A Study of Mutual Information for GMM-Based Spectral Conversion

Hsin-Te Hwang[1], *Yu Tsao*[2], *Hsin-Min Wang*[3], *Yih-Ru Wang*[1], *Sin-Horng Chen*[1]

[1]Dept. of Electrical Engineering, National Chiao Tung University, Hsinchu, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[3]Institute of Information Science, Academia Sinica, Taipei, Taiwan
hwanght@iis.sinica.edu.tw, yu.tsao@citi.sinica.edu.tw, whm@iis.sinica.edu.tw,
yrwang@cc.nctu.edu.tw, schen@mail.nctu.edu.tw

Abstract

The Gaussian mixture model (GMM)-based method has dominated the field of voice conversion (VC) for last decade. However, the converted spectra are excessively smoothed and thus produce muffled converted sound. In this study, we improve the speech quality by enhancing the dependency between the source (natural sound) and converted feature vectors (converted sound). It is believed that enhancing this dependency can make the converted sound closer to the natural sound. To this end, we propose an integrated maximum a posteriori and mutual information (MAPMI) criterion for parameter generation on spectral conversion. Experimental results demonstrate that the quality of converted speech by the proposed MAPMI method outperforms that by the conventional method in terms of formal listening test.

Index Terms: Voice conversion, mutual information, GMM.

1. Introduction

The task of voice conversion (VC) is to transform a source voice to a specific target voice [1-6]. Generally, the VC operation can be divided into two parts—spectral conversion and prosody conversion. In this study, we focus our discussion on spectral conversion. Many spectral conversion algorithms have been proposed in recent years. Among them, the Gaussian mixture model (GMM)-based method is a successful one [1-6]. The GMM-based method first prepares a GMM set to characterize the mapping from the source to target voices. Then, the conversion of spectral parameters is performed online in a frame by frame basis based on the GMM set. Although the GMM-based method has been proven effective, there are some issues to overcome.

One of the issues is the time independent mapping. Because the GMM-based method conducts conversion frame by frame, the discontinuity in the converted feature sequence usually occurs. To handle this issue, Toda *et al*. [3] proposed to combine dynamic features with static features to train a joint density GMM (JDGMM). Then, the maximum likelihood parameter generation algorithm (MLPG) (also called the MAP-based mapping in [4]) is employed to generate the converted spectrum sequence. Another issue is the over-smoothing effect. It is observed that the converted spectra by the GMM-based method are excessively smoothed, thereby degrading the quality of the converted sound. Several methods have been proposed to handle this over-smoothing issue. One notable method tackles this problem by incorporating the global variance (GV) into the conversion process [3-5]. The quality of the converted voice is enhanced by refining the converted feature sequence to match the GV of the target speaker. In this study, we propose a new criterion to handle the over-smoothing issue from another viewpoint.

A previous study reported that the true mapping from source to target voice features is complex and nonlinear [6]. However, the GMM-based method performs VC by simplifying the mapping with a linear transformation. It is believed that this simplification reduces the correlation of the converted sound and the natural sound characteristics from the source voice and accordingly produces muffled converted sound. Here, we study to improve the correlation (dependency) of the converted sound and the natural sound characteristics by incorporating the maximum mutual information (MMI) criterion in the GMM-based VC framework. Since the proposed parameter generation algorithm is performed based on a combination of MAP and MMI, we call the integrated method the MAPMI-mapping method.

The remainder of this paper is organized as follows. Section 2 reviews the conventional GMM-based spectral conversion and discusses the existing issues in GMM-based method. Section 3 describes the proposed MAPMI-mapping method. Section 4 presents our experimental setup and result analysis. Finally, our conclusion is given in section 5.

