Alleviating the Over-Smoothing Problem in GMM-Based Voice Conversion with Discriminative Training

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

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

In this paper, we propose a discriminative training (DT) method to alleviate the muffled sound effect caused by over smoothing in the Gaussian mixture model (GMM)-based voice conversion (VC). For the conventional GMM-based VC, we often observed a large degree of ambiguities among acoustic classes (generative classes), determined by the source feature vectors for generating the converted feature vectors, causing the "muffled sound" effect on the converted voice. The proposed DT method is applied to refine the parameters in the maximum likelihood (ML)-trained joint density GMM (JDGMM) in the training stage to reduce the ambiguities among acoustic classes (generative classes) to alleviate the muffled sound effect. Experimental results demonstrate that the DT method significantly enhances the discriminative power between acoustic classes (generative classes) in the objective evaluation and effectively alleviates the muffled sound effect in the subjective evaluation.

Index Terms: Voice conversion, discriminative training, GMM.

1. Introduction

Voice conversion (VC) is a technique that transforms a source speaker's voice to that of a specific target speaker [1-12]. The overall VC includes two parts, namely spectral and prosody conversions. This study focuses on the spectral conversion (SC). Many SC approaches have been proposed in the past. Among them, the GMM-based SC is a successful one. Although the GMM-based SC has been proven effective, several issues remain, e.g., the muffled sound effect [3-8], time independent mapping [3,9], and the limited flexibility of the GMM-based SC framework [10]. In this study, we focus our attention on handling the muffled sound effect. The muffled sound effect occurs when the spectra are smoothed overly. Several methods have been proposed to handle this problem [3-8]. Erro *et al.* [4] and Godoy *et al.* [5] tackle this problem by enhancing the converted spectral envelope via dynamic frequency warping (DFW). Toda *et al.* [3] handle the problem by incorporating the global variance (GV) into the maximum likelihood (ML)-based conversion.

Our previous works [6, 7] proposed to tackle the over-smoothing problem by increasing the dependency between the source and target feature vectors. The maximum mutual information (MMI) criterion was incorporated in the conversion [6] and training phases [7] to increase the quality of the converted voice. Different from our previous works, this study investigates to tackle the problem by increasing the discriminative power between acoustic classes (referred to as generative classes in the following discussion).

By analyzing the conventional GMM-based SC and its generated sounds, we observe that a large degree of ambiguities among generative classes, determined by the source feature vectors, can cause the converted voices sound "muffled". Therefore, an effective way to handle the muffled sound effect is to reduce the degree of ambiguities among generative classes. In this study, we propose a discriminative training (DT) method to enhance the discriminative power among generative classes and to overcome the over-smoothing problem. To evaluate the proposed DT method, we adopt the posterior probability, denoted as Posterior, of a Gaussian component (generative class) given the source voice as the measure. Additionally, a formal listening test is conducted as the subjective evaluation. Evaluation results confirm that with the proposed DT method, the Posterior score is enhanced, and the muffled sound effect is effectively alleviated.

The remainder of this paper is organized as follows. Section 2 reviews the GMM-based SC with ML-based trajectory mapping, and discusses the over-smoothing problem. Section 3 describes the proposed method. Section 4 presents our experimental results. Our conclusion is given in the last section.

2. GMM-based spectral conversion

2.1. Training a JDGMM

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 feature vectors 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 joint density GMM (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

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

where ω_m is the prior probability of the mth mixture component; $\boldsymbol{\mu}_m^{(\mathbf{Z})}=[(\boldsymbol{\mu}_m^{(\mathbf{X})})^{\mathrm{T}},(\boldsymbol{\mu}_m^{(\mathbf{Y})})^{\mathrm{T}}]^{\mathrm{T}}$ 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. ML-based trajectory mapping

