Voice Conversion Using Deep Neural Networks With Layer-Wise Generative Training

Ling-Hui Chen, Zhen-Hua Ling, *Member, IEEE*, Li-Juan Liu, and Li-Rong Dai

***Abstract*—This paper presents a new spectral envelope conversion method using deep neural networks (DNNs). The conventional joint density Gaussian mixture model (JDGMM) based spectral conversion methods perform stably and effectively. However, the speech generated by these methods suffer severe quality degradation due to the following two factors: 1) inadequacy of JDGMM in modeling the distribution of spectral features as well as the non-linear mapping relationship between the source and target speakers, 2) spectral detail loss caused by the use of high-level spectral features such as mel-cepstra. Previously, we have proposed to use the mixture of restricted Boltzmann machines (MoRBM) and the mixture of Gaussian bidirectional associative memories (MoGBAM) to cope with these problems. In this paper, we propose to use a DNN to construct a global non-linear mapping relationship between the spectral envelopes of two speakers. The proposed DNN is generatively trained by cascading two RBMs, which model the distributions of spectral envelopes of source and target speakers respectively, using a Bernoulli BAM (BBAM). Therefore, the proposed training method takes the advantage of the strong modeling ability of RBMs in modeling the distribution of spectral envelopes and the superiority of BAMs in deriving the conditional distributions for conversion. Careful comparisons and analysis among the proposed method and some conventional methods are presented in this paper. The subjective results show that the proposed method can significantly improve the performance in terms of both similarity and naturalness compared to conventional methods.**

***Index Terms*—Bidirectional associative memory, deep neural network, Gaussian mixture model, restricted Boltzmann machine, spectral envelope conversion, voice conversion.**

I. Introduction

VOICE conversion is a technique that attempts to modify one type of speech (source speech) to make it sound like another type of speech (target speech), while retaining the linguistic information. There are many applications for voice conversion, such as converting from one speaker (source speaker) to another speaker (target speaker) [1], from impaired speech to normal speech [2], from speaking speech to singing speech [3], whisper speech enhancement [4], etc. In the narrow sense, voice conversion means speaker conversion. Most of the state-of-art research on voice conversion focus on spectral conversion techniques.

Many approaches have been proposed for spectral conversion during the last decades. Generally, these approaches can be divided into two categories: rule based methods and statistical methods. The rule based methods directly modify the speech spectra based on the acoustical knowledge of the speech signal, e.g. simply moving the position of formants according to the difference between the formants of two speakers [5], [6]. Therefore, those methods can retain most of the details in the spectra and hence lead to good naturalness in the converted speech. However, the performance of these rule based methods, especially the similarity of the converted speech toward target speech, is not stable, e.g., they may work very well between some speaker pairs but are unsatisfactory between other speaker pairs. This is because the information (e.g. formants) for building the conversion function is usually difficult to be extracted accurately. On the other hand, the statistical methods [7], [8], [9] employ statistical models to automatically estimate the mapping relationship between the spectral features of the source and target speakers. By contrast to rule based methods, statistical methods can construct more precise mapping functions using complex statistical models. Therefore, statistical methods, especially the Gaussian mixture model (GMM) based spectral conversion methods, have become the mainstream research techniques nowadays.

However, there are several problems with the classical joint density GMM (JDGMM)-based spectral conversion method. Firstly, the conversion function derived from a JDGMM is a piece-wise linear transformation [8], which is insufficient to model the highly non-linear mapping relationship between the spectra of source and target speakers. Secondly, the inter-speaker covariances of JDGMM are weaker than intra-speaker covariances. The source features that belongs to the same mixture thus tend to be converted to the same target, which is a weighted average of means of the target GMM model. Therefore the converted features have smaller variance than the source or target features, which could be a reason for the "muffled" sound in converted speech [10]. Besides, the GMM are usually estimated on data of low-dimensional spectral features, such as mel-cepstra, line spectral pairs (LSPs), and etc. Although these low-dimensional spectral features can well represent speech spectra, some important spectral details are lost during the feature extraction. These problems may affect the similarity and quality of generated speech.

Manuscript received February 20, 2014; revised July 01, 2014; accepted August 26, 2014. Date of publication September 04, 2014; date of current version September 13, 2014. This work was supported in part by the National Nature Science Foundation of China under Grants 61273264 and 61273032 and the National 973 program of China under Grant 2012CB326405 . The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Junichi Yamagishi.

The authors are with the National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei 230027, China (e-mail: chenlh@ustc.edu.cn; zhling@ustc.edu.cn; ljliu037@mail.ustc.edu.cn; lrdai@ustc.edu.cn).