2. GMM-based spectral conversion considering the dynamic features

2.1. Training a JDGMM

In a typical VC framework, a parallel speech corpus must be prepared beforehand for training a conversion function. Let $\mathbf{X}=\left[\mathbf{X}_1^{\mathrm{T}},\cdots,\mathbf{X}_t^{\mathrm{T}},\cdots,\mathbf{X}_T^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\mathbf{Y}=\left[\mathbf{Y}_1^{\mathrm{T}},\cdots,\mathbf{Y}_t^{\mathrm{T}},\cdots,\mathbf{Y}_T^{\mathrm{T}}\right]^{\mathrm{T}}$ be the source and target feature vectors, respectively both consist of T feature frames; $\mathbf{X}_t=[\mathbf{x}_{S_t}^{\mathrm{T}},\mathbf{x}_{D_t}^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{Y}_t=[\mathbf{y}_{S_t}^{\mathrm{T}},\mathbf{y}_{D_t}^{\mathrm{T}}]^{\mathrm{T}}$ are the 2D-dimensional source and target feature vectors consisting of D static and D dynamic feature vectors at frame t. The superscript T denotes the vector transposition. A JDGMM is employed to model the joint feature vector $\mathbf{Z}_t=[\mathbf{X}_t^{\mathrm{T}},\mathbf{Y}_t^{\mathrm{T}}]^{\mathrm{T}}$ as

$$P(\mathbf{Z}_t\mid\Theta^{(\mathbf{Z})})=\sum_{m=1}^{M}\omega_m N(\mathbf{Z}_t;\boldsymbol{\mu}_m^{(\mathbf{Z})},\boldsymbol{\Sigma}_m^{(\mathbf{Z})}) \tag{1}$$

where ω_m is the prior probability of the mth mixture component with subject to $\sum_{m=1}^{M}\omega_m=1$; $\boldsymbol{\mu}_m^{(\mathbf{Z})}=[(\boldsymbol{\mu}_m^{(\mathbf{X})})^{\mathrm{T}},(\boldsymbol{\mu}_m^{(\mathbf{Y})})^{\mathrm{T}}]^{\mathrm{T}}$ and $\boldsymbol{\Sigma}_m^{(\mathbf{Z})}=\begin{bmatrix}\boldsymbol{\Sigma}_m^{(\mathbf{XX})} & \boldsymbol{\Sigma}_m^{(\mathbf{XY})}\\ \boldsymbol{\Sigma}_m^{(\mathbf{YX})} & \boldsymbol{\Sigma}_m^{(\mathbf{YY})}\end{bmatrix}$ are the mean vector and covariance matrix of the mth mixture component. The covariance matrices $\boldsymbol{\Sigma}_m^{(\mathbf{XX})}$, $\boldsymbol{\Sigma}_m^{(\mathbf{XY})}$, $\boldsymbol{\Sigma}_m^{(\mathbf{YX})}$, and $\boldsymbol{\Sigma}_m^{(\mathbf{YY})}$ are usually diagonal. The parameter set $\Theta^{(\mathbf{Z})}=\{\omega_m,\boldsymbol{\mu}_m^{(\mathbf{Z})},\boldsymbol{\Sigma}_m^{(\mathbf{Z})}\}_{m=1,2,\cdots,M}$ can be estimated via the expectation maximization (EM) algorithm.

2.2. Estimating the conditional probability density function

Based on JDGMM, the conditional probability density function (PDF), $P(\mathbf{Y}_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})})$, can be represented by:

$$P(\mathbf{Y}_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) = \sum_{m=1}^{M} P(m \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid \mathbf{X}_t, m, \Theta^{(\mathbf{Z})}) \quad (2)$$

where

$$P(m \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) = \frac{\omega_m N(\mathbf{X}_t; \boldsymbol{\mu}_m^{(\mathbf{X})}, \boldsymbol{\Sigma}_m^{(\mathbf{XX})})}{\sum_{l=1}^{M} \omega_l N(\mathbf{X}_t; \boldsymbol{\mu}_l^{(\mathbf{X})}, \boldsymbol{\Sigma}_l^{(\mathbf{XX})})} \quad (3)$$

$$P(\mathbf{Y}_t \mid \mathbf{X}_t, m, \Theta^{(\mathbf{Z})}) = N(\mathbf{Y}_t; \boldsymbol{\mu}_{m,t}^{(\mathbf{Y}|\mathbf{X})}, \boldsymbol{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})}). \quad (4)$$