In the conversion phase, the ML-based trajectory mapping [3] is adopted to obtain the converted static feature sequence $\hat{\mathbf{y}}_S$ as

$$\hat{\mathbf{y}}_S = \arg\max P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})}), \text{ s.t. } \mathbf{Y} = \mathbf{W}\mathbf{y}_S \quad , \tag{2}$$

where $\mathbf{W}$ is a 2*DT*-by-*DT* weighting matrix (given in [3]) for computing the joint static and dynamic features; $P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})})$ is the conditional probability density function (PDF) given as

$$P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})}) = \sum_{\text{all } \mathbf{m}} P(\mathbf{m} \mid \mathbf{X}, \Theta^{(\mathbf{Z})}) P(\mathbf{Y} \mid \mathbf{X}, \mathbf{m}, \Theta^{(\mathbf{Z})}) , \tag{3}$$

where $\mathbf{m} = [m_1, \cdots, m_t, \cdots, m_T]$ is an arbitrary mixture component sequence. (3) can further be represented as

$$P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})}) = \prod_{t=1}^{T} \sum_{m=1}^{M} P(m \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid \mathbf{X}_t, m, \Theta^{(\mathbf{Z})}) , \tag{4}$$

where

$$P(m \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) = \frac{\omega_m N(\mathbf{X}_t; \boldsymbol{\mu}_m^{(\mathbf{X})}, \boldsymbol{\Sigma}_m^{(\mathbf{XX})})}{\sum_{l=1}^{M} \omega_l N(\mathbf{X}_t; \boldsymbol{\mu}_l^{(\mathbf{X})}, \boldsymbol{\Sigma}_l^{(\mathbf{XX})})} , \tag{5}$$

$$P(\mathbf{Y}_t \mid \mathbf{X}_t, m, \Theta^{(\mathbf{Z})}) = N(\mathbf{Y}_t; \boldsymbol{\mu}_{m,t}^{(\mathbf{Y}|\mathbf{X})}, \boldsymbol{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})}) . \tag{6}$$

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

$$\boldsymbol{\mu}_{m,t}^{(\mathbf{Y}|\mathbf{X})} = \mathbf{A}_m \mathbf{X}_t + \mathbf{b}_m, \ \boldsymbol{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})} = \boldsymbol{\Sigma}_m^{(\mathbf{YY})} - \mathbf{A}_m \boldsymbol{\Sigma}_m^{(\mathbf{XX})} \mathbf{A}_m^{\mathrm{T}} , \tag{7}$$

where the transformation matrix $\mathbf{A}_m$ and the bias vector $\mathbf{b}_m$ are given as

$$\mathbf{A}_m = \boldsymbol{\Sigma}_m^{(\mathbf{YX})} \boldsymbol{\Sigma}_m^{(\mathbf{XX})-1}, \ \mathbf{b}_m = \boldsymbol{\mu}_m^{(\mathbf{Y})} - \mathbf{A}_m \boldsymbol{\mu}_m^{(\mathbf{X})} . \tag{8}$$

In practical implementation, a suboptimal solution is often employed to ML-based trajectory mapping [3]. The conditional PDF, $P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})})$ in (2), can be approximated as $P(\mathbf{Y} \mid \mathbf{X}, \Theta^{(\mathbf{Z})}) \approx P(\hat{\mathbf{m}} \mid \mathbf{X}, \Theta^{(\mathbf{Z})}) P(\mathbf{Y} \mid \mathbf{X}, \hat{\mathbf{m}}, \Theta^{(\mathbf{Z})})$, where the mixture component sequence is determined by $\hat{\mathbf{m}} = \arg\max P(\mathbf{m} \mid \mathbf{X}, \Theta^{(\mathbf{Z})})$. As a result, the converted static feature sequence $\hat{\mathbf{y}}_S$ can be obtained by substituting the approximated conditional PDF into (2) as

$$\begin{aligned} \hat{\mathbf{y}}_S &= \arg\max P(\hat{\mathbf{m}} \mid \mathbf{X}, \Theta^{(\mathbf{Z})}) P(\mathbf{Y} \mid \mathbf{X}, \hat{\mathbf{m}}, \Theta^{(\mathbf{Z})}), \text{ s.t. } \mathbf{Y} = \mathbf{W}\mathbf{y}_S \\ &= \arg\max P(\mathbf{Y} \mid \mathbf{X}, \hat{\mathbf{m}}, \Theta^{(\mathbf{Z})}), \text{ s.t. } \mathbf{Y} = \mathbf{W}\mathbf{y}_S . \end{aligned} \tag{9}$$

2.3. The over-smoothing problem

From (9), each converted feature vector is mainly generated from its corresponding acoustic class (mixture component) modeled by the conditional PDF, $P(\mathbf{Y}_t \mid \mathbf{X}_t, \hat{m}, \Theta^{(\mathbf{Z})})$, or equivalently, $P(\mathbf{Z}_t \mid \hat{m}, \Theta^{(\mathbf{Z})})$; where the acoustic class (generative class) $\hat{m}$ is determined by the source feature vector at frame *t* by $\hat{m} = \arg\max P(m \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})})$.

