INTERSPEECH
2009 BRIGHTON

Cross-Language Voice Conversion Based on Eigenvoices

Malorie Charlier[1,2], *Yamato Ohtani*[1], *Tomoki Toda*[1], *Alexis Moinet*[2], *Thierry Dutoit*[2]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]Faculté Polytechnique de Mons, Belgium

Malorie.CHARLIER@student.fpms.ac.be, yamato-o@is.naist.jp, tomoki@is.naist.jp,
Alexis.Moinet@fpms.ac.be, Thierry.Dutoit@fpms.ac.be

Abstract

This paper presents a novel cross-language voice conversion (VC) method based on eigenvoice conversion (EVC). Cross-language VC is a technique for converting voice quality between two speakers uttering different languages each other. In general, parallel data consisting of utterance pairs of those two speakers are not available. To deal with this problem, we apply EVC to cross-language VC. First, we train an eigenvoice GMM (EV-GMM) using many parallel data sets by a source speaker and many pre-stored other speakers who can utter the same language as the source speaker. And then, the conversion model between the source speaker and a target speaker who cannot utter the source speaker's language is developed by adapting the EV-GMM using a few arbitrary sentences uttered by the target speaker in a different language. The experimental results demonstrate that the proposed method yields significant performance improvements in both speech quality and conversion accuracy for speaker individuality compared with a conventional cross-language VC method based on frame selection.

Index Terms: speech synthesis, voice conversion, cross-language, eigenvoice conversion, unsupervised adaptation

1. Introduction

Voice conversion (VC) is a technique to modify voices of a given speaker, called the source speaker, so that they sound like those of another speaker, called the target speaker. A conversion method based on Gaussian mixture model (GMM) proposed by Stylianou *et al.* [1] is one of the most popular statistical approaches to VC. 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 the source and target voices [2]. The trained GMM allows the conversion from the source into the target based on minimum mean square error [1] or maximum likelihood criterion [3]. Although this method is very effective, this training framework is difficult to use if parallel data are not available.

One of promising VC applications is cross-language VC [4] of which the goal is to preserve voice quality of the speaker when synthesizing another language. This technique is very effective for a speech translation system, a language training system, and so on. Abe *et al.* [4] proposed a cross-language VC method between a Japanese speaker and an English synthesizer. To synthesize English speech as if uttered by the Japanese speaker, parallel data are created by synthesizing speech samples sounding like Japanese with the English speech synthesizer, and then the mapping function is trained using those pseudo-parallel data. Mashimo *et al.* [5] proposed another method based on a source bilingual speaker. The conversion model is trained using parallel data in the target speaker's language. And then, the trained conversion model is straightforwardly adopted for converting speech uttered in the other language by the source speaker as if uttered by the target speaker. These methods are not enough convenient to be applied to any language-pair or any speaker-pair.

Several attempts at training the conversion model using non-parallel data have been proposed for making the VC training framework more flexible. One typical approach is to create pseudo-parallel data from non-parallel data. Suendermann *et al.* [6] proposed a text-independent training method based on frame selection. This method adopts unit selection [7] to find corresponding time frames in the source and target speech data. Erro and Moreno [8] applied the frame selection method to cross-language VC and reported that this method is very effective for cross-language VC.

Another promising approach is to use the model adaptation techniques. Mouchtaris *et al.* [9] proposed a non-parallel training method based on maximum likelihood constrained adaptation of a GMM trained with an existing parallel data set of a different speaker-pair. Inspired by this method, Toda *et al.* [10] proposed eigenvoice conversion by integrating an eigenvoice technique [11] to the GMM-based VC framework. One-to-many EVC, which is one of main frameworks of EVC, allows the conversion from one pre-defined source speaker's voice into an arbitrary target speaker's voice. In the training process, multiple parallel data sets consisting of utterance pairs of the source speaker and many pre-stored target speakers are used for training an eigenvoice GMM (EV-GMM). In the adaptation process, the EV-GMM is adapted to arbitrary target speakers using only their speech data without any linguistic restrictions. One of the main advantages of EVC is to exploit voices of many other speakers as prior information to develop the conversion model for the adapted target speaker, and thus a significant reduction of the amount of adaptation data is allowed.

