COST REDUCTION OF TRAINING MAPPING FUNCTION BASED ON MULTISTEP VOICE CONVERSION

Tsuyoshi Masuda and Makoto Shozakai

Speech Solutions, New Business Development, Asahi Kasei Corporation
Atsugi AXT Maintower 22F, 3050 Okata, Atsugi, Kanagawa 243-0021 JAPAN

ABSTRACT

Several approaches based on a statistical method for voice conversion from one speaker to another have been developed. In a statistical spectral mapping method which is a typical one in these approaches, a mapping function which represents a correlation between different speakers is determined using spectral features. This technique has the problem that it is necessary to train the mapping function for each speaker pair. The training cost must become a serious issue in case that the number of speakers increases significantly.

This paper describes a novel voice conversion method for reducing the training cost. This technique is easily implemented and can use conventional techniques directly. Experimental results demonstrate that the converted speech is almost maintaining the conventional quality despite the significant training cost reduction by the proposed method.

Index Terms— Voice conversion, speech synthesis, training cost, multistep voice conversion

1. INTRODUCTION

Voice conversion is expected to be applied to various applications because of its flexibility. These applications includes, for instance, a personalization of a text-to-speech synthesis system [1], cellular applications, a speech enhancement for telecommunications [2] [3] and handsfree systems, and a cross-language speaker conversion[4], and so on.

As a typical voice conversion method, statistical spectral mapping has been proposed. An early attempt was based on a vector quantization approach by Abe et al. [5]. This method realizes discrete codebook mapping based on hard clustering. Using a probabilistic approach, Stylianou et al. [6] have proposed a mapping method based on a Gaussian Mixture Model (GMM). Main concepts of the method are soft clustering and continuous transformation. In 2005, the performance of GMM-based conversion technique has significantly improved by using maximum likelihood estimation (MLE) considering dynamic features and global variance by Toda et al. [7].

Common to these techniques [5] [6] [7] is that these have a training procedure for creating a mapping function which indicates a correlation between different speakers. The training procedure has the potential problem that it is necessary to train the mapping function for each speaker pair. It is imaginable that the training cost becomes a significant problem when a lot of people use voice conversion in cellular applications. To reduce the training cost, Mouchtaris et al. [8] have proposed a non-parallel training method based on maximum likelihood constrained adaptation. Although it relaxes a constraint that we have to use a parallel corpus in training, the number of training (or adaptation) does not decrease.

In order to reduce the training cost when the number of speakers is greatly large, this paper describes a novel method based on Multistep Voice Conversion (MVC). This algorithm is easy to be implemented and can use conventional techniques directly , e.g. [5] [6] [7]. This paper's purposes are to reduce the training cost, i.e. to decrease the number of training, and to maintain the quality of converted speech by MVC equivalent to the one by conventional method. Experimental results demonstrate that this technique achieves a performance equivalent of converted speech quality by the conventional technique in spite of the significant training cost reduction.

This paper is organized as follows. In **Section 2**, a framework of MVC is described. In **Section 3**, experimental evaluations are described. Finally, conclusions are provided in **Section 4**.

2. MULTISTEP VOICE CONVERSION

Figure 1 shows a schematic diagram of conventional voice conversion, where $F(Src, Tgt)$ denotes a mapping function to convert Src into Tgt. Src and Tgt are source and target speakers, respectively. The number of mapping functions is $N \times M$ when there are N source speakers and M target speakers in conventional method.

A schematic diagram of MVC is shown in Fig. 2, where $G(Src, Tgt)$ denotes a mapping function to convert an intermediate speaker Int into Tgt. After Src's features are converted into Int's ones using $F(Src, Int)$, $G(Src, Tgt)$ is estimated by these converted ones and Tgt's ones. The number of mapping functions is only $N + M$ under the same condition.

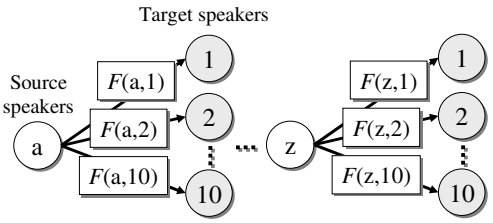

Fig. 1. Schematic diagram of conventional voice conversion.

