EXPLORING MUTUAL INFORMATION FOR GMM-BASED SPECTRAL CONVERSION

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

[1]Dept. of Electrical and Computer 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

ABSTRACT

In this paper, we propose a maximum mutual information (MMI) training criterion to refine the parameters of the joint density GMM (JDGMM) set to tackle the over-smoothing issue in voice conversion (VC). Conventionally, the maximum likelihood (ML) criterion is used to train a JDGMM set, which characterizes the joint property of the source and target feature vectors. The MMI training criterion, on the other hand, updates the parameters of the JDGMM set to increase its capability on modeling the dependency between the source and target feature vectors, and thus to make the converted sounds closer to the natural ones. The subjective listening test demonstrates that the quality and individuality of the converted speech by the proposed ML followed by MMI (ML+MMI) training method is better that by the ML training method.

Index Terms— Voice conversion, mutual information, GMM.

1. INTRODUCTION

The task of voice conversion (VC) is to transform a voice to another specific voice [1-5]. Generally, we can divide the overall VC procedure into two parts: spectral and prosody conversions. In this study, we focus on spectral conversion (SC). However, we will use VC and SC alternatively in this paper.

Many SC approaches have been proposed in the past. Among them, the Gaussian mixture model (GMM)-based method is a successful one [1-5]. The GMM-based method first estimates a joint density GMM (JDGMM) set to characterize the joint property of the source and target voices in the training phase. Then, the conversion of spectral parameters is performed online in a frame by frame manner using the estimated JDGMM set. Although the GMM-based method has been proven effective, several issues remain to overcome.

One major issue of the GMM-based method is the over-smoothing problem. It is often 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 issue. A notable one is to incorporate the global variance (GV) into the conversion process [3]. The quality of the converted voice is clearly enhanced by refining the converted feature vector sequence to match the GV of the target speaker.

A previous study has reported that the true mapping from source to target voice features is complex and nonlinear [4]. However, the GMM-based method performs SC by simplifying the mapping with a linear transformation. This simplification might reduce the correlation of the converted sound and the natural sound characteristics from the source voice, and accordingly produces muffled converted sound. To overcome this problem, we have proposed an MAPMI (maximum a posteriori and mutual information)-based mapping criterion in our previous study [5]. The MMI criterion to capture the dependency between the source and converted feature vectors effectively works as a penalty term to enhance the dependence of the converted feature vectors on the source feature vectors. It is verified that the quality of the converted speech is dramatically enhanced by the MAPMI-based mapping criterion [5].

In this paper, we further investigate the introduction of the mutual information (MI) criterion to improve the training procedure for GMM-based SC. In the conventional GMM-based SC, the maximum likelihood (ML) criterion is applied to estimate a JDGMM set. Here, we propose using a maximum mutual information (MMI) training criterion to update the parameters of the ML-trained JDGMM set to increase its capability on modeling the dependency between the source and target feature vectors, and thus to make the converted sounds closer to the natural ones.

The remainder of this paper is organized as follows. Section 2 reviews the conventional GMM-based SC and discusses the existing issues. Section 3 describes the proposed MMI training criterion. Section 4 presents our experimental setup and result analysis. Finally, we conclude this work in Section 5.

2. GMM-BASED SPECTRAL CONVERSION CONSIDERING THE DYNAMIC FEATURES

2.1. Training a JDGMM Set

In a SC system, a parallel speech corpus must be prepared to train 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 vector sequences aligned by dynamic time warping, where T is the number of 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 source and target feature vectors at frame t, each comprising D static and D dynamic features. A JDGMM set is estimated 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 . $\boldsymbol{\mu}_m^{(\mathbf{Z})}=[(\boldsymbol{\mu}_m^{(\mathbf{X})})^{\mathrm{T}},(\boldsymbol{\mu}_m^{(\mathbf{Y})})^{\mathrm{T}}]^{\mathrm{T}}$, 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 prior, 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 set to be 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 by the EM algorithm.