Digital Object Identifier 10.1109/TASLP.2014.2353991

Many approaches have been proposed in order to cope with these problems. Some approaches attempt to compensate the inadequacy of GMM in modeling by using trajectory-level information, e.g., integrating dynamic features and global variance (GV) into the conventional parameter generation criterion [11], reformulating the frame-level GMM as a trajectory model for modeling and generating sentence-level feature trajectories [12], etc. Some approaches directly use real spectral samples to generate speech waveforms [13], [14], etc. There are also some approaches to construct non-linear mapping relationships, such as building mapping relationships in the high-order kernel space [15], [16], using artificial neural networks (ANNs) to model the mapping relationships between the mel-cepstra of two speakers [17], [18], using a conditional restricted Boltzmann machine (CRBM) to directly model the conditional distributions of target spectral features given source spectral features [19], etc.

Previously, we have proposed to use the mixture of restricted Boltzmann machines (MoRBM) [20] and mixture of Gaussian bidirectional associative memories (MoGBAM) [21] to model the distribution of raw spectral envelopes instead of the GMM. The restricted Boltzmann machine (RBM) has a strong ability to model the distribution of spectral envelopes. It has been successfully applied to spectral envelope modeling for statistical parametric speech synthesis [22]. However, in spectral conversion using the MoRBM, it is not straightforward to derive the conditional distributions from the joint distributions given by the RBMs. The performance is degraded by the approximations made at the conversion stage [20]. On the other hand, although the Gaussian bidirectional associative memory (GBAM) is a two-layer stochastic neural network that can model the inter-dimensional correlations, it is theoretically equivalent to a single Gaussian distribution [21]. Therefore, the modeling ability of MoGBAM is limited and the derived mapping relationship is still a piece-wise linear transformation.

In this paper, we propose a new method that uses a deep neural network (DNN) to construct the mapping relationship between the spectral envelopes of source and target speakers. The proposed four-layer DNN is trained layer-by-layer from a cascade of a Bernoulli BAM (BBAM) and two RBMs. The RBMs are employed to model the distributions of spectral envelopes of the source and target speakers respectively. The BBAM is employed to model the joint distribution of hidden variables extracted from the two RBMs. Previously, voice conversion using neural networks has been studied in [17], [18], where the neural networks were trained using the back propagation (BP) algorithm with the minimum mean square error (MMSE) criterion. In [17], the parameters of the network were not pre-trained but randomly initialized. In [18], generative pre-trainings were performed for speaker dependent RBMs/ deep belief networks (DBNs) and the speaker dependent networks were concatenated by a neural network with one hidden layer. However, our proposed method is different from [18] in three points: 1) the model is proposed to convert the spectral envelopes directly instead of mel-cepstra, 2) the intermediate network that connects the two RBMs is a BBAM, which is a generative model trained using the contrastive divergence (CD) algorithm [23], 3) no further fine-tuning using the BP algorithm is performed for jointly optimizing the parameters in all layers. The third point is an important feature in the proposed method. Commonly, the MMSE criterion is used for fine-tuning the DNN. However, as we observed in our experiments, this is inconsistent with human subjective perception. Smaller spectral distortion doesn't always mean better speech quality. This phenomenon has also been observed in other related research [11], [24]. Besides, for MMSE criteria, the conditional distribution is assumed to be unimodal, which is inconsistent with the multimodal nature of data distribution, because humans can speak the same text in many different ways. On the contrary, the proposed generative training considers the DNN as a stochastic neural network to model this multimodality [25]. The proposed model is a combination of RBMs and BBAMs, which can therefore take advantage of both the strong modeling and generating abilities of RBMs as well as the superiority of BBAMs in deriving the conditional distributions. All the converted spectral envelopes are generated from an RBM of the target speaker. Previous studies have revealed that in general RBMs/DBNs have the ability to generate spectral features with good speech quality [24], [26].

This paper is organized as follows. In Section II, we will briefly review the basic techniques of RBMs and BAMs. In Section III, the details of the conventional methods and our proposed method will be described. Our experimental results and discussions will be presented in Section IV. Section V gives the conclusion and discussion of future works.