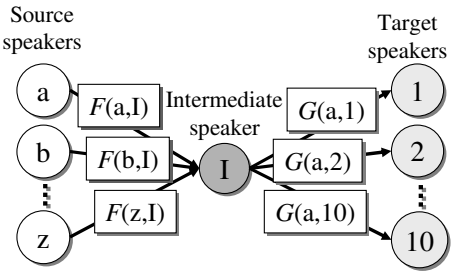

Fig. 2. Schematic diagram of MVC. In this case, a representative source speaker is "**a**" for obtaining $G(\ )$s.

In the MVC method, the source speaker is converted into the target speaker via the intermediate speaker. In order to reduce the training cost, mapping functions which convert intermediate into targets are shared between different source speakers. Therefore, this technique requires only one training of $F(\ )$ for the source speaker who is not a representative source speaker "**a**" in Fig. 2. The more the number of source or target speakers increases, the less the training cost is required in comparison with the conventional method.

2.1. Training process

In the MVC method, two kinds of training methods are adopted. One is the conversion dependent training, and the other is the conversion independent training. Figure 3 shows the training process of the MVC method, to obtain $G(\ )$, which consists of the following steps:

Step 1 : Select a representative source speaker S.

Step 2 : $F(\mathrm{S},\mathrm{I})$ is trained by features of S and features of an intermediate speaker I.

Step 3 : The same features of S in Step 2 are converted using $F(\mathrm{S},\mathrm{I})$.

Step 4 : $G(\mathrm{S},\mathrm{T})$ is trained by the converted features in Step 3 and features of a target speaker T.

Step 4 is performed when $G(\ )$s for converting to other target speakers are required. The training is called "conversion dependent training". On the other hand, we can also train $G(\ )$

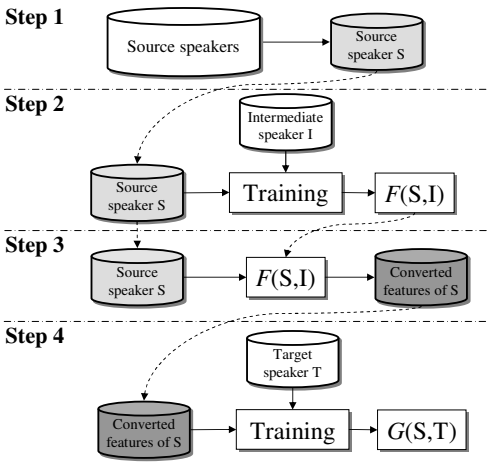

Fig. 3. Training process of MVC to obtain $G(\ )$. It indicates "conversion dependent training".

independent from the source speaker using features of S and T. In this case, $G(\mathrm{S},\mathrm{T})$ replaces $F(\mathrm{I},\mathrm{T})$ which converts I to T. This is called "conversion independent training". In the present work, "conversion dependent training" is employed considering the following conversion process.

2.2. Conversion process

In conversion process of the MVC method, firstly, source features are converted into intermediate ones using $F(\ )$. After that, these are converted to target ones using $G(\ )$. Amount of calculation shall be doubled because this technique needs to convert features twice. However, by introducing a composite mapping function that consists of a combination of $F(\ )$ and $G(\ )$, it is possible to achieve the same amount of calculation as that of the conventional method.

2.3. Spectral conversion

It is widely recognized that the GMM-based mapping method [6] is the most popular one. The GMM-based technique is employed as a spectral mapping method in the present work. In this technique, a mapping function from a source feature vector $\boldsymbol{x}_k$ to a target feature vector $\boldsymbol{y}_k$ in frame k is defined as

$$\hat{\boldsymbol{y}}_k = \sum_{i=1}^{M} p(m_i|\boldsymbol{x}_k,\boldsymbol{\lambda})\boldsymbol{E}(\boldsymbol{y}_k|\boldsymbol{x}_k,m_i,\boldsymbol{\lambda}), \quad (1)$$

$$\boldsymbol{E}(\boldsymbol{y}_k|\boldsymbol{x}_k,m_i,\boldsymbol{\lambda}) = \boldsymbol{\mu}_i^{(y)} + \boldsymbol{\Sigma}_i^{(yx)}{\boldsymbol{\Sigma}_i^{(xx)}}^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}_i^{(x)}), \quad (2)$$

where $\hat{\boldsymbol{y}}_k$ denotes an estimated spectral feature vector. The i^{th} mixture has a weight w_i, mean vectors $\boldsymbol{\mu}_i^{(x)}$ and $\boldsymbol{\mu}_i^{(y)}$, and covariance matrices $\boldsymbol{\Sigma}_i^{(xx)}$ and $\boldsymbol{\Sigma}_i^{(yx)}$. The total number of mixtures is M. The multivariate normal distribution with $\boldsymbol{\mu}_i^{(x)}$