2.2. Estimating the Conditional PDF

Given a JDGMM set, 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 for mapping (also called the ML-based mapping in [3]), the converted static feature vector sequence $\hat{\mathbf{y}}_S$ is obtained as

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

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

In practical implementations [3,5], $P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})})$ in (7), is often 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 vector 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

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

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 $\hat{\mathbf{m}} = [\hat{m}_1, \cdots, \hat{m}_t, \cdots, \hat{m}_T]$.

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 [4], the authors have explained that the over-smoothing issue comes from the complex and nonlinear true mapping from the source to the target speech features. However, the GMM-based method may have limited capability to characterize the true mapping precisely. In (5), the correlation term $\boldsymbol{\Sigma}_m^{(\mathbf{YX})} \boldsymbol{\Sigma}_m^{(\mathbf{XX})-1}$ can be very small, which results in $\boldsymbol{\mu}_{m,t}^{(\mathbf{Y}|\mathbf{X})} \approx \boldsymbol{\mu}_m^{(\mathbf{Y})}$. This can make the converted feature vector sequence given by minimum mean square error (MMSE)-based mapping [2] converge to the weighted mean of the target mixture components, and make the converted voice sound muffled.

The same problem exists in MAP-based mapping. We often observe that, in (10), $\boldsymbol{\mu}_{\hat{m}_t,t}^{(\mathbf{Y}|\mathbf{X})} \approx \boldsymbol{\mu}_{\hat{m}_t}^{(\mathbf{Y})}$. Obviously, some information in the source feature vectors is missing during conversion. This can cause the converted voice to lose natural characteristics. To enhance the dependency between the source and converted voices, we have recently proposed a new MAPMI-based mapping criterion in the conversion phase and obtained significant quality improvements in the converted speech [5]. In this study, we further try to enhance the dependency in the training phase.

3. THE PROPOSED MMI TRAINING METHOD

The MMI training criterion is to refine the parameters of the JDGMM set to increase its capability of characterizing the dependency between the source and target feature vectors. In this section, we first review the MI criterion, and then present the proposed MMI-training method.

3.1. Mutual Information

The mutual information (MI) between two continuous random variables can be defined as [6]:

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

In this study, $\mathbf{X}$ and $\mathbf{Y}$ are the source and target feature vector sequences, respectively; $P(\mathbf{X}, \mathbf{Y} \mid \Theta^{(\mathbf{Z})})$ is the likelihood of $\mathbf{X}$ and $\mathbf{Y}$ given a JDGMM set; $P(\mathbf{X} \mid \Theta^{(\mathbf{Z})})$ and $P(\mathbf{Y} \mid \Theta^{(\mathbf{Z})})$ are the marginal PDF of the source and target feature vectors, respectively. The direct computation in (11) is difficult. With the law of large numbers, MI in (11) can be approximated by

$$MI(\mathbf{X}, \mathbf{Y}) \approx \frac{1}{T} \sum_{t=1}^{T} \log \frac{P(\mathbf{Y}_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})})}{P(\mathbf{Y}_t \mid \Theta^{(\mathbf{Z})})} \quad (12)$$

3.2. MMI Training

The goal of MMI training is to determine a parameter set, $\Theta^{(\mathbf{Z})}$ ($\Theta^{(\mathbf{Z})} = \{\omega_m, \boldsymbol{\mu}_m^{(\mathbf{Z})}, \boldsymbol{\Sigma}_m^{(\mathbf{Z})}\}_{m=1}^{M}$), that maximizes the MI given in (12). For ease of computation, we approximate the conditional PDF $P(\mathbf{Y}_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})})$ by a single mixture component $\hat{m}_t$ as

$$P(\mathbf{Y}_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) \approx P(\hat{m}_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid \mathbf{X}_t, \hat{m}_t, \Theta^{(\mathbf{Z})}), \quad (13)$$

where $\hat{m}_t$ is determined by $\hat{m}_t = \arg\max P(m_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})})$. By substituting (13) into (12), the MI function is further approximated as

$$MI'(\mathbf{X}, \mathbf{Y}) = \frac{1}{T} \sum_{t=1}^{T} \log \frac{P(\hat{m}_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid \mathbf{X}_t, \hat{m}_t, \Theta^{(\mathbf{Z})})}{P(\mathbf{Y}_t \mid \Theta^{(\mathbf{Z})})} \quad (14)$$