II. Restricted Boltzmann Machines and Bidirectional Associative Memories

A. Restricted Boltzmann Machines

An RBM [27] is a bipartite undirected graphical model as demonstrated in Fig. 1(a). It has a two-layer structure with one visible layer corresponding to a set of visible stochastic variables $\mathbf{v} = [v_1, \cdots, v_V]^\top$ and one hidden layer corresponding to a set of hidden stochastic variables $\mathbf{h} = [h_1, \cdots, h_H]^\top$, where V and H denote the numbers of units in the visible and hidden layers respectively. The RBM is also a type of stochastic neural network, which is usually considered as a probabilistic model. This paper adopts RBMs to model the distributions of spectral envelopes, which are continuous and real-valued. Therefore, a Gaussian-Bernoulli RBM (GBRBM) is employed. Since only the GBRBM is involved in this paper, it is called RBM for short. The units in the visible layer of the RBM represent Gaussian stochastic variables, while those in the hidden layer represent Bernoulli stochastic variables. The distribution of the stochastic variables $\mathbf{v}$ described by the RBM is defined by an energy function

$$E(\mathbf{v}, \mathbf{h}) = \sum_{i=1}^{V} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j=1}^{H} b_j h_j - \sum_{i=1}^{V}\sum_{j=1}^{H} \frac{v_i}{\sigma_i} w_{ij} h_j, \quad (1)$$

where $\mathbf{W} = \{w_{ij}\} \in \mathcal{R}^{V \times H}$ is the weight matrix that interacts between the visible and hidden layers, $\mathbf{a} = [a_1, \cdots, a_V]^\top$ and $\mathbf{b} = [b_1, \cdots, b_H]^\top$ are the bias terms of the visible and hidden layers respectively, $\mathbf{\Sigma} = \mathrm{diag}\{\sigma_1^2, \cdots \sigma_V^2\}$ is usually fixed to the diagonal covariance matrix of the training samples and does not need to be estimated during the model training [28]. Therefore the parameter set of an RBM is $\theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}\}$.

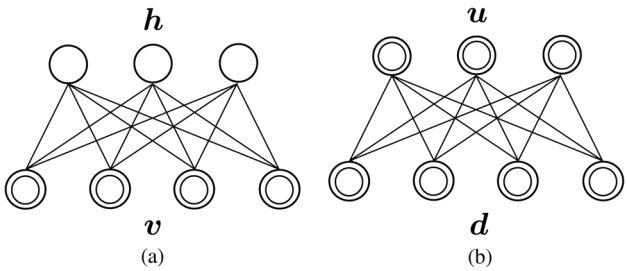

Fig. 1. The graphical model representations for (a) an RBM and (b) a BAM. The double circles represent visible units while the single circles represent hidden units.

As an energy based model (EBM), the joint distribution over $\mathbf{v}$ and $\mathbf{h}$ is defined by the energy function (1) as

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}}\exp\{-E(\mathbf{v},\mathbf{h})\}, \tag{2}$$

where

$$\mathcal{Z} = \sum_{\mathbf{h}}\int \exp\{-E(\mathbf{v},\mathbf{h})\}d\mathbf{v} \tag{3}$$

is the partition function, which is an intractable computation. However in practice, $\mathcal{Z}$ can be approximated using the annealed importance sampling (AIS) algorithm [29]. The distribution of $\mathbf{v}$ described by the RBM can be written as

$$P(\mathbf{v}) = \frac{1}{\mathcal{Z}}\sum_{\mathbf{h}}\exp\{-E(\mathbf{v},\mathbf{h})\}. \tag{4}$$

The parameters of the RBM can be estimated by maximum likelihood (ML) learning using the CD algorithm [23].

The probabilistic distribution given by an RBM can be written as

$$P(\mathbf{v}) = \sum_{\mathbf{h}} P(\mathbf{v}|\mathbf{h})P(\mathbf{h}), \tag{5}$$

where

$$P(\mathbf{v}|\mathbf{h}) = \mathcal{N}(\mathbf{v};\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{W}\mathbf{h}+\mathbf{b},\boldsymbol{\Sigma}), \tag{6}$$

$$P(\mathbf{h}) = \int_{\mathbf{v}} P(\mathbf{v},\mathbf{h})d\mathbf{v}. \tag{7}$$

It can be seen that an RBM is equivalent to a GMM whose components are determined by $\mathbf{h}$. Since there are 2^H possible combinations of the variables in $\mathbf{h}$, an RBM is equivalent to a GMM with 2^H structured components, whose weights are given by $P(\mathbf{h})$, and a global diagonal covariance matrix which is shared among all components.

B. Bidirectional Associative Memories

A BAM [30] is another type of two-layer stochastic neural network [30]. As shown in Fig. 1(b), the topological structure of the graphical model of a BAM resembles that of an RBM except that there are no hidden units in a BAM. The units in the upper-layer denote a set of stochastic variables $\mathbf{u} = [u_1,\cdots,u_U]^\top$ while those in the lower layer denote another set of stochastic variables $\mathbf{d} = [d_1,\cdots,d_D]^\top$, where U and D denote the numbers of units in the upper and lower layers respectively. Similarly, the BAM can be considered as a probabilistic model defined by an energy function. In this paper, according to the different assumptions on the distributions of the stochastic variables, two types of BAMs are involved: the BBAM and the GBAM.