and $\mathbf{\Sigma}_i^{(xx)}$ is represented as $N(\boldsymbol{x}_k; \boldsymbol{\mu}_i^{(x)}, \mathbf{\Sigma}_i^{(xx)})$. $\boldsymbol{\lambda}$ denotes a set of model parameters, i.e. weights, mean vectors and covariance matrices. The conditional probabilities $p(m_i|\boldsymbol{x}_k, \boldsymbol{\lambda})$ are given from

$$p(m_i|\boldsymbol{x}_k, \boldsymbol{\lambda}) = \frac{w_i N(\boldsymbol{x}_k; \boldsymbol{\mu}_i^{(x)}, \mathbf{\Sigma}_i^{(xx)})}{\sum_{j=1}^{M} w_j N(\boldsymbol{x}_k; \boldsymbol{\mu}_j^{(x)}, \mathbf{\Sigma}_j^{(xx)})}. \quad (3)$$

Kain and Macon [9] have proposed a training method for estimating model parameters based on joint density estimation. In this method, the following GMM on joint vector $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ is trained in advance with training data consisting of time-aligned features determined by Dynamic Time Warping (DTW).

$$p(\boldsymbol{z}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} w_i N(\boldsymbol{z}; \boldsymbol{\mu}_i^{(z)}, \mathbf{\Sigma}_i^{(z)}), \quad (4)$$

$$\mathbf{\Sigma}_i^{(z)} = \begin{bmatrix} \mathbf{\Sigma}_i^{(xx)} & \mathbf{\Sigma}_i^{(xy)} \\ \mathbf{\Sigma}_i^{(yx)} & \mathbf{\Sigma}_i^{(yy)} \end{bmatrix}, \quad \boldsymbol{\mu}_i^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(x)} \\ \boldsymbol{\mu}_i^{(y)} \end{bmatrix}. \quad (5)$$

All the parameters of the GMM are estimated using the EM algorithm. The covariance matrices $\mathbf{\Sigma}_i^{(xx)}$, $\mathbf{\Sigma}_i^{(xy)}$, $\mathbf{\Sigma}_i^{(yx)}$ and $\mathbf{\Sigma}_i^{(yy)}$ are diagonal.

3. EXPERIMENTS

3.1. Experimental conditions

A set of 50 sentences which is phonetically balanced by Japanese speakers is used as a training set, and another set of 50 sentences which does not include these training sentences is used for evaluation. One male and one female are used as a source speaker and a target speaker, respectively. As intermediate speakers, two males (male1, male2) and two females (female1, female2) are used, respectively. The experimental sentences are converted into all gender pairs: male-to-male, male-to-female, female-to-male and female-to-female. The first through 41^{st} cepstral coefficients extracted by the STRAIGHT analysis method [10] from 16 kHz sampling speech data are used as spectral features. The shift length is 5 ms and the number of mixtures is 64.

To objectively gauge spectral conversion performance by intermediate speakers, the cepstral distortion (CD) given by the following equation is employed,

$$\mathrm{CD} = \frac{20}{\ln 10} \cdot \sqrt{2 \sum_{d=1}^{41} \left(c_d^{(t)} - c_d^{(e)}\right)^2}, \quad (6)$$

where $c_d^{(t)}$ and $c_d^{(e)}$ denote the d^{th} coefficient of the target and the estimated cepstra, respectively.

To subjectively investigate spectral conversion performance on a degradation of naturalness, degradation mean opinion score (DMOS) test is conducted. In order to measure only the performance of the spectral conversion, the converted speech is synthesized using natural prosodic features automatically extracted from the target speech as follows. A time-alignment for modifying duration is performed with DTW, and then at each frame, F_0 and total power of a converted linear spectrum are set to each target value. As post-processing processes, post-filtering and moving average method for frame smoothing are applied to the converted linear spectrum. The STRAIGHT synthesis method [10] is employed as a speech synthesizer.

In our evaluations, we compare the following methods:

- **conventional**: The conventional method.
- **mvc_own**: The MVC method using source speaker's own $G(\ )$, e.g. using the mapping function G(A, B) when the source speaker A is converted into the target speaker B.
- **mvc_oth**: The MVC method using other source speaker's $G(\ )$, e.g. using the mapping function G(C, B) which has been trained from the source speaker C when the source speaker A is converted into the target speaker B.