The mean vector and the covariance matrix of the conditional PDF, $P(\mathbf{Y}_t \mid \mathbf{X}_t, m, \Theta^{(\mathbf{Z})})$, are given as

$$\boldsymbol{\mu}_{m,t}^{(\mathbf{Y}|\mathbf{X})} = \boldsymbol{\mu}_m^{(\mathbf{Y})} + \boldsymbol{\Sigma}_m^{(\mathbf{YX})} \boldsymbol{\Sigma}_m^{(\mathbf{XX})-1} (\mathbf{X}_t - \boldsymbol{\mu}_m^{(\mathbf{X})}) \quad (5)$$

$$\boldsymbol{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})} = \boldsymbol{\Sigma}_m^{(\mathbf{YY})} - \boldsymbol{\Sigma}_m^{(\mathbf{YX})} \boldsymbol{\Sigma}_m^{(\mathbf{XX})-1} \boldsymbol{\Sigma}_m^{(\mathbf{XY})}. \quad (6)$$

2.3. MAP-based mapping

When using the MAP criterion considering the dynamic features for mapping (also called the ML-based mapping in [3]), the converted static feature sequence $\hat{\mathbf{y}}_S$ is obtained as

$$\hat{\mathbf{y}}_S = \arg\max P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})}) \quad \text{s.t. } \mathbf{Y} = \mathbf{W}\mathbf{y}_S \quad (7)$$

where $\mathbf{W}$ is the $2DT$-by-DT weighting matrix (given in [3]) for calculating the joint static and dynamic features.

In practical implementation, the conditional PDF, $P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})})$ in (7), can be approximated with a single sequence of mixture components

$$P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})}) \approx P(\hat{\mathbf{m}} \mid \mathbf{X}, \Theta^{(\mathbf{Z})}) P(\mathbf{Y} \mid \mathbf{X}, \hat{\mathbf{m}}, \Theta^{(\mathbf{Z})}) \quad (8)$$

where the mixture component sequence $\hat{\mathbf{m}}$ is determined by $\hat{\mathbf{m}} = \arg\max P(\mathbf{m} \mid \mathbf{X}, \Theta^{(\mathbf{Z})})$. Finally, the converted static feature sequence can be obtained by

$$\hat{\mathbf{y}}_S = (\mathbf{W}^{\mathrm{T}} \mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1} \mathbf{W})^{-1} \mathbf{W}^{\mathrm{T}} \mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1} \mathbf{E}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})} \quad (9)$$

where

$$\mathbf{E}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})} = [(\boldsymbol{\mu}_{\hat{m}_1,1}^{(\mathbf{Y}|\mathbf{X})})^{\mathrm{T}}, \cdots, (\boldsymbol{\mu}_{\hat{m}_t,t}^{(\mathbf{Y}|\mathbf{X})})^{\mathrm{T}}, \cdots, (\boldsymbol{\mu}_{\hat{m}_T,T}^{(\mathbf{Y}|\mathbf{X})})^{\mathrm{T}}]^{\mathrm{T}}$$
$$\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1} = \mathrm{diag}[\boldsymbol{\Sigma}_{\hat{m}_1}^{(\mathbf{Y}|\mathbf{X})-1}, \cdots, \boldsymbol{\Sigma}_{\hat{m}_t}^{(\mathbf{Y}|\mathbf{X})-1}, \cdots, \boldsymbol{\Sigma}_{\hat{m}_T}^{(\mathbf{Y}|\mathbf{X})-1}]. \quad (10)$$

In (10), $\boldsymbol{\mu}_{\hat{m}_t,t}^{(\mathbf{Y}|\mathbf{X})}$ and $\boldsymbol{\Sigma}_{\hat{m}_t}^{(\mathbf{Y}|\mathbf{X})-1}$, can be calculated by (5) and (6), respectively, with determined $\hat{\mathbf{m}}$, where $\hat{\mathbf{m}} = [\hat{m}_1, \cdots, \hat{m}_t, \cdots, \hat{m}_T]$.

