^{(tar)} | \lambda^{(EV)}) p(X | Y^{(tar)}, \lambda^{(EV)}) dX \\ &= \arg\max p(Y^{(tar)} | \lambda^{(EV)}), & (15) \end{aligned}$$

where $Y^{(tar)}$ is a time sequence of the target features for the training. Because the probability density is modeled with a GMM, we iteratively maximize the following auxiliary function,

$$Q(w, \hat{w}) = \sum_{\text{all } m} p(m | Y^{(tar)}, \lambda^{(EV)}) \log p(Y^{(tar)}, m | \hat{\lambda}^{(EV)}). \quad (16)$$

The estimated weight vector is written as

$$\hat{w} = \left\{ \textstyle\sum_{i=1}^{M} \overline{\gamma}_i^{(tar)} B_i^\top \Sigma_i^{(yy)^{-1}} B_i \right\}^{-1} \textstyle\sum_{i=1}^{M} B_i^\top \Sigma_i^{(yy)^{-1}} \overline{Y}_i^{(tar)}, \quad (17)$$

where

$$\begin{aligned} \overline{\gamma}_i^{(tar)} &= \textstyle\sum_{t=1}^{T} p(m_i | Y_t^{(tar)}, \lambda^{(EV)}), & (18) \\ \overline{Y}_i^{(tar)} &= \textstyle\sum_{t=1}^{T} p(m_i | Y_t^{(tar)}, \lambda^{(EV)}) (Y_t^{(tar)} - b_i^{(0)}). & (19) \end{aligned}$$

We use the target independent GMM as an initial model.

An advantage of EVC compared with the constrained linear regression approach [7] is the robust parameter estimation when using a quite small amount of target data due to a smaller number of free parameters to be estimated.

3.3.2. Manual Control of Converted Speaker Individuality

The other advantage of EVC is that we can flexibly control the speaker individuality of the converted speech by manually modifying the weight vector. This causes a novel framework of VC allowing fine tuning.

If we would like to control not only spectral parameters but also the other speech parameters such as an F_0, it is possible to realize the weight vector simultaneously affecting all of those parameters by extracting basis vectors from supervectors consisting of their mean vectors. It might be possible to improve the controllability for the converted speaker individuality by perceptually designing the weight vector [11].

4. Experimental Evaluations

We objectively evaluated the spectral conversion accuracy of EVC compared with that of the conventional VC when varying the amount of target training data.

4.1. Experimental Conditions

In order to train the canonical EV-GMM, we used 160 speakers consisting of 80 male and 80 female speakers as the pre-stored target speakers. These speakers were included in Japanese Newspaper Article Sentences (JNAS) database [13]. Each of them uttered a set of phonetically balanced 50 sentences. Because 7 sub-sets were uttered by them as shown in **Table 1**, we used a male speaker not included in JNAS as the source speaker, who uttered 10 subsets including the 7 sub-sets. We automatically performed DTW between utterances of the source and each pre-stored target speaker for preparing parallel data sets.

We performed voice conversion from the source speaker to other 10 target speakers consisting of five male and five female speakers, who were not included in the pre-stored speakers. Every speaker uttered the sub-set J including 53 sentences. We varied the number of target training utterances from 1 to 32. The remaining 21 utterances were used for the evaluation. In the EVC, we estimated the weight vector of the canonical EV-GMM using only the target training data. In the conventional VC, we individually trained GMMs for the conversion between the source and 10 target speakers using parallel training data sets.

We used mel-cepstrum as a spectral feature. The first through 24th mel-cepstral coefficients were extracted from 16 kHz sampling speech data. The STRAIGHT analysis method [14] was employed for the spectral extraction.

In EVC, we used all eigenvectors (159 vectors) as the basis vectors without any loss of the information caused by truncating eigenvectors. The number of mixtures of the EV-GMM was constantly set to 512. On the other hand, we optimized the number of mixtures of the conventional GMM so that the mel-cpestral distortion between the converted and target mel-cepstra was minimized

Table 1: *The number of pre-stored target speakers uttering each sub-set (A, B, $\cdots$, or G). Each sub-set consists of phonetically balanced 50 sentences.*

Sub-sets	A	B	C	D	E	F	G	Total
Number of male speakers	15	11	15	13	15	11	0	80
Number of female speakers	15	11	15	13	12	0	14	80

Table 2: *The optimum number of mixtures for each size of training data in the conventional VC. Each number shows the average number of mixtures for 10 target speakers.*

Number of utterance-pairs	1	2	4	8	16	32
Number of mixtures	7.6	18.4	20.8	54.4	76.8	224.0