In the implementation, we first prepare a ML-trained JDGMM set with the parameter set, $\Theta^{(\mathbf{Z})}$, and update the parameter set by maximizing the MI function in (14), i.e.,

$$\hat{\Theta}^{(\mathbf{Z})} = \arg \max \frac{1}{T} \sum_{t=1}^{T} \log \frac{P(\hat{m}_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid \mathbf{X}_t, \hat{m}_t, \Theta^{(\mathbf{Z})})}{\sum_{l=1}^{M} P(l \mid \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid l, \Theta^{(\mathbf{Z})})}. \quad (15)$$

In this study, we only update the mean and covariance parameters of the target voice, i.e., $\{\boldsymbol{\mu}_m^{(\mathbf{Y})}, \boldsymbol{\Sigma}_m^{(\mathbf{YY})}\}_{m=1}^M$; the remaining parameters are kept the same as those of the ML-trained JDGMM set.

For simplicity, we denote the parameter set $\{\boldsymbol{\mu}_m^{(\mathbf{Y})}, \boldsymbol{\Sigma}_m^{(\mathbf{YY})}\}$ by Φ_m in the following discussions. We adopt the generalized gradient ascent algorithm to iteratively update Φ_m

$$\Phi_m(n+1) = \Phi_m(n) + \varepsilon \sum_{t=1}^{T} \frac{\partial I_{\hat{m}_t}(\mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \Phi_m(n)} \tag{16}$$

where n denotes the nth iteration number, ε is the step size, and

$$I_{\hat{m}_t}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})}) = \log \frac{P(\hat{m}_t \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid \mathbf{X}_t, \hat{m}_t, \Theta^{(\mathbf{Z})})}{\sum_{l=1}^{M} P(l \mid \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid l, \Theta^{(\mathbf{Z})})}. \tag{17}$$

To calculate the partial derivatives of (17) with respect to $\boldsymbol{\mu}_m^{(\mathbf{Y})}$ and $\boldsymbol{\Sigma}_m^{(\mathbf{YY})}$, with $\boldsymbol{\Sigma}_m^{(\mathbf{YY})}$ being positive definite, we first transform the parameters to an unconstrained domain [8]. Then, we have transformed parameters, $\tilde{\mu}_{m,d}^{(\mathbf{Y})} = \mu_{m,d}^{(\mathbf{Y})} / \sigma_{m,d}^{(\mathbf{YY})}$ and $\tilde{\sigma}_{m,d}^{(\mathbf{YY})} = \log \sigma_{m,d}^{(\mathbf{YY})}$, respectively, for $\mu_{m,d}^{(\mathbf{Y})}$ and $\sigma_{m,d}^{(\mathbf{YY})}$; d represents the dth dimension. Finally, by applying a partial derivative on (17) with respect to $\tilde{\mu}_{m,d}^{(\mathbf{Y})}$, we have

- If $m = \hat{m}_t$

$$\frac{\partial I_{\hat{m}_t}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \tilde{\mu}_{m,d}^{(\mathbf{Y})}} = \frac{y_t^{(d)} - \mu_{m,t,d}^{(\mathbf{Y}|\mathbf{X})}}{(\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})})^2} \cdot \sigma_{m,d}^{(\mathbf{YY})} - \frac{P(m \mid \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid m, \Theta^{(\mathbf{Z})})}{\sum_{l=1}^{M} P(l \mid \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid l, \Theta^{(\mathbf{Z})})} \cdot \frac{y_t^{(d)} - \mu_{m,d}^{(\mathbf{Y})}}{\sigma_{m,d}^{(\mathbf{YY})}},$$