In this paper, we apply one-to-many EVC to cross-language VC. We assume that a large amount of speech data of the source speaker is previously available but quick and flexible development of the conversion model for various target speakers is desired. This situation would be observed in some applications, e.g., converting output speech of a foreign language speech synthesizer as if uttered by an arbitrary input speaker in a speech translation system or converting a teacher's voice into an arbitrary student's voice in language training. In such a situation, the EV-GMM can be trained in advance by using multiple parallel data sets consisting of utterance pairs of the source speaker and many pre-stored other speakers speaking the same language as the source speaker because such parallel data sets could be generally available. And then, the EV-GMM is adapted to the target speaker who cannot speak the source speaker's language using a small amount of speech data uttered by the target speaker in a different language. We conduct experimental evaluations in cross-language VC from a Japanese male speaker

Copyright © 2009 ISCA

into a French female speaker. The experimental results demonstrate that the proposed method yields significant improvements in both converted speech quality and conversion accuracy for speaker individuality compared with the conventional method based on frame selection.

2. Conventional method based on frame selection

We apply the conventional frame selection method to our cross-language VC task from the source speaker, of which a large amount of speech data is available, to the target speaker, of which a relatively small amount of speech data is available. First we create parallel data, of which the size is equal to the size of the source speech data, by selecting frames in the target speech data to be aligned to individual frames in the source speech data [6]. And then, a GMM of the joint probability density of source and target features is trained. In this paper, we adopt the trajectory-based conversion method considering dynamic features and the global variance (GV) based on maximum likelihood criterion [3].

2.1. Selection process

We use the Viterbi algorithm to select a sequence of frames $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T\}$ from the target database so that it best matches a sequence of source frames $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$.

First all of the target frames are divided into C clusters with K-means algorithm. And then pre-selection is performed for each source frame $\boldsymbol{x}_t$ by selecting the cluster with the closest centroid as follows:

$$\hat{c}_t = \underset{c}{\mathrm{argmin}} \sum_{d=1}^{D} w^{(d)} \left(x_t(d) - y_c(d)\right)^2 \quad (1)$$

where D is the dimension of the feature vector, $x_t(d)$ is the d^{th} component of the source feature vector at frame t, $y_c(d)$ is the d^{th} component of the c^{th} centroid of the target, and $w^{(d)}$ is the weighting factor associated with these $d^{(th)}$ components. The target frames $(\boldsymbol{y}_1^{(\hat{c}_t)}, \cdots, \boldsymbol{y}_N^{(\hat{c}_t)})$ belonging to the selected cluster $\hat{c}_t$ are used as candidates to be selected for frame t.

The selection process is performed by minimizing a global distance, which is defined as the weighted sum of two types of distance. One is the target distance capturing the quality degradation caused by the difference between the selection target $\boldsymbol{x}_t$ and a candidate $\boldsymbol{y}_n^{(\hat{c}_t)}$, which is given by

$$\mathcal{D}(\boldsymbol{x}_t, \boldsymbol{y}_n^{(\hat{c}_t)}) = \sum_{d=1}^{D} w^{(d)} \left(x_t(d) - y_n^{(\hat{c}_t)}(d)\right)^2. \quad (2)$$

The other is the concatenation distance capturing the quality degradation caused by concatenating two candidate frames $\boldsymbol{y}_m^{(\hat{c}_t-1)}$ and $\boldsymbol{y}_n^{(\hat{c}_t)}$, which is given by

$$\mathcal{D}(\boldsymbol{y}_m^{(\hat{c}_{t-1})}, \boldsymbol{y}_n^{(\hat{c}_t)}) = \sum_{d=1}^{D} w^{(d)} \left(y_m^{(\hat{c}_{t-1})}(d) - y_n^{(\hat{c}_t)}(d)\right)^2. \quad (3)$$

2.2. Training process

From the time-aligned feature vectors we generate $2D$-dimensional acoustic feature vectors $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ (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 feature vectors, respectively, where $\top$ denotes transposition of the vector. The joint probability density of the source and target feature vectors is modeled by a GMM [2] as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right) \quad (4)$$

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (5)$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The mixture component index is m. The total number of mixture components is M. A parameter set of the GMM is $\boldsymbol{\lambda}$, which consists of weights α_m, mean vectors $\boldsymbol{\mu}_m^{(X,Y)}$ and covariance matrices $\boldsymbol{\Sigma}_m^{(X,Y)}$ for individual mixture components. This paper employs diagonal covariance matrices for the individual block covariance matrices in Eq. (5).