To estimate the degree of ambiguities among generative classes, we use the posterior probability as the measure score

$$F(\mathbf{Z}_t; \Theta^{(\mathbf{Z})}) = \frac{1}{T} \sum_{t=1}^{T} \log P(\hat{m} \mid \mathbf{Z}_t, \Theta^{(\mathbf{Z})}) . \tag{10}$$

Notably, a lower posterior probability score indicates a larger degree of ambiguities among generative classes. In our experimental results, as will be shown later, we often observed a large degree of ambiguities among generative classes in the ML-trained JDGMM. A possible reason for this is due to the one-to-many problem [11, 12], which occurs when a single group of source features is associated with multiple groups of target features. Consequently, the converted voice is generated from ambiguous generative classes and sounds "muffled". To overcome the "muffled sound" effect, this study derives a discriminative training (DT) method to refine the ML-trained JDGMM to alleviate the ambiguities among generative classes and accordingly the "muffled sound" effect.

3. The proposed discriminative training method

The goal of the proposed DT method is to determine the parameter set $\Theta^{(\mathbf{Z})}$ in the JDGMM that maximizes the posterior probability in (10). Notably, maximizing the posterior probability to alleviate the ambiguities among classes is known as one of the most popular discriminative training methods in pattern recognition (equivalent to the maximum mutual information (MMI) training in the field of automatic speech recognition (ASR) [13]). Different from the discriminative training methods employed in ASR [13], which aim to improve the recognition performance, the DT method here is to alleviate the "muffled sound" effect in SC.

In the implementation, we first prepare an ML-trained JDGMM with the parameter set, $\Theta^{(\mathbf{Z})}$. Then, the generative class $\hat{m}$ is determined by the source feature vector at frame *t* as $\hat{m} = \arg\max P(m \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})})$. Finally, the updated parameter set $\Theta^{(\mathbf{Z})}$ can be obtained by maximizing the objective function in (10). For ease of deriving the parameters, the objective function in (10) can be further written by using the Bayes' rule as

$$F(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})}) = \frac{1}{T} \sum_{t=1}^{T} \log \frac{P(\hat{m} \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid \mathbf{X}_t, \hat{m}, \Theta^{(\mathbf{Z})})}{\sum_{l=1}^{M} P(l \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid \mathbf{X}_t, l, \Theta^{(\mathbf{Z})})}, \tag{11}$$

where $\Theta^{(\mathbf{Z})} = \{\mathbf{A}_m, \mathbf{b}_m, \omega_m, \boldsymbol{\mu}_m^{(\mathbf{X})}, \boldsymbol{\Sigma}_m^{(\mathbf{XX})}, \boldsymbol{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})}\}_{m=1}^{M}$. The proposed DT method adopts a gradient-based approach to update the parameters in $\Theta^{(\mathbf{Z})}$. In this study, we only update the transformation, bias, and covariance parameters, i.e., $\{\mathbf{A}_m, \mathbf{b}_m, \boldsymbol{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})}\}_{m=1}^{M}$; the remaining parameters are kept the same as those of the ML-trained JDGMM set. Our preliminary experiments show that applying DT on the complete parameter set $\Theta^{(\mathbf{Z})}$ only slightly improves the objective function in (11). For simplicity, we denote the parameter set $\{\mathbf{A}_m, \mathbf{b}_m, \boldsymbol{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})}\}$ by $\boldsymbol{\Phi}_m$ in the following discussion. The gradient-based method updates $\boldsymbol{\Phi}_m$ by

$$\boldsymbol{\Phi}_m(n+1) = \boldsymbol{\Phi}_m(n) + \varepsilon \sum_{t=1}^{T} \frac{\partial I_{\hat{m}}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \boldsymbol{\Phi}_m(n)} , \tag{12}$$