2.4. Problem in GMM-based method

A major problem to the GMM-based method is that the converted spectra are excessively smoothed, thereby producing muffled converted sounds. In [6], the authors explained that the over-smoothing issue comes from the true mapping from the source to the target speech features is complex and nonlinear, and the GMM-based method may have limitation to characterize the true mapping exactly. Thus, in (5), the correlation item $\boldsymbol{\Sigma}_m^{(\mathbf{YX})} \boldsymbol{\Sigma}_m^{(\mathbf{XX})-1}$ can be very small, which causes $\boldsymbol{\mu}_{m,t}^{(\mathbf{Y}|\mathbf{X})} \approx \boldsymbol{\mu}_m^{(\mathbf{Y})}$. This can make the converted feature sequence converge to the weighted mean of the target mixture components and the converted voice sound muffled for the MMSE-based mapping [2].

The same problem exists in the MAP-based mapping. We often observe that $\boldsymbol{\mu}_{\hat{m}_t,t}^{(\mathbf{Y}|\mathbf{X})} \approx \boldsymbol{\mu}_{\hat{m}_t}^{(\mathbf{Y})}$ in (10) due to the previous explanation. Therefore, it is believed that some information from source feature vector to converted feature vector is missing. In other words, the dependency between the source and the converted feature vectors is weak. This leads to the converted sound losing some natural sound characteristics contributed from the source speech. To enhance this dependency, we incorporated MMI into parameter generation and proposed an integrated maximum a posteriori and mutual information criteria (MAPMI) method for the GMM-based VC framework.

3. Proposed parameter generation algorithm

For the proposed MAPMI method, we follow the same training and conversion procedures to that used in the MAP-based mapping with replacing the MAP criterion in (7) by the integrated MAP and MMI criterion. In this section, we first review the MI criterion and then present the proposed MAPMI-based mapping.

3.1. Mutual Information

The mutual information (MI) between two continuous random variables can be defined as [7]:

$$MI(\mathbf{X}, \mathbf{Y}) = \int_{\mathbf{Y}} \int_{\mathbf{X}} P(\mathbf{X}, \mathbf{Y} \mid \Theta^{(\mathbf{Z})}) \log \frac{P(\mathbf{X}, \mathbf{Y} \mid \Theta^{(\mathbf{Z})})}{P(\mathbf{Y} \mid \Theta^{(\mathbf{Z})}) P(\mathbf{X} \mid \Theta^{(\mathbf{Z})})} d\mathbf{X} d\mathbf{Y}. \quad (11)$$

In this study, $\mathbf{X}$ and $\mathbf{Y}$ are the source and target feature sequences, respectively; $P(\mathbf{X}, \mathbf{Y} \mid \Theta^{(\mathbf{Z})})$ is a JDGMM given in (1); $P(\mathbf{X} \mid \Theta^{(\mathbf{Z})})$ and $P(\mathbf{Y} \mid \Theta^{(\mathbf{Z})})$ are the corresponding marginal PDF of the source and target feature vectors, respectively. Computing the MI directly from PDF modeled by the GMM and integration shown in (11) is difficult. Thus, a numerical method is adopted here to approximate (11). The (11) can be rewritten as

$$MI(\mathbf{X}, \mathbf{Y}) = E_{\mathbf{XY}} \left[\log \frac{P(\mathbf{X}, \mathbf{Y} \mid \Theta^{(\mathbf{Z})})}{P(\mathbf{Y} \mid \Theta^{(\mathbf{Z})}) P(\mathbf{X} \mid \Theta^{(\mathbf{Z})})} \right] = E_{\mathbf{XY}} \left[\log \frac{P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})})}{P(\mathbf{Y} \mid \Theta^{(\mathbf{Z})})} \right] \quad (12)$$

where $E_{\mathbf{XY}}$ denotes the expectation operation. By the law of large numbers, the expectation operation can be approximated by a sample mean as

$$MI(\mathbf{X}, \mathbf{Y}) \approx \frac{1}{T} \sum_{t=1}^{T} \log \frac{P(\mathbf{Y}_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})})}{P(\mathbf{Y}_t \mid \Theta^{(\mathbf{Z})})}. \quad (13)$$