- If $m \neq \hat{m}_t$

$$\frac{\partial I_{\hat{m}_t}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \tilde{\mu}_{m,d}^{(\mathbf{Y})}} = -\frac{P(m \mid \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid m, \Theta^{(\mathbf{Z})})}{\sum_{l=1}^{M} P(l \mid \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid l, \Theta^{(\mathbf{Z})})} \cdot \frac{y_t^{(d)} - \mu_{m,d}^{(\mathbf{Y})}}{\sigma_{m,d}^{(\mathbf{YY})}}. \tag{18}$$

Similarly, by applying the partial derivative on (17) with respect to $\tilde{\sigma}_{m,d}^{(\mathbf{YY})}$, we have

- If $m = \hat{m}_t$

$$\frac{\partial I_{\hat{m}_t}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \tilde{\sigma}_{m,d}^{(\mathbf{YY})}} = \left(-(\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})})^{-2} + \left(\frac{y_t^{(d)} - \mu_{m,t,d}^{(\mathbf{Y}|\mathbf{X})}}{(\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})})^2} \right)^2 \right) \cdot (\sigma_{m,d}^{(\mathbf{YY})})^2 - \frac{P(m \mid \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid m, \Theta^{(\mathbf{Z})})}{\sum_{l=1}^{M} P(l \mid \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid l, \Theta^{(\mathbf{Z})})} \cdot \left(\left(\frac{y_t^{(d)} - \mu_{m,d}^{(\mathbf{Y})}}{\sigma_{m,d}^{(\mathbf{YY})}} \right)^2 - 1 \right), \tag{19}$$

- If $m \neq \hat{m}_t$

$$\frac{\partial I_{\hat{m}_t}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \tilde{\sigma}_{m,d}^{(\mathbf{YY})}} = -\frac{P(m \mid \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid m, \Theta^{(\mathbf{Z})})}{\sum_{l=1}^{M} P(l \mid \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid l, \Theta^{(\mathbf{Z})})} \cdot \left(\left(\frac{y_t^{(d)} - \mu_{m,d}^{(\mathbf{YY})}}{\sigma_{m,d}^{(\mathbf{YY})}} \right)^2 - 1 \right).$$

$\{\mu_{m,d}^{(\mathbf{X})}, \sigma_{m,d}^{(\mathbf{XX})}\}$ and $\{\mu_{m,d}^{(\mathbf{Y})}$ and $\sigma_{m,d}^{(\mathbf{YY})}\}$ are the mean and standard deviation values of the source and target PDFs; $\sigma_{m,d}^{(\mathbf{XY})}$ and $\sigma_{m,d}^{(\mathbf{YX})}$ are the cross standard deviation values of the joint feature vectors; $\mu_{m,t,d}^{(\mathbf{Y}|\mathbf{X})}$ and $\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}$ are the mean and standard deviation values of the conditional PDF; $x_t^{(d)}$ and $y_t^{(d)}$ are the source and target feature vectors, respectively. Based on (16) - (19), we can iteratively update $\{\boldsymbol{\mu}_m^{(\mathbf{Y})}, \boldsymbol{\Sigma}_m^{(\mathbf{YY})}\}_{m=1}^M$.

4. EXPERIMENTAL RESULTS

4.1. Experimental conditions

We compare the VC performance using the JDGMM sets trained by the ML and the proposed ML followed by MMI training criteria (referred to as ML and ML+MMI, respectively, in the following discussion). A parallel Mandarin speech corpus was chosen for evaluation. 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 to establish the conversion system and the remaining 40 sentences to perform conversion and conduct evaluation.