The probability density of the GV of the output static feature vectors over an utterance is also modeled by a Gaussian distribution,

$$P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) = \mathcal{N}(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(v)}) \quad (6)$$

where the GV $\boldsymbol{v}(\boldsymbol{y}) = [v(1), \cdots, v(D)]^\top$ is calculated by

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right)^2. \quad (7)$$

A parameter set $\boldsymbol{\lambda}^{(v)}$ consists of a mean vector $\boldsymbol{\mu}^{(v)}$ and a diagonal covariance matrix $\boldsymbol{\Sigma}^{(v)}$.

2.3. Conversion process

Let $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ be a time sequence of the source feature vectors and that of the target feature vectors, respectively. The converted static feature vector sequence $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ is determined by maximizing a product of the conditional probability density of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and the GV probability density under a constraint $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\mathrm{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})^{\omega} P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) \quad (8)$$

where $\boldsymbol{W}$ is a window matrix to extend the static feature vector sequence to the feature vector sequence consisting of static and dynamic features. A balance between $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ and $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$ is controlled by the weight ω ($= 1/2T$ in this paper).

3. Proposed method based on eigenvoice conversion (EVC)

We apply one-to-many EVC to our cross-language VC task.

3.1. Eigenvoice GMM (EV-GMM)

The joint probability density of the source and target feature vectors is modeled by the EV-GMM as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w}), \boldsymbol{\Sigma}_m^{(X,Y)}\right) \quad (9)$$

where the mean vector $\boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w})$ is written as

$$\boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w}) = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(\boldsymbol{w}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{B}_m^{(Y)}\boldsymbol{w} + \boldsymbol{b}_m^{(Y)}(0) \end{bmatrix} \quad (10)$$

In one-to-many EVC, the target mean vector of the m^{th} mixture component is represented as a linear combination of a bias vector $\boldsymbol{b}_m^{(Y)}(0)$ and representative vectors $\boldsymbol{B}_m^{(Y)} = [\boldsymbol{b}_m^{(Y)}(1), \cdots, \boldsymbol{b}_m^{(Y)}(J)]$, where the number of representative vectors is J. The J-dimensional weight vector $\boldsymbol{w} = [w(1), \cdots, w(J)]^\top$ is adapted to arbitrary target speakers while the parameter set of the EV-GMM $\boldsymbol{\lambda}^{(EV)}$ is tied over different target speakers.

3.2. Training of EV-GMM

The tied parameter set of the EV-GMM is trained in advance using the multiple parallel data sets consisting of the single source speaker and many pre-stored target speakers. We employ speaker adaptive training (SAT) [12] to construct a *canonical model* causing significant improvements of the model adaptation performance [13]. Let $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t^{(s)}$ be the feature vector of the source speaker and that of the s^{th} pre-stored target speaker at frame t. SAT estimates not only the tied parameter set $\boldsymbol{\lambda}^{(EV)}$ but also a set of the weight vectors $\boldsymbol{w}_1^S = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_S\}$ adapted to individual pre-stored target speakers as follows:

$$\hat{\boldsymbol{\lambda}}^{(EV)}, \hat{\boldsymbol{w}}_1^S = \mathop{\arg\max}_{\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}_1^S} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}_s). \quad (11)$$

To enable maximum a posteriori (MAP) estimation in the adaptation process [14] described bellow, we also train the following Gaussian distribution for the weight vector:

$$P(\boldsymbol{w}|\boldsymbol{\lambda}^{(w)}, \tau) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}^{(w)}, \tau^{-1}\boldsymbol{\Sigma}^{(w)}) \quad (12)$$

where τ is a hyper-parameter. A model parameter set $\boldsymbol{\lambda}^{(w)}$ consisting of the mean vector $\boldsymbol{\mu}^{(w)}$ and the covariance matrix $\boldsymbol{\Sigma}^{(w)}$ is estimated using a set of the weight vectors estimated for individual pre-stored target speakers in SAT (see Eq. (11)) as follows:

$$\hat{\boldsymbol{\lambda}}^{(w)} = \mathop{\arg\max}_{\boldsymbol{\lambda}^{(w)}} \prod_{s=1}^{S} P(\hat{\boldsymbol{w}}_s | \boldsymbol{\lambda}^{(w)}, \tau = 1). \quad (13)$$