In the BBAM, the stochastic variables are assumed to obey the Bernoulli distribution. Therefore, the energy function of the BBAM can be written as

$$E(\mathbf{d},\mathbf{u}) = -\mathbf{d}^\top\mathbf{a} - \mathbf{u}^\top\mathbf{b} - \mathbf{d}^\top\mathbf{W}\mathbf{u}, \tag{8}$$

where $\eta = \{\mathbf{W},\mathbf{a},\mathbf{b}\}$ is the parameter set of the BAM, $\mathbf{W}\in\mathcal{R}^{D\times U}$ is the weight matrix, $\mathbf{a}\in\mathcal{R}^{D\times 1}$ and $\mathbf{b}\in\mathcal{R}^{U\times 1}$ are the bias terms. On the other hand, in the cases of modeling continuous real-valued stochastic variables, for which the Gaussian assumptions are usually made, the GBAM is adopted. Accordingly, the energy function for a GBAM can be written as

$$E(\mathbf{d},\mathbf{u}) = \sum_{i=1}^{D}\frac{(d_i-a_i)^2}{2\sigma_{d,i}^2} + \sum_{j=1}^{U}\frac{(u_j-b_j)^2}{2\sigma_{u,j}^2} - \sum_{i=1}^{D}\sum_{j=1}^{U} w_{ij}\frac{d_i}{\sigma_{d,i}}\frac{u_j}{\sigma_{u,j}}. \tag{9}$$

Similarly the parameters $\boldsymbol{\Sigma}_u = \text{diag}\{\sigma_{u,1}^2,\cdots,\sigma_{u,U}^2\}$ and $\boldsymbol{\Sigma}_d = \text{diag}\{\sigma_{d,1}^2,\cdots,\sigma_{d,D}^2\}$ are fixed to the diagonal covariance matrices of the training samples for $\mathbf{u}$ and $\mathbf{d}$ respectively.

Conventionally, a BAM is considered as a memory for storing and recalling paired data $(\mathbf{d},\mathbf{u})$. The training criterion for BAMs is to minimize the energy defined by (8) or (9). The parameters of a BAM are estimated using Hebbian learning [30], or some improved learning algorithms such as pseudo-relaxation learning algorithm for BAM (PRLAB) [31], quick learning for BAM (QLBAM) [32], etc.

III. Spectral Envelope Conversion Using DNNs

A. A General Framework of Statistical Spectral Conversion

At first, a general framework of statistical spectral conversion method is briefly reviewed. Let $\mathbf{x} = [\mathbf{x}_1^\top,\cdots\mathbf{x}_T^\top]^\top$ and $\mathbf{y} = [\mathbf{y}_1^\top,\cdots\mathbf{y}_T^\top]^\top$ be the aligned T-frame observation sequences of spectral features for the source and target speakers respectively. Each frame in the sequence consists of the static, velocity and acceleration components, e.g.

$$\mathbf{x}_t = \left[\mathbf{x}_t^{(s)\top},\Delta\mathbf{x}_t^{(s)\top},\Delta^2\mathbf{x}_t^{(s)\top}\right]^\top, \tag{10}$$

$$\mathbf{y}_t = \left[\mathbf{y}_t^{(s)\top},\Delta\mathbf{y}_t^{(s)\top},\Delta^2\mathbf{y}_t^{(s)\top}\right]^\top, \tag{11}$$

where $\mathbf{x}_t^{(s)}$ and $\mathbf{y}_t^{(s)}$ are the static features at t-th frame, $\Delta(\cdot)$ and $\Delta^2(\cdot)$ are the corresponding velocity and accelerate features. The observation sequences $\mathbf{x}$ and $\mathbf{y}$ are fixed linear transformations of the static sequences $\mathbf{x}^{(s)} = [\mathbf{x}_1^{(s)\top},\cdots\mathbf{x}_T^{(s)\top}]^\top$ and $\mathbf{y}^{(s)} = [\mathbf{y}_1^{(s)\top},\cdots\mathbf{y}_T^{(s)\top}]^\top$, e.g. $\mathbf{x} = \mathbf{M}\mathbf{x}^{(s)}$, where $\mathbf{M}\in$

$\mathcal{R}^{3TN\times TN}$ is a fixed matrix [33] and N is the dimensionality of static feature vectors.