3.2. MAPMI-based mapping

In this section, we describe the MAPMI-based mapping method to improve the sound characteristics of the converted feature vectors. We define the objective function, which combines the log-scaled conditional PDF and MI, as

$$L = \frac{1}{T}\log\left\{P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Z})})^{1-\alpha}\times\left(\frac{P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Z})})}{P(\mathbf{Y}\mid\Theta^{(\mathbf{Z})})}\right)^{\alpha}\right\}$$
$$= (1-\alpha)\times\frac{1}{T}\sum_{t=1}^{T}\log P(\mathbf{Y}_t\mid\mathbf{X}_t,\Theta^{(\mathbf{Z})}) + \alpha\times\frac{1}{T}\sum_{t=1}^{T}\log\frac{P(\mathbf{Y}_t\mid\mathbf{X}_t,\Theta^{(\mathbf{Z})})}{P(\mathbf{Y}_t\mid\Theta^{(\mathbf{Z})})}. \quad (14)$$

In (14), the first term is the log-scaled conditional PDF in (2), and the second term is the MI in (13). The power weight α controls the weighting between the log-scaled conditional PDF and MI. Thus, in the same manner of MAP-based mapping considering the dynamic features, the MAPMI-based mapping can be formulated as

$$\hat{\mathbf{y}}_S = \arg\max\ \frac{1}{T}\log\left\{P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Z})})^{1-\alpha}\times\left(\frac{P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Z})})}{P(\mathbf{Y}\mid\Theta^{(\mathbf{Z})})}\right)^{\alpha}\right\}$$
$$\text{s.t. } \mathbf{Y} = \mathbf{W}\mathbf{y}_S. \quad (15)$$

It is noted that in (15), the proposed parameter generation algorithm is performed based on an integration of MAP and MMI criteria. Similar to the MAP-based mapping, the log-scaled conditional PDF and MI shown in (15) can be approximated with a single sequence of mixture components

$$\log P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Z})}) \approx \log P(\hat{\mathbf{m}}\mid\mathbf{X},\Theta^{(\mathbf{Z})})P(\mathbf{Y}\mid\mathbf{X},\hat{\mathbf{m}},\Theta^{(\mathbf{Z})})$$
$$\log\frac{P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Z})})}{P(\mathbf{Y}\mid\Theta^{(\mathbf{Z})})} \approx \log\frac{P(\hat{\mathbf{m}}\mid\mathbf{X},\Theta^{(\mathbf{Z})})P(\mathbf{Y}\mid\mathbf{X},\hat{\mathbf{m}},\Theta^{(\mathbf{Z})})}{P(\hat{\mathbf{m}})P(\mathbf{Y}\mid\hat{\mathbf{m}},\Theta^{(\mathbf{Z})})} \quad (16)$$

where the mixture component sequence $\hat{\mathbf{m}}$ is determined by $\hat{\mathbf{m}} = \arg\max P(\mathbf{m}\mid\mathbf{X},\Theta^{(\mathbf{Z})})$. By substituting the approximation results (16) into (15), we can obtain the converted static feature sequence by

$$\hat{\mathbf{y}}_S = (\mathbf{W}^{\mathrm{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1}\mathbf{W} - \alpha\cdot\mathbf{W}^{\mathrm{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y})-1}\mathbf{W})^{-1} \times(\mathbf{W}^{\mathrm{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1}\mathbf{E}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})} - \alpha\cdot\mathbf{W}^{\mathrm{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y})-1}\mathbf{E}_{\hat{\mathbf{m}}}^{(\mathbf{Y})}) \quad (17)$$

where

$$\mathbf{E}_{\hat{\mathbf{m}}}^{(\mathbf{Y})} = [(\boldsymbol{\mu}_{\hat{m}_1}^{(\mathbf{Y})})^{\mathrm{T}},\cdots,(\boldsymbol{\mu}_{\hat{m}_t}^{(\mathbf{Y})})^{\mathrm{T}},\cdots,(\boldsymbol{\mu}_{\hat{m}_T}^{(\mathbf{Y})})^{\mathrm{T}}]^{\mathrm{T}}$$
$$\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y})-1} = \mathrm{diag}[\boldsymbol{\Sigma}_{\hat{m}_1}^{(\mathbf{YY})-1},\cdots,\boldsymbol{\Sigma}_{\hat{m}_t}^{(\mathbf{YY})-1},\cdots,\boldsymbol{\Sigma}_{\hat{m}_T}^{(\mathbf{YY})-1}]; \quad (18)$$

$\mathbf{E}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})}$ and $\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1}$ are given in (10). Please note that when setting α =0 in (17), the converted static feature sequence is the same to that derived by the MAP-based mapping method in (9).