3.3. Unsupervised adaptation and conversion

The EV-GMM is adapted for an arbitrary target speaker by estimating the optimum weight vector for given speech samples in a completely unsupervised manner, i.e., using neither parallel data nor linguistic information. For a time sequence of the given target feature vectors $\boldsymbol{Y}'_1, \cdots, \boldsymbol{Y}'_T$, the MAP adaptation of the EV-GMM is performed as follows:

$$\begin{aligned} \hat{\boldsymbol{w}} &= \mathop{\arg\max}_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{Y}'_1, \cdots, \boldsymbol{Y}'_T, \boldsymbol{\lambda}) \\ &= \mathop{\arg\max}_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{\lambda}^{(w)}, \tau) \prod_{t=1}^{T} P(\boldsymbol{Y}'_t | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}). \quad (14) \end{aligned}$$

The conversion process is straightforwardly performed with the adapted EV-GMM in a manner described in **Section 2.3**. Note that the probability density of the GV is also modeled by eigenvoices and it is adapted to the arbitrary target speaker as described in [15].

4. Experimental evaluations

We conducted subjective evaluations in cross-language VC from a Japanese male speaker into a French female speaker.

4.1. Experimental conditions

In order to train one-to-many EV-GMM, we used 160 speakers consisting of 80 male and 80 female Japanese speakers in the Japanese Newspaper Article Sentences (JNAS) database as pre-stored target speakers. Each speaker uttered 50 phoneme-balanced sentences. The source Japanese speaker not included in JNAS uttered the same sentence sets as uttered by pre-stored target speakers. The total number of training sentences uttered by the source speaker was 350 (details in [16]).

We used 2, 8, and 32 sentences uttered by the target French speaker who cannot speak Japanese to adapt the EV-GMM. The number of mixture components was set to 128 and the number of representative vectors was set to 159. The hyper-parameter for the MAP estimation τ was set to a constant value manually determined in our preliminary experiment [14].

In the frame selection[1], we created parallel data sets consisting of 350 sentences by aligning 2, 8, and 32 sentences of the target French speaker, which were the same as used in the EV-GMM adaptation, to 350 Japanese sentences of the source Japanese speaker, which were the same as used in training of the EV-GMM. The number of mixture components was set to 128.

All speech data were sampled at 16 kHz. As a spectral parameter, we employed 1^{st} through 24^{th} mel-cepstral coefficients extracted by STRAIGHT analysis [17]. We converted the source Japanese speaker's voice into the target French speaker's voice. The number of Japanese test sentences was 21, which were not included in the training data. The source fundamental frequency F_0 was converted into target fundamental frequency $\hat{F}_0$ as follows:

$$\log \hat{F}_0 = \frac{\sigma^{(y)}}{\sigma^{(x)}} \left(\log F_0 - \mu^{(x)} \right) + \mu^{(y)} \quad (15)$$

where $\mu^{(x)}$ and $\sigma^{(x)}$ are mean and standard deviation of log-scaled F_0 of the source, and $\mu^{(y)}$ and $\sigma^{(y)}$ are those of the target, respectively.

We conducted an opinion test on speech quality and an XAB test on speaker similarity. In the opinion test, listeners evaluated speech quality of the converted speech samples using a 5-point-scale opinion score (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). In each trial of the XAB test, analysis-synthesized French speech of the target French speaker was presented as reference, and then a pair of two types of the converted speech generated by EVC and frame selection were presented in random order. Listeners were asked which voices sounded more similar to the reference in terms of speaker individuality. Ten Japanese listeners participated in each test.

4.2. Experimental results

Figure 1 shows the result of the opinion test on speech quality. One-to-many EVC significantly outperforms the frame selection. Even if using only two sentences of the target speaker, one-to-many EVC yields much better converted speech quality than the frame selection using 32 sentences of the target speaker. One-to-many EVC effectively exploits many other speakers'

[1] We used software available from http://www.enterface.net/enterface06/docs/results/sources/project4_sources.zip