The statistical spectral conversion methods consist of two stages: the training stage and the conversion stage. At the training stage, a generative model $\lambda^{(v)}$ is constructed to describe the distribution of the joint spectral space. The features of the joint space are concatenations of the time-aligned source and target spectral features, which are $\mathbf{v} = [\mathbf{x}^\top, \mathbf{y}^\top]^\top$. Then, during the conversion stage, for converting a source sequence $\mathbf{x}$, a sequence of conditional distributions $\{P(\mathbf{y}_t|\mathbf{x}_t, \lambda^{(v)})\}$ can be derived from the joint distribution frame-by-frame. As the last step before synthesizing speech using the vocoder, the maximum output probability parameter generation (MOPPG) algorithm is applied to generate a sequence of converted static spectral features $\tilde{\mathbf{y}}^{(s)}$ at sentence-level as

$$\tilde{\mathbf{y}}^{(s)} = \arg\max_{\mathbf{y}^{(s)}} P\left(\mathbf{y}|\mathbf{x}, \lambda^{(v)}\right), \tag{12}$$

$$= \arg\max_{\mathbf{y}^{(s)}} \prod_{t=1}^{T} P\left(\mathbf{y}_t|\mathbf{x}_t, \lambda^{(v)}\right),$$

$$s.t. \quad \mathbf{y} = \mathbf{M}\mathbf{y}^{(s)}. \tag{13}$$

Therefore, the difference among the spectral conversion methods using different generative models exists in the derived conditional distributions.

B. Spectral Conversion Based on JDGMM

In the training stage of the conventional JDGMM-based methods, a GMM is employed as the generative model to describe the distribution of the joint spectral feature space. The distribution of the GMM is defined as

$$P(\mathbf{v}_t) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\mathbf{v}_t; \boldsymbol{\mu}_m^{(v)}, \boldsymbol{\Sigma}_m^{(v)}\right), \sum_{m=1}^{M} \alpha_m = 1, \tag{14}$$

where $\mathcal{N}(\cdot)$ denotes a Gaussian distribution, α_m, $\boldsymbol{\mu}_m^{(v)}$ and $\boldsymbol{\Sigma}_m^{(v)}$ are the weight, mean vector and covariance matrix of the m-th Gaussian component, and

$$\boldsymbol{\mu}_m^{(v)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(v)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}. \tag{15}$$

$\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\mu}_m^{(y)}$ are the mean vectors of the m-th mixture for source and target speakers' spaces, $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ are the corresponding covariance matrices. $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$ are the cross covariance matrices. The diagonal covariance matrices for source and target speaker's space and the cross-diagonal covariance matrices, i.e. $\boldsymbol{\Sigma}_m^{(xx)}$, $\boldsymbol{\Sigma}_m^{(yy)}$, $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ being diagonal [11], are usually adopted to reduce the complexity of the model. This kind of covariance matrices are suitable for modeling the mel-cepstra with weak inter-dimensional correlations.

At the conversion stage, given the input $\mathbf{x}_t$, the conditional distribution $P(\mathbf{y}_t|\mathbf{x}_t, \lambda^{(v)})$ derived from the JDGMM can also be described by a GMM with M components

$$P(\mathbf{y}_t|\mathbf{x}_t, \lambda^{(v)}) = \sum_{m=1}^{M} \beta_{m,t} \mathcal{N}\left(\mathbf{y}_t; \boldsymbol{\mu}_{m,t}^{(y|x)}, \boldsymbol{\Sigma}_m^{(y|x)}\right), \tag{16}$$

where

$$\beta_{m,t} = P\left(m|\mathbf{x}_t, \lambda^{(v)}\right), \tag{17}$$

$$\boldsymbol{\mu}_{m,t}^{(y|x)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)-1}\left(\mathbf{x}_t - \boldsymbol{\mu}_m^{(x)}\right), \tag{18}$$

$$\boldsymbol{\Sigma}_m^{(y|x)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)-1}\boldsymbol{\Sigma}_m^{(xy)}, \tag{19}$$

are the weight, mean vector and covariance matrix of the m-th mixture. In practice, a sequence of sub-optimal Gaussian distributions is usually adopted to improve the conversion efficiency without obviously degrading the performance [11]. The Gaussian component with the maximal posteriori probability $\tilde{m}_t = \arg\max_m \beta_{m,t}$ is selected from the conditional GMM for each frame. The distribution of the conditional GMM for each frame is approximated with a single Gaussian. According to (13), the spectral feature sequence can be generated with a closed-form solution

$$\tilde{\mathbf{y}}^{(s)} = (\mathbf{M}^\top\mathbf{U}^{-1}\mathbf{M})^{-1}\mathbf{M}^\top\mathbf{U}^{-1}\mathbf{E}, \tag{20}$$

in which $\mathbf{E} = [\boldsymbol{\mu}_{\tilde{m}_1,1}^{(y|x)^\top}, \cdots, \boldsymbol{\mu}_{\tilde{m}_T,T}^{(y|x)^\top}]^\top$ and $\mathbf{U} = \text{diag}\{\boldsymbol{\Sigma}_{\tilde{m}_1}^{(y|x)}, \cdots, \boldsymbol{\Sigma}_{\tilde{m}_T}^{(y|x)}\}$ are the concatenations of all mean vectors and diagonal covariance matrices of the conditional Gaussian distributions. It can be seen from (18) and (20) that, using the sub-optimal mixture sequence, the converted feature sequence $\tilde{\mathbf{y}}^{(s)}$ is a linear transformation of the input static feature sequence $\mathbf{x}^{(s)}$.