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 [7]. The analysis window was the pitch synchronous window. A dynamic time warping (DTW) algorithm was performed within each syllable boundary to obtain a joint feature vector sequence in the training phase. The number of Gaussian mixtures in each JDGMM set was 64. The maximum number of iterations in MMI training was set to 5. The step size was empirically determined. In the conversion phase, MAP-based mapping (based on (9)) was adopted to generate the converted spectrum sequence for both training methods. We report both the objective and subjective evaluations on the female to male VC task.

4.2. Objective evaluations

We conducted two objective evaluations, namely, conversion accuracy and dependency, to compare the ML and ML+MMI training methods. For the conversion accuracy evaluation, we calculate the difference of target and converted Mel-cepstral feature vectors, i.e., the Mel-cepstral distortion (MCD), $D_{MCD}(\mathbf{y}_{S_t}, \hat{\mathbf{y}}_{S_t})$, defined as

$$D_{MCD}(\mathbf{y}_{S_t}, \hat{\mathbf{y}}_{S_t}) = \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^{D} (y_{S_t}^d - \hat{y}_{S_t}^d)^2} \tag{20}$$

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 calculate the mutual information (MI) of the source and target feature vectors based on (14). The unit of the MI measure is nat. A larger MI value suggests a higher dependency. Table 1 shows the MCD [dB] and MI [nat] results of the ML and ML+MMI training methods.

From Table 1, we first observe that ML+MMI produces higher MI than ML. This result confirms that ML+MMI can enhance the dependency of the source and target sounds comparing to ML. We also observe that ML+MMI gives higher MCD than ML. Although

Table 1: The objective test results of the ML and ML+MMI training methods. The MCD value before VC is 9.37 dB.

Methods	MI	MCD
ML	-3.43	5.12
ML+MMI	-1.78	5.54

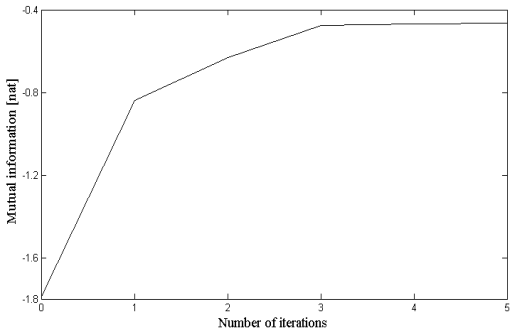

Figure 1: The mutual information curve of MMI training on the training set.

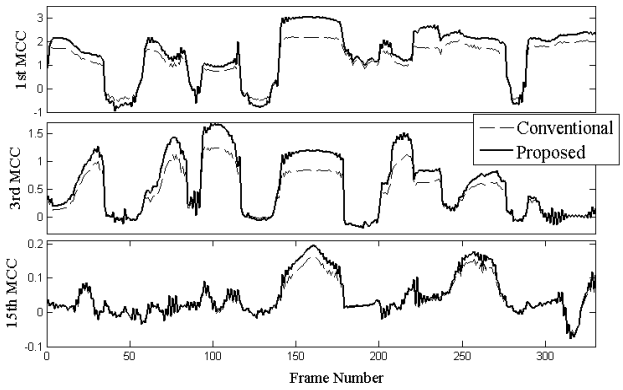

Figure 2: Example trajectories of MCC features converted by models trained by the conventional (ML) and proposed (ML+MMI) training methods.

this result seems to imply that the conversion accuracy might be degraded by the MMI training, many studies have shown that MCD might not be consistent with the subjective test results [3], [5]. We will discuss the subjective evaluations later.

Fig. 1 shows the MI value as a function of the number of iterations of MMI training, where the MI value at the 0th iteration is that of the ML-trained JDGMM set. We can see that the MI value consistently improved by MMI training and saturates after three iterations.