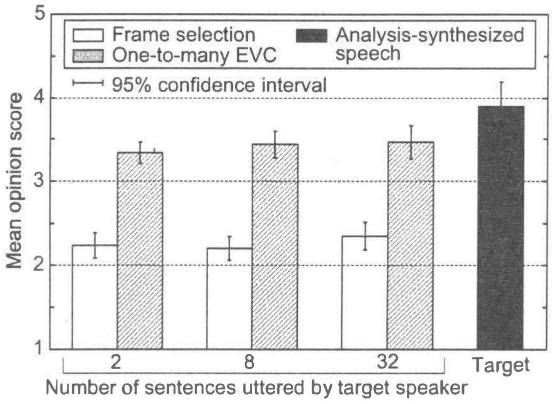

Figure 1: *Result of opinion test on speech quality.*

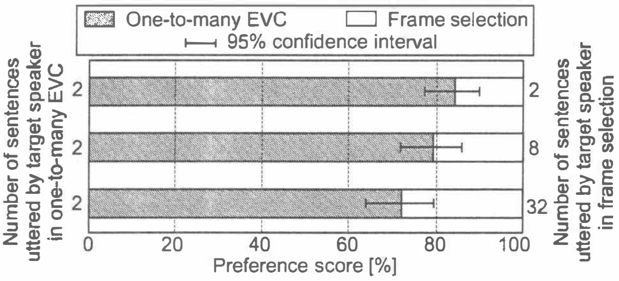

Figure 2: *Result of preference test on speaker individuality.*

voices to develop the conversion model for a target speaker even if their language is different from the target speaker's. In other words, several parameters of the conversion model are effectively shared between multiple speakers even if they utter different languages. These parameters are robustly estimated using many other speaker's voices in the EVC.

Figure 2 shows the result of the XAB test on speaker similarity. Note that only two sentences of the target French speaker were always used in the one-to-many EVC. We can observe that one-to-many EVC yields much better conversion accuracy for speaker individuality than the frame selection. The performance of frame selection is improved by increasing the amount of target speaker's data but it is still significantly inferior to that of one-to-many EVC using only two target sentences.

These results suggest that the proposed method is very effective in cross-language VC and significantly outperforms the conventional method based on frame selection.

5. Conclusions

This paper has described cross-language voice conversion (VC) based on eigenvoice conversion (EVC). An eigenvoice GMM (EV-GMM) is trained in advance using parallel data sets consisting of a source and many pre-stored other speakers in the same language. And then, the conversion model between the source speaker and an arbitrary target speaker uttering a different language is effectively developed by unsupervised adaptation of the EV-GMM using a very limited amount of target speech samples. The results of subjective evaluations have demonstrated that our proposed cross-language EVC significantly outperforms the conventional cross-language VC based on frame selection.

Acknowledgment: The authors are grateful to Prof. Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method.

6. References

[1] 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.

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

[3] T. Toda, A.W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[4] 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.

[5] 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, July 2002.

[6] D. Suendermann, H. Hoege, A. Bonafonte, H. Ney, A. W. Black, and S. Narayanan. Text-independent voice conversion based on unit selection. *Proc. ICASSP*, Vol. 1, pp. 81–84, Toulouse, France, USA, Mar. 2006.

[7] A.J. Hunt and A.W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. *Proc. ICASSP*, pp. 373–376, Atlanta, USA, May 1996.

[8] D. Erro and A. Moreno. Frame alignment method for cross-lingual voice conversion. *Proc. INTERSPEECH*, pp. 1969–1972, Antwerp, Belgium, Aug. 2007.

[9] 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.

[10] T. Toda, Y. Ohtani, and K. Shikano. One-to-many and many-to-one voice conversion based on eigenvoices. *Proc. ICASSP*, pp. 1249–1252, Hawaii, USA, Apr. 2007.

[11] 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.

[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] Y. Ohtani, T. Toda, H. Saruwatari, K. Shikano. Speaker adaptive training for one-to-many eigenvoice conversion based on Gaussian mixture model. *Proc. INTERSPEECH*, pp. 1981–1984, Antwerp, Belgium, Aug. 2007.

[14] D. Tani, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano. Maximum a posteriori adaptation for many-to-one eigenvoice conversion. *Proc. INTERSPEECH*, pp. 1461–1464, Brisbane, Australia, Sep. 2008.

[15] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. An improved one-to-many eigenvoice conversion system. *Proc. INTERSPEECH*, pp. 1080–1083, Brisbane, Australia, Sep. 2008.

[16] T. Toda, Y. Ohtani, and K. Shikano. Eigenvoice conversion based on Gaussian mixture model. *Proc. ICSLP*, pp. 2446–2449, Pittsburgh, USA, Sep. 2006.

[17] 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.