C. Spectral Conversion Based on MoRBM

In [20], a spectral envelope conversion method using MoRBM was introduced. The MoRBM is constructed by replacing each Gaussian component in the JDGMM with an RBM. For the joint spectral modeling, the parameters of the RBM θ_m for the m-th mixture component can be written as $\mathbf{W}_m = [\mathbf{W}_{m,x}^\top, \mathbf{W}_{m,y}^\top]^\top$ and $\mathbf{a}_m = [\mathbf{a}_{m,x}^\top, \mathbf{a}_{m,y}^\top]^\top$, where $(\cdot)_x$ and $(\cdot)_y$ denote the parameters corresponding to the source and target speaker respectively.

During conversion, in order to use the conventional MOPPG algorithm, the conditional distribution derived from the joint space RBM θ_m is approximated by a single Gaussian distribution

$$P(\mathbf{y}_t|\mathbf{x}_t, \theta_m) \simeq \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_{m,t}, \boldsymbol{\Sigma}_{m,y}), \tag{21}$$

where

$$\boldsymbol{\mu}_{m,t} = \arg\max_{\mathbf{y}_t} P(\mathbf{x}_t, \mathbf{y}_t|\theta_m) \tag{22}$$

is the mode of the conditional distribution derived from θ_m and $\boldsymbol{\Sigma}_{m,y}$ is the diagonal covariance matrix of the training samples for the m-th mixture component. $\boldsymbol{\mu}_{m,t}$ can be solved by the gradient descent algorithm. The first derivative for iteratively updating $\boldsymbol{\mu}_{m,t}$ is given by

$$\frac{\partial \log P(\mathbf{x}_t, \mathbf{y}_t|\theta_m)}{\partial \mathbf{y}_t} = \boldsymbol{\Sigma}_{m,y}^{-\frac{1}{2}}\mathbf{W}_{m,y}g(\mathbf{s}) - \boldsymbol{\Sigma}_{m,y}^{-1}(\mathbf{y}_t - \mathbf{a}_y), \tag{23}$$

where $g(\cdot)$ is the element-wise sigmoid function, and

$$\mathbf{s} = \mathbf{x}_t^\top\boldsymbol{\Sigma}_{m,x}^{-\frac{1}{2}}\mathbf{W}_{m,x} + \mathbf{y}_t^\top\boldsymbol{\Sigma}_{m,y}^{-\frac{1}{2}}\mathbf{W}_{m,y} + \mathbf{b}_m. \tag{24}$$

Before applying the gradient descent algorithm, the mode of the joint distribution RBM θ_m is estimated using the method introduced in [24]. Then the portion of the estimated mode which corresponds to the target speaker is adopted to initialize the gradient descent optimization in order to solve (22). An approximation is also made to reduce the computational cost during conversion by replacing the sigmoid function $g(\cdot)$ with a step function [20].

D. Spectral Conversion Based on MoGBAMs

In [21], we have introduced a spectral envelope conversion method using MoGBAM. The energy function (9) of a GBAM can be simply written as

$$E(\mathbf{d},\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top \boldsymbol{\Sigma}_u^{-1}\mathbf{u} + \frac{1}{2}\mathbf{d}^\top \boldsymbol{\Sigma}_d^{-1}\mathbf{d} - \mathbf{d}^\top \mathbf{W}\mathbf{u} \tag{25}$$

by omitting the bias terms. We can see that the distribution defined by the energy function (25) can be reformulated into a Gaussian-like form with zero mean

$$P(\mathbf{v}) = \frac{1}{\mathcal{Z}} \exp\left\{-\frac{1}{2}\mathbf{v}^\top \mathbf{P}\mathbf{v}\right\}, \tag{26}$$

where $\mathbf{v} = [\mathbf{d}^\top, \mathbf{u}^\top]^\top$ is the joint stochastic variable of $\mathbf{u}$ and $\mathbf{d}$. This distribution is exactly equivalent to a Gaussian distribution if the precision matrix

$$\mathbf{P} = \boldsymbol{\Sigma}_v^{-\frac{1}{2}} \begin{bmatrix} \mathbf{I}_{D\times D} & -\mathbf{W} \\ -\mathbf{W}^\top & \mathbf{I}_{U\times U} \end{bmatrix} \boldsymbol{\Sigma}_v^{-\frac{1}{2}} \tag{27}$$

is positive definite (which means $|\mathbf{I} - \mathbf{W}^\top\mathbf{W}| > 0$), where $\boldsymbol{\Sigma}_v = \mathrm{diag}\{\boldsymbol{\Sigma}_d, \boldsymbol{\Sigma}_u\}$ is the diagonal covariance matrix of the joint stochastic variable $\mathbf{v}$.