4. Experimental results

4.1. Evaluation conditions

We evaluated the conventional MAP-based (9) and proposed MAPMI-based (17) mapping methods (referred to as MAP and MAPMI, respectively, in the following discussion) on a parallel Mandarin speech corpus. This corpus consisted of 2 speakers, one female and one male. Eighty parallel sentences were selected from both speakers. Among the 80 sentences, we used 40 sentences for establishing the conversion system and the rest 40 sentences for performing conversion and testing evaluations.

Speech signals were firstly recorded in a 20kHz sampling rate, and then down-sampled to 16kHz. The resolution per sample was 16 bits. The spectral features were the first through 24th Mel-cepstral coefficients extracted from the STRAIGHT smoothed spectra [8]. The analysis window was the pitch synchronous window with a 5ms window shift. A dynamic time warping (DTW) algorithm was performed within each syllable boundary to obtain a joint feature sequence during the training phases. The number of Gaussian mixtures was set to 64 for both mapping methods. In this study, the power weight α in the MAPMI in (17) was set to 0.3, which was determined based on a preliminary experiment on the training set. In the following, we report the objective and subjective evaluations on the female to male spectral conversion.

4.2. Objective evaluations

We conducted two objective evaluations, namely, conversion accuracy and dependency, to compare the conventional and the proposed methods. For the conversion accuracy evaluation, we calculate the difference of the target and converted Mel-cepstral feature vectors by Mel-cepstral distortion (MCD), $D_{MCD}(\mathbf{y}_{S_t},\hat{\mathbf{y}}_{S_t})$, which is defined by

$$D_{MCD}(\mathbf{y}_{S_t},\hat{\mathbf{y}}_{S_t}) = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(y_{S_t}^d - \hat{y}_{S_t}^d)^2} \quad (19)$$

where $\mathbf{y}_{S_t}$ and $\hat{\mathbf{y}}_{S_t}$ are the target and converted feature vectors, respectively. The unit of the MCD measure is dB; a lower MCD value indicates a more accurate conversion. For the dependency evaluation, we calculated the mutual informal (MI) of the converted and source feature vector based on (13). The unit of the MI measure is nat; a larger MI value suggests a higher dependency. We listed the MCD [dB] and MI [nat] results of MAP and MAPMI in Table 1.

From Table 1, we first observed that MAPMI produces converted sounds with higher MI than MAP. This result confirms that MAPMI can enhance the dependency of the source and converted sounds comparing to MAP. Next, we observed that MAPMI produces a higher MCD. This result suggests that the incorporation of MI can degrade the conversion accuracy. It should be noted that many studies have shown that there was no strong relationship between the MCD results and the subjective test results [3] and [5]. We will investigate this part in our sub-

Table 1: *Objective tests of the conventional* (MAP) *and proposed* (MAPMI) *methods. The MCD before the conversion is 9.37 dB.*

Methods	MI [nat]	MCD [dB]
MAP	2.42	5.12
MAPMI	3.25	6.01

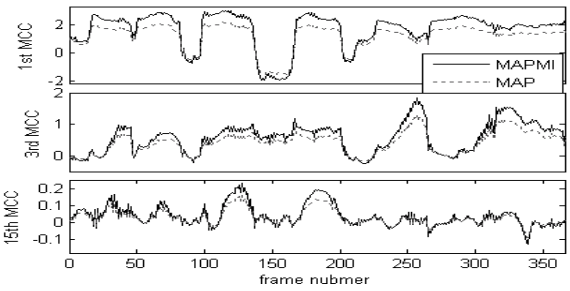

Figure 1: *An example of trajectories converted by the conventional* (MAP) *and proposed* (MAPMI) *methods.*

jective evaluations in the next section.