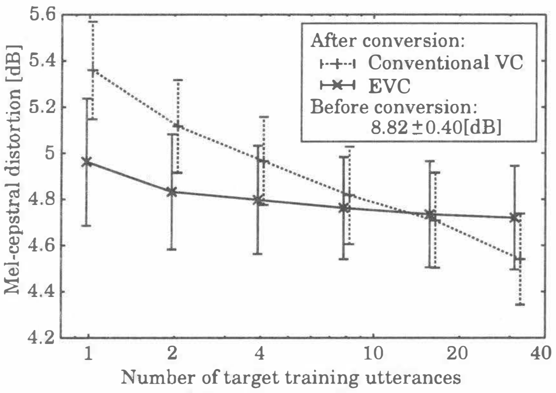

Figure 3: *Mel-cepstral distortion as a function of the number of target training utterances. We show mean distortions and its standard deviations for 10 target speakers.*

in the evaluation set. Such an optimization was separately performed for each target speaker and each size of training data. The optimization results are shown in **Table 2**. We used the diagonal covariance matrices for both GMMs.

4.2. Experimental Result

Figure 3 shows mel-cepstral distortion as a function of the number of target training utterances. EVC outperforms the conventional VC when using a small amount of training data. This is because EVC effectively uses the information extracted from a lot of pre-stored speakers as a prior knowledge.

We can see that in the conventional VC an increase of the amount of training data causes a large decrease of the distortion. The joint probability density is modeled more accurately as the optimum number of mixtures increases according to a larger number of the training data as shown in **Table 2**. On the other hand, we can see a tendency that the distortion decrease in EVC is not so large when increasing more than two target utterances due to the constant model complexity. Consequently, the conventional VC outperforms EVC when using dozens of target utterances. Note that EVC still has an advantage of unsupervised training compared with the conventional VC even in such cases.

It is observed that inter-speaker variances of the distortion in EVC are larger than those in the conventional VC. This might be caused by setting several parameters of EV-GMM to those of the target independent GMM. Applying SAT to the canonical EV-GMM training is very promising.

5. Conclusions

We proposed a novel voice conversion (VC) framework called eigenvoice conversion (EVC). We applied the eigenvoice technique to the conversion method with a Gaussian mixture model (GMM) for realizing VC from a source speaker to arbitrary target speakers (one-to-many VC). Statistics extracted from multiple parallel data sets consisting of the source's voices and the multiple pre-stored target speakers' voices were effectively used as a prior knowledge in EVC. We conducted an experimental evaluation on the spectral conversion accuracy. As a result, it was demonstrated that EVC works very well even if we have only a few utterances of the target speaker. We need to perform subjective evaluations of the converted speech with EVC as well.

The proposed idea of constructing a canonical VC model from multiple speaker-pairs' data sets seems to cause explosive spread of VC applications. We will apply it to various types of VC, e.g., many-to-one VC and many-to-many VC.

Acknowledgment: This research was supported in part by MIC SCOPE-S and MEXT e-Society leading project.

6. References

[1] H. Kuwabara and Y. Sagisaka. Acoustic characteristics of speaker individuality: control and conversion. *Speech Communication*, Vol. 16, No. 2, pp. 165–173, 1995.

[2] 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.

[3] H. Valbret, E. Moulines and J.P. Tubach. Voice transformation using PSOLA technique. *Speech Communication*, Vol. 11, No. 2–3, pp. 175–187, 1992.

[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, A.W. Black, and K. Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. *Proc. ICASSP*, Vol. 1, pp. 9–12, Philadelphia, USA, Mar. 2005.

[6] A. Kain and M.W. Macon. Spectral voice conversion for text-to-speech synthesis. *Proc. ICASSP*, pp. 285–288, Seattle, USA, May 1998.

[7] A. Mouchtaris, J.V. der Spiegel, and P. Mueller. Non-parallel training for voice conversion by maximum likelihood constrained adaptation. *Proc. ICASSP*, Vol. 1, pp. 1–4, Montreal, Canada, May 2004.

[8] N. Iwahashi and Y. Sagisaka. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. *Speech Communication*, Vol. 16, No. 2, pp. 139–151, 1995.

[9] 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.

[10] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. KItamura. Eigenvoice for HMM-based speech synthesis. *Proc. ICSLP*, Vol. 1, pp. 1269–1272, Denver, USA, Sep. 2002.

[11] K. Miyanaga, T. Masuko, T. Kobayashi. A style control technique for HMM-based speech synthesis. *Proc. ICSLP*, Jeju Island, Korea, Oct. 2004.

[12] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. *Proc. ICSLP*, pp. 1137–1140, Philadelphia, Oct. 1996.

[13] JNAS: Japanese Newspaper Article Sentences. *http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html*

[14] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an 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.