3.2. Objective evaluation

Figure 4 shows the CD on different methods, where **unconv** shows the result between target and source cepstra without any conversion. It is observed that the spectral conversion accuracy of **mvc_own** is slightly deteriorated. This is because **mvc_own** conducted conversion twice compared with **conventional**. Slight degradation of **mvc_oth** compared with **mvc_own**, which is caused by using a mapping function $G(\ )$ which includes a feature mapping between another speakers-pair. However, the objective experiment result doesn't indicate remarkable degradation although the training cost is reduced. In Fig. 5, the result of the objective test as to different intermediate speakers is given, which demonstrates almost comparable performance for all intermediate speakers.

3.3. Subjective evaluation

The DMOS test is performed on the degradation of naturalness. In the DMOS test, an opinion score is set to a 5-point scale (5: imperceptible, 4: perceptible, but not annoying, 3: slightly annoying, 2: annoying, 1: very annoying). An original target speech is presented as reference speech. From evaluation sentences, 3 sentences of each pair are selected. The number of listeners is five. In this evaluation, female2 is used as the representative intermediate speaker among the objective evaluation.

Figure 6 shows the result of the DMOS test. **ana_syn** shows the result for analysis-by-synthesized target speech using the 0^{th} through 41^{st} cepstral coefficients by the STRAIGHT

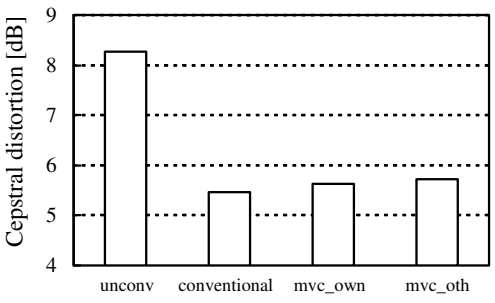

Fig. 4. Result of the objective evaluation for different methods. **mvc_own** and **mvc_oth** show average of cepstral distortion of all case of intermediate speakers.

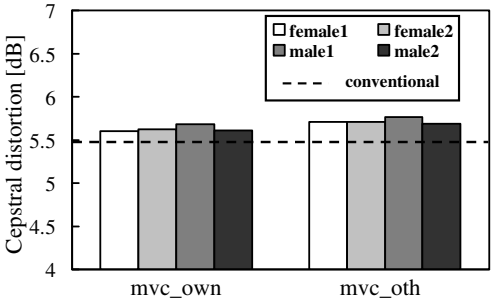

Fig. 5. Result of the objective evaluation by intermediate speaker. The dashed line corresponds to conventional result in Fig. 4.

analysis method. The result has a tendency similar to the objective experiment. We can see that the proposed technique achieves satisfactory performance on speech quality as the conventional technique. We expect that the performance will be improved more by applying MLE-based approach [7].

4. CONCLUSION

We have proposed the new method in order to reduce the training cost of mapping function for voice conversion based on Multistep Voice Conversion (MVC). In MVC, a source speaker is transformed to a target speaker via an intermediate speaker. This technique has the following advantages: 1) it is easy to implement, and 2) we can introduce some conventional techniques directly. Experimental results based on spectral distortion measure and perceptual evaluation have demonstrated that the converted speech is maintaining the quality of the conventional converted speech despite the significant training cost reduction.

5. REFERENCES

[1] M. Eichner, M. Wolff, and R. Hoffmann, "Voice characteristics conversion for TTS using reverse VTLN," *Proc. ICASSP*, vol. 1, pp. 17–20, May 2004.

[2] A. Mouchtaris, J.V. der Spiegel, and P. Mueller, "A spectral conversion approach to feature denoising and speech enhancement," *Proc. INTERSPEECH*, pp. 2057–2060, 2005.

[3] K. Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," *Proc. ICASSP*, pp. 1843–1846, June 2000.

[4] 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.

[5] 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.

[6] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 132–142, 1998.

[7] 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, March 2005.

[8] A. Mouchtaris, J.V. der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 952–963, 2006.

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

[10] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

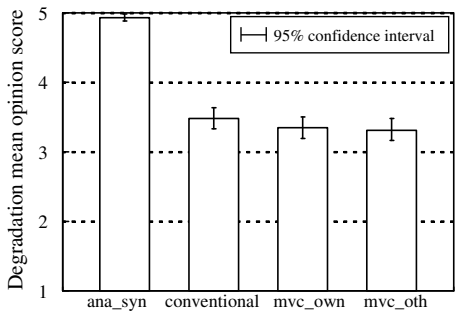

Fig. 6. Result of the subjective evaluation on the naturalness degradation.