In addition to MCD and MI, we plotted the trajectories converted by MAP and MAPMI as another objective comparison. In Fig. 1, the trajectories of the 1st, 3rd, and 15th MCC (Mel-cepstral coefficient) are plotted in the first, second, and third panels, respectively, all from a same utterance. From Fig. 1, it can be seen that MAP generated overly smoothed trajectories, and the trajectory movements were greatly enhanced by MAPMI. We observed the same phenomena for other MCC elements in all the 40 evaluation utterances. In the next section, we present the results of the subjective evaluations.

4.3. Subjective evaluations

We conducted a formal listening test and used the mean opinion score (MOS) method to evaluate the speech quality and speaker individuality of converted speech. The evaluation was performed by16 subjects; all of them have research experience in the speech processing field. Twenty five test sentences were randomly chosen from the test set for 16 subjects. Samples were presented in random order for the 25 test sentences. The subjects were asked to give a five point scale on both quality and similarity tests (5: very good/similar, 1: very bad/dissimilar) after listening to the analysis-synthesized target voice. Since this study focused on spectral conversion, the F_0 conversion was applied for both methods using the simple linear transformation as [3].

Fig.2 shows the results of the MOS test. From Fig. 2, we can see that the proposed MAPMI method outperformed the conventional MAP method on both speech quality and speaker individuality tests. Particularly, the proposed method achieved significant gains on speech quality. This result implies that MI plays an important cue to speech quality and speaker individuality. For a further analysis, we noted that the proposed method effectively overcomes the muffled sound issue. Moreover, we found that the intelligibility of the converted sound is improved according to the responses of the subjects. Note that in the objective evaluations in Sec. 4.2, MAPMI gives higher MI and lower MCD measures comparing to MAP. Therefore, it is confirmed that by increasing the dependency to the source sound, the quality of the converted sound can be enhanced, although the conversion accuracy is marginally degraded.

5. Conclusion

In this paper, we incorporated MI into the parameter generation algorithm and proposed the MAPMI-based mapping for spectral conversion of the VC task. The proposed MAPMI improved the

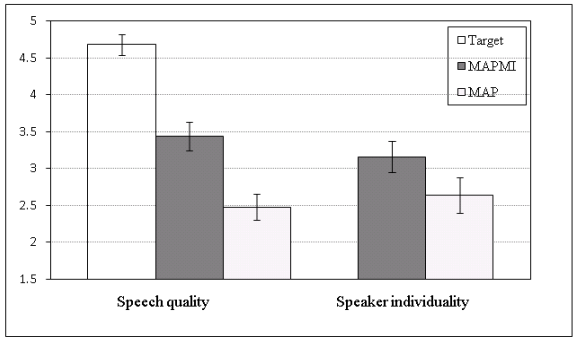

Figure 2: *Subjective test results of converted speech with conventional* (MAP) *and proposed* (MAPMI) *methods. Error bars indicate 95% confidence intervals.* "Target" *shown in speech quality denotes the analysis-synthesized target speech.*

dependency between source and converted sounds to overcome the over-smoothing issue in the conventional GMM-based methods. Experimental results confirmed that the proposed method dramatically improved the quality and similarity in terms of subjective tests. Because the mechanism of overcoming the over-smoothing issue by the proposed method differs from that by GV [3-5], our first future work will focus on combining the proposed method with GV. Next, we will investigate the effectiveness of applying MI for the F_0 conversion to further enhance the MAPMI-based mapping for the VC task.

6. Acknowledgements

The authors would like to thank Prof. H. Kawahara of Wakayama University, Japan, for the permission to use the STRAIGHT method.

7. References

[1] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, pp. 131-142, Mar. 1998.

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

[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, Nov. 2007.

[4] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous Stochastic Feature Mapping Based on Trajectory HMMs," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 2, pp. 417-430, Feb. 2011.

[5] H. Benisty and D. Malah, "Voice Conversion using GMM with Enhanced Global Variance", in *Proc. Interspeech*, Florence, Aug 2011, pp. 669-672.

[6] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice Conversion with Smoothed GMM and MAP Adaptation", in *Proc. Interspeech*, Geneva, Sep 2003, pp. 2413-2416.

[7] T. M. Cover and J. A. Thomas, "Elements of Information Theory", Wiley, 1991.

[8] 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.