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

$$I_{\hat{m}}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})}) = \log \frac{P(\hat{m} \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid \mathbf{X}_t, \hat{m}, \Theta^{(\mathbf{Z})})}{\sum_{l=1}^{M} P(l \mid \mathbf{X}_t, \Theta^{(\mathbf{Z})}) P(\mathbf{Y}_t \mid \mathbf{X}_t, l, \Theta^{(\mathbf{Z})})} . \tag{13}$$

To calculate the partial derivatives of (13) with respect to $\mathbf{A}_m$, $\mathbf{b}_m$, and $\mathbf{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})}$ (where $\mathbf{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})}$ is positive definite), we first transform the parameters to an unconstrained domain [14], $\tilde{a}_{m,d} = a_{m,d} / \sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}$, $\tilde{b}_{m,d} = b_{m,d} / \sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}$ and $\tilde{\sigma}_{m,d}^{(\mathbf{Y}|\mathbf{X})} = \log \sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}$, respectively, for $a_{m,d}$, $b_{m,d}$ and $\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}$; d represents the dth dimension; $a_{m,d}$ is the dth diagonal element of the transformation matrix $\mathbf{A}_m$; $b_{m,d}$ is the dth element of the bias vector $\mathbf{b}_m$; $\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}$ is the standard deviation value of the conditional PDF, $P(\mathbf{Y}_t \mid \mathbf{X}_t, m, \Theta^{(\mathbf{Z})})$. Then, by applying the partial derivative on (13) with respect to $\tilde{a}_{m,d}$, $\tilde{b}_{m,d}$, and $\tilde{\sigma}_{m,d}^{(\mathbf{Y}|\mathbf{X})}$, we have

• if $m = \hat{m}$

$$\frac{\partial I_{\hat{m}}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \tilde{a}_{m,d}} = \left(1 - P(m \mid \mathbf{Z}_t, \Theta^{(\mathbf{Z})})\right) \cdot \left(\frac{y_t^{(d)} - \mu_{m,t,d}^{(\mathbf{Y}|\mathbf{X})}}{\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}} \cdot x_t^{(d)} \right)$$

$$\frac{\partial I_{\hat{m}}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \tilde{b}_{m,d}} = \left(1 - P(m \mid \mathbf{Z}_t, \Theta^{(\mathbf{Z})})\right) \cdot \left(\frac{y_t^{(d)} - \mu_{m,t,d}^{(\mathbf{Y}|\mathbf{X})}}{\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}} \right)$$

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

• if $m \neq \hat{m}$

$$\frac{\partial I_{\hat{m}}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \tilde{a}_{m,d}} = -P(m \mid \mathbf{Z}_t, \Theta^{(\mathbf{Z})}) \cdot \left(\frac{y_t^{(d)} - \mu_{m,t,d}^{(\mathbf{Y}|\mathbf{X})}}{\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}} \cdot x_t^{(d)} \right)$$

$$\frac{\partial I_{\hat{m}}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \tilde{b}_{m,d}} = -P(m \mid \mathbf{Z}_t, \Theta^{(\mathbf{Z})}) \cdot \left(\frac{y_t^{(d)} - \mu_{m,t,d}^{(\mathbf{Y}|\mathbf{X})}}{\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}} \right)$$

$$\frac{\partial I_{\hat{m}}(\mathbf{X}_t, \mathbf{Y}_t; \Theta^{(\mathbf{Z})})}{\partial \tilde{\sigma}_{m,d}^{(\mathbf{Y}|\mathbf{X})}} = -P(m \mid \mathbf{Z}_t, \Theta^{(\mathbf{Z})}) \cdot \left(\left(\frac{y_t^{(d)} - \mu_{m,t,d}^{(\mathbf{Y}|\mathbf{X})}}{\sigma_{m,d}^{(\mathbf{Y}|\mathbf{X})}} \right)^2 - 1 \right), \quad (14)$$

where

$$P(m \mid \mathbf{Z}_t, \Theta^{(\mathbf{Z})}) = \frac{\omega_m N(\mathbf{Z}_t; \boldsymbol{\mu}_m^{(\mathbf{Z})}, \mathbf{\Sigma}_m^{(\mathbf{Z})})}{\sum_{l=1}^{M} \omega_l N(\mathbf{Z}_t; \boldsymbol{\mu}_l^{(\mathbf{Z})}, \mathbf{\Sigma}_l^{(\mathbf{Z})})}, \quad (15)$$