Therefore, for the spectral envelope modeling, the GBAMs can model the inter-dimensional correlations because the covariance matrix $\mathbf{P}^{-1}$ of the Gaussian distribution described by the GBAM is a full matrix. For the joint spectral envelope modeling, the stochastic variables in the upper and lower layers represent the acoustic features of source and target speakers respectively. Similar to the MoRBM introduced in Section III-C, the MoGBAM used here is constructed by replacing each Gaussian component in the JDGMM with a GBAM. Since the conditional distributions derived from GBAMs are single Gaussian distributions with diagonal covariance matrices

$$P(\mathbf{u}|\mathbf{d}) = \mathcal{N}\left(\mathbf{u}; \boldsymbol{\Sigma}_u^{\frac{1}{2}}\mathbf{W}^\top \boldsymbol{\Sigma}_d^{-\frac{1}{2}}\mathbf{d}, \boldsymbol{\Sigma}_u\right), \tag{28}$$

the spectral envelope conversion is exactly the same as that of the JDGMM-based method.

The rest of this section focuses on the training of BAMs. As introduced in Section II-B, there are many conventional algorithms for training BAMs. But in these conventional algorithms, the BAM is not considered to be a probabilistic model as we do in this paper. Therefore, unlike conventional training methods, the ML learning is employed for training BAMs in this paper. It is straightforward to employ the gradient descent algorithm to estimate the parameters of a GBAM because the distribution has an analytic expression for the partition function. But this algorithm could make the model easily become over-fitted because of the huge number of parameters in the GBAM for high-dimensional spectral envelope modeling. On the other hand, the partition function of the distribution described by the BBAM is intractable. Therefore, the gradient descent algorithm cannot be directly applied to train BBAMs.

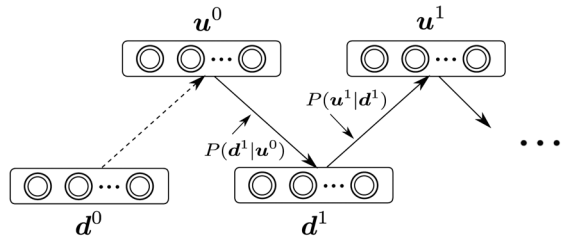

Fig. 2. The Gibbs chain for training a BAM, the dashed arrow indicates that the first step of the chain is given by the training data pair.

To cope with this problem, similar to the training of the Gaussian RBMs [34], the CD algorithm is employed for the ML training of both the GBAM and the BBAM in this paper. Specifically for the GBAM, the gradient descent algorithm is employed to update the parameters. The first derivative for updating $\mathbf{W}$ is derived as

$$\delta\mathbf{W} = \boldsymbol{\Sigma}_d^{-\frac{1}{2}}\left(E_s\left[\mathbf{d}\mathbf{u}^\top\right] - E_m\left[\mathbf{d}\mathbf{u}^\top\right]\right)\boldsymbol{\Sigma}_u^{-\frac{1}{2}}, \tag{29}$$

where $E_s[\cdot]$ denotes the expectation with respect to the distribution of the training samples and $E_m[\cdot]$ denotes the expectation with respect to the distribution defined by the model [23]. A Gibbs chain is run for approximating $E_m[\cdot]$ as illustrated in Fig. 2. Benefiting from the conditional independence between units in the same layer, the sampling process can be run iteratively between the two layers. The difference from the CD algorithm used in the training of RBMs is that the first step of sampling $\mathbf{u}$ given $\mathbf{d}$ is determined by the joint training samples.