In addition to MCD and MI, we evaluate the trajectory of the converted features. Fig. 2 shows the trajectories of the 1st, 3rd, and 15th MCC (Mel-cepstral coefficient) of a speech utterance. From Fig. 2, it can be seen that the ML-trained model generated overly smoothed trajectories, while the ML+MMI-trained model clearly enhanced the trajectory movements. The same phenomenon can be observed in other MCC features of all the 40 evaluation utterances.

4.3. Subjective evaluations

We conducted a formal listening test and used the preference scores to evaluate the speech quality and speaker individuality of converted speech. The evaluation was performed by 12 subjects; all of them have research experience in the speech processing field. Twenty five test sentences were randomly selected from the test set for the 12 subjects. Samples were presented in random order for the 25 test sentences in the test of speech quality and speaker individuality. The subjects were asked which sample sounded more natural in the test of speech quality. In the test of speaker individuality, an ABX test was conducted; X denotes the analysis-

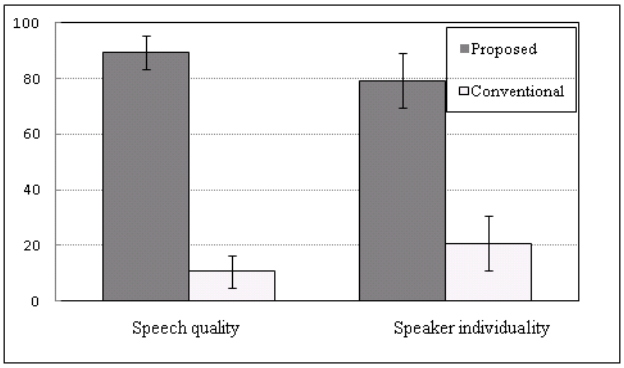

Figure 3: Preference test results of speech converted by models trained by the conventional (ML) and proposed (ML+MMI) training methods. Error bars indicate 95% confidence intervals. The unit of preference score is [%].

synthesized target speech; the speech converted by the ML-trained and ML+MMI-trained models were presented to the subjects in random order as A and B. The subjects were asked which sample sounded more similar to X. Since this study focuses on spectral conversion, a simple linear transformation method [3] is used for F_0 conversion for both training methods.

Fig.3 shows the results of the preference test. From Fig. 3, we can see that ML+MMI outperforms ML on both speech quality and speaker individuality tests. Particularly, the proposed ML+MMI method achieves 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 ML+MMI effectively overcomes the muffled sound problem, which is often observed in the ML-based system. Recall that, in the objective evaluations, ML+MMI gives higher MI measures and lower MCD values than ML. Therefore, it is confirmed that, by increasing the dependency between the source and target voices with the MMI training, the quality of the converted sound can be enhanced, although the MCD is slightly decreased.

5. CONCLUSIONS

In this paper, we have proposed a maximum mutual information (MMI) training criterion to refine the parameters in a JDGMM set in the training phase of VC. The refined JDGMM set is then used to perform VC in the conversion phase. Our subjective listening tests demonstrate that the MMI training provides clear improvements on speech quality and speaker individuality. In our previous study, it has been shown that incorporating the MI criterion in the conversion phase of VC can also enhances the quality of converted voices [5]. Thus, we believe that MI is an important factor in the GMM-based VC framework. In the future, we will further study the effectiveness of MI for GMM-based VC. We will also try to incorporate MI in F_0 conversion to further enhance the VC performance.

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] 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.

[5] H. T. Hwang, Y. Tsao, H. M. Wang, Y. R. Wang and S. H. Chen, "A Study of Mutual Information for GMM-Based Spectral Conversion," To Appear in *Interspeech2012*.

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

[7] 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.

[8] B. H. Juang, W. Chou, and C. H. Lee, " Minimum classification error rate methods for speech recognition," *IEEE Trans. Audio, Speech and Language Processing,* vol. 5, no. 3, pp. 257-265, May. 1997.