$\mu_{m,t,d}^{(\mathbf{Y}|\mathbf{X})}$ is the mean value of the conditional PDF, $P(\mathbf{Y}_t \mid \mathbf{X}_t, m, \Theta^{(\mathbf{Z})})$, at frame t; $x_t^{(d)}$ and $y_t^{(d)}$ are the source and target feature vectors at frame t, respectively. The mean vector $\boldsymbol{\mu}_m^{(\mathbf{Z})}$ and covariance matrix $\mathbf{\Sigma}_m^{(\mathbf{Z})}$ in (15) can be obtained by

$$\boldsymbol{\mu}_m^{(\mathbf{Z})} = \begin{bmatrix} \boldsymbol{\mu}_m^{(\mathbf{X})} \\ \mathbf{A}_m \boldsymbol{\mu}_m^{(\mathbf{X})} + \mathbf{b}_m \end{bmatrix}, \mathbf{\Sigma}_m^{(\mathbf{Z})} = \begin{bmatrix} \mathbf{\Sigma}_m^{(\mathbf{XX})} & \mathbf{\Sigma}_m^{(\mathbf{XX})} \mathbf{A}_m^{\mathrm{T}} \\ \mathbf{A}_m \mathbf{\Sigma}_m^{(\mathbf{XX})} & \mathbf{A}_m \mathbf{\Sigma}_m^{(\mathbf{XX})} \mathbf{A}_m^{\mathrm{T}} + \mathbf{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})} \end{bmatrix}. \quad (16)$$

Based on (12)-(16), we can update $\{\mathbf{A}_m, \mathbf{b}_m, \mathbf{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})}\}_{m=1}^{M}$ iteratively.

4. Experimental results

4.1. Evaluation setup

We evaluate SC using the JDGMMs trained by the conventional ML training method and the proposed ML followed by DT method (denoted as ML+DT hereafter) 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 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 24 Mel-cepstral coefficients extracted from the STRAIGHT smoothed spectra [15]. 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 ($N_{\max}$) in the proposed DT method was set to 10. The step size ε was empirically determined and set to $\varepsilon \times (1 - n / N_{\max})$. In the conversion phase, ML-based trajectory mapping (9) (w/o considering the GV) was adopted to generate the converted static feature sequence for both training methods. We report both the objective and subjective evaluation results on the female to male SC task.

4.2. Objective evaluations

We conducted the objective evaluation in terms of two metrics, namely, the conversion accuracy and the degree of ambiguities among generative classes, to compare the ML and ML+DT trained JDGMMs. To evaluate the conversion accuracy, the Mel-cepstral distortion (MCD), $D_{MCD}(\mathbf{y}_{S_t}, \hat{\mathbf{y}}_{S_t})$, was used to calculate the difference of the target and converted Mel-cepstra in the evaluation set, which is given by

$$D_{MCD}(\mathbf{y}_{S_t}, \hat{\mathbf{y}}_{S_t}) \text{ [dB]} = \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^{D} (y_{S_t}^d - \hat{y}_{S_t}^d)^2}, \quad (17)$$

where $\mathbf{y}_{S_t}$ and $\hat{\mathbf{y}}_{S_t}$ are the target and converted static feature vectors, respectively. A lower MCD value indicates a more accurate conversion. Moreover, we calculate the posterior probability (Posterior) based on (10) to measure the degree of ambiguities among generative classes. A lower Posterior value indicates a larger degree of ambiguities among generative classes. Note that the maximum Posterior value is zero if $P(\hat{m} \mid \mathbf{Z}_t, \Theta^{(\mathbf{Z})}) = 1$ for all t, indicating no ambiguities among generative classes. Moreover, the average Posterior value is $\log(1/M)$ if $P(\hat{m} \mid \mathbf{Z}_t, \Theta^{(\mathbf{Z})}) = 1/M$ for all t (M is the total number of generative classes). In this study, the average Posterior value is -4.16 ($M = 64$).