To be specific, the Gibbs chain sampling is run as follows:

1) Initialize the Gibbs chain with the training sample pair: $(\mathbf{d}^0, \mathbf{u}^0) \leftarrow (\mathbf{d}, \mathbf{u})$, and set the sampling step $k \leftarrow 1$,
2) At the k-th step, firstly draw a sample $\mathbf{d}^k$ from the conditional distribution

$$P(\mathbf{d}^k|\mathbf{u}^{k-1}) = \mathcal{N}\left(\mathbf{d}^k; \boldsymbol{\Sigma}_d^{\frac{1}{2}}\mathbf{W}\boldsymbol{\Sigma}_u^{-\frac{1}{2}}\mathbf{u}^{k-1}, \boldsymbol{\Sigma}_d\right). \tag{30}$$

Then, draw a sample $\mathbf{u}^k$ from the corresponding conditional distribution

$$P(\mathbf{u}^k|\mathbf{d}^k) = \mathcal{N}\left(\mathbf{u}^k; \boldsymbol{\Sigma}_u^{\frac{1}{2}}\mathbf{W}^\top\boldsymbol{\Sigma}_d^{-\frac{1}{2}}\mathbf{d}^k, \boldsymbol{\Sigma}_u\right). \tag{31}$$

3) Set $k \leftarrow k+1$ and return to 2) to continue running the Gibbs chain.

When $k = +\infty$, the CD algorithm achieves the exact ML learning [29]. But considering the computational efficiency, one-step Gibbs sampling is adopted as it works well in the training of RBMs. The mean-field approximation [35] is adopted to sample data on the Gibbs chain, e.g., $\mathbf{d}^k = E[P(\mathbf{d}^k|\mathbf{u}^{k-1})]$, where $E[\cdot]$ denotes the expectation of a distribution.

In order to make the precision matrix $\mathbf{P}$ positive definite, a regularization term $\mathrm{tr}\{\mathbf{W}\mathbf{W}^\top\}$ is added in the likelihood function [34]. This regularization term also has the same effect as the weight decay that can prevent the model from over-fitting. And further, the momentum, which considers the last update of

the parameter, is also adopted to accelerate the updating of parameters [36]. Therefore, the updating formula for the weight matrix $\mathbf{W}$ at the i-th step is given by

$$\mathbf{W}^{i+1} = \mathbf{W}^i + \Delta\mathbf{W}^i, \tag{32}$$

where

$$\Delta\mathbf{W}^i = \epsilon\Delta\mathbf{W}^{i-1} + \gamma(\delta\mathbf{W}^i + \rho\mathbf{W}^i), \tag{33}$$

γ is the learning rate, ϵ and ρ are hyper-parameters for momentum and regularization terms respectively, $\Delta\mathbf{W}$ is initialized with zero ($\Delta\mathbf{W}^0 = \mathbf{0}$).

E. Spectral Envelope Conversion Using GTDNNs

As discussed in Section III-C and III-D, RBMs and GBAMs can be used to model the distributions of joint spectral envelopes instead of Gaussian distributions. But both of these two models have their limitations in spectral envelope conversion. The advantages of the RBM are obvious. It has a strong ability in modeling the distribution of spectral envelopes. It can also behave as a feature extractor that can effectively extract high-order distributed binary representations [37] for raw spectral envelopes. But it is not straightforward to derive the conditional distributions from the joint distributions described by the RBMs for conversion. Some inappropriate approximations are made during the conversion stage as introduced in Section III-C. The performance of MoRBM based methods could be degraded by those approximations. On the other hand, the MoGBAM method is superior to the GMM and MoRBM based methods in some aspects. Different from RBMs, GBAMs have no hidden units and therefore it is straightforward to derive the conditional distributions from GBAMs. Although GBAMs can effectively model the inter-dimensional correlations, they are equivalent to Gaussian distributions whose modeling ability is limited for complex distributions. Besides, the mapping relationship constructed by MoGBAM is still a piece-wise linear transformation.

In this section, in order to take advantage of both models and avoid their disadvantages, we propose a new model to combine BAMs and RBMs. As illustrated in the right part of Fig. 3, the proposed model is a four-layer feedforward DNN, including an input layer, an output layer and two hidden layers. The input and output layers denote the stochastic variables in the spectral envelope vectors of source and target speakers respectively. The parameter set of the proposed DNN is defined as

$$\psi = \{\mathbf{W}_x, \mathbf{W}_y, \mathbf{W}_h, \mathbf{b}_x, \mathbf{b}_y, \mathbf{b}_{h_x}, \mathbf{b}_{h_y}\}, \tag{34}$$

where $\mathbf{W}_x \in \mathcal{R}^{3N\times H_x}$, $\mathbf{W}_h \in \mathcal{R}^{H_x\times H_y}$ and $\mathbf{W}_y \in \mathcal{R}^{3N\times H_y}$ are the weight matrices that interact between the input and first hidden layers, between the two hidden layers, and between the output and second hidden layers respectively, $\mathbf{b}_x \in \mathcal{R}^{3N\times 1}$, $\mathbf{b}_{h_x} \in \mathcal{R}