Table 1 shows the MCD and Posterior results of the ML and ML+DT trained JDGMMs. From Table 1, we observe that the Posterior value given by the ML-trained JDGMM is even slightly lower than the average Posterior value. This result implies a serious mismatch/inconsistence between the generative classes selected by the source and joint feature vectors. In other words, the degree of ambiguities among generative classes is large in the ML-trained JDGMM. This result might be caused by the one-to-many problem detailed in [11, 12]. It is obvious that ML+DT produces a significantly higher Posterior value than ML. The Posterior value given by the ML+DT trained JDGMM is -2.25, while for the oracle case when the generative class $\hat{m}$ in (10) is determined by the joint feature vector $\mathbf{Z}_t$, the Posterior value is -0.11. The result confirms that the ambiguities among generative classes can be

Table 1: *Objective tests of the conventional* (*ML*) *and proposed ML followed by discriminative training* (*ML+DT*) *methods. The MCD before the conversion is 9.37 dB. The average Posterior value is -4.16.*

Methods	Posterior	MCD
ML	-4.57	5.12
ML+DT	-2.25	5.85

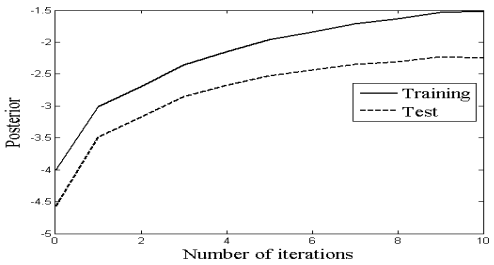

Figure 1: *The Posterior value curve of the proposed discriminative training (DT) method on the training and test sets.*

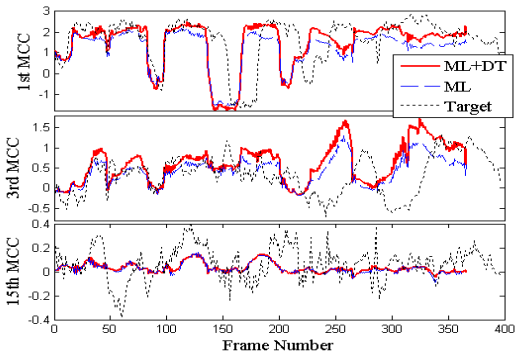

Figure 2: *Example trajectories of MCC features. Target: extracted from the target speech; ML: generated from the ML trained JDGMM; ML+DT: generated from the ML+DT trained JDGMM.*

effectively alleviated by the ML+DT method. From Table 1, we also observe that ML+DT gives slightly higher MCD than ML. Although this result seems to imply that ML+DT could degrade the voice quality, we obtain different observations from the subjective evaluation results, as will be discussed later.

Fig. 1 shows the Posterior value as a function of the number of iterations of the DT method, where the Posterior value at the 0th iteration is obtained by the ML-trained JDGMM. We can see that the Posterior value consistently increases, and then saturates after nine iterations. In addition to MCD and Posterior, 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 is obvious that the ML-trained JDGMM generates over-smoothed trajectories. Notably, the enhancement of the trajectory movements by the proposed DT method becomes unclear for high order MCCs, as can be observed from the 15th MCC plot in Fig. 2. The similar phenomena were observed for all of the 40 evaluation utterances. A previous study has shown that the dynamic ranges of the higher order MCCs are usually smaller than those of the lower order MCCs in natural speech [8]. Therefore, to improve the voice quality, enhancing the dynamic ranges of the lower order MCCs is more effective than enhancing the dynamic ranges of the higher order MCCs. This will be verified in the subjective evaluation.

4.3. Subjective evaluations

A formal listening test is conducted as the subjective evaluation in this study. The preference scores are used to evaluate the speech quality and speaker individuality of the converted speech. Ten subjects were involved in the evaluation, and all of these subjects have research background in the speech processing field. Twenty five sentences were randomly selected from the test set for each subject, and the samples for each sentence were presented in a random order in the test of speech quality and speaker individuality. In the speech quality test, the subjects were asked to determine which voice sample sounded more natural. In the speaker individuality test, an ABX test was performed. X denotes the analysis-synthesized target speech. The speech utterances generated by the ML and ML+DT trained JDGMMs were presented to the subjects in random order as A and B. The subjects were asked which sample (A or B) sounded more similar to X. Since we focus our attention on SC in this study, a same linear transformation [3] is applied for the prosody conversion throughout the evaluations.

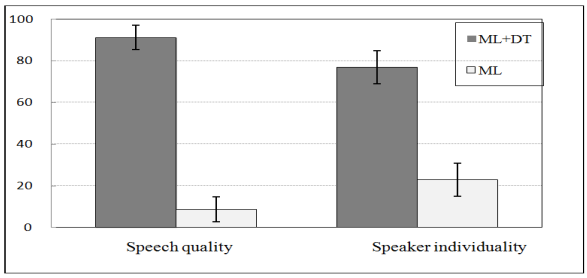

Figure 3: *Preference test results of speech converted by models trained by the conventional and proposed discriminative training methods. Error bars indicate 95% confidence intervals. The unit of preference score is* [%].

Fig.3 illustrates the evaluation results of the two preference tests. From Fig. 3, we observe that ML+DT outperforms ML in both speech quality and speaker individuality tests. This result implies that the Posterior score plays an important cue to both speech quality and speaker individuality. After an interview with the subjects, they all agreed that the proposed DT method effectively alleviates the "muffled sound" effect. Although yielding a slightly worse MCD value, applying the DT method to refine the JDGMM trained by the ML training method can improve the Posterior scores, reduce the ambiguities among discriminative classes, and thus enhance the quality of the converted speech.

5. Conclusion

In this study, we propose a discriminative training (DT) method to enhance the discriminative power among generative classes for the joint density GMM (JDGMM) and to overcome the over-smoothing problem in spectral conversion. Experimental results demonstrate that the discriminative power between generative classes can be indeed enhanced and the muffled sound effect can be effectively alleviated. In order to better understand the over-smoothing problem in GMM-based VC, our future works will investigate the connection between the one-to-many and weak correlation problems with the proposed DT method. Moreover, evaluations on more speaker pairs will be conducted in the future.

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. Speech Audio Process.*, vol. 6, no. 2, pp.131-142, Mar. 1998.

[2] A. Kain, and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, 1998, vol. 1, 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, Lang., Process.*, vol. 15, no. 8, pp. 2222-2235, Nov. 2007.

[4] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 18, no. 5, pp. 922-931, July. 2010.

[5] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans. Audio, Speech, Lang., Process*, vol. 20, no. 4, pp. 1313-1323, May. 2012.

[6] 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," *Proc. INTERSPEECH*, 2012.

[7] H. T. Hwang, Y. Tsao, H. M. Wang, Y. R. Wang and S. H. Chen, "Exploring mutual information for GMM-based spectral conversion," *Proc. ISCSLP*, 2012, pp. 50-54.

[8] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced global variance", *Proc. INTERSPEECH*, 2011, pp. 669-672.

[9] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 19, no. 2, pp. 417-430, Feb. 2011.

[10] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Statistical voice conversion based on noisy channel model," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 20, no. 6, pp. 1784-1794, Aug. 2012.

[11] A. Mouchtaris, Y. Agiomyrgiannakis, and Y. Stylianou, "Conditional vector quantization for voice conversion," *Proc. ICASSP*, 2007, vol. 4, pp. 505-508.

[12] E. Godoy, O. Rosec, and T. Chonavel, "Alleviating the one-to-many mapping problem in voice conversion with context-dependent modeling", *Proc. INTERSPEECH*, 2009, pp. 1627-1630.

[13] L. R. Bahl, P.F. Brown, P. V. De Souza, and L. R., Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Proc. ICASSP*, 1986, vol. 11, pp. 49-52.

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

[15] 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 Commun.*, vol. 27, no. 3-4, pp.187-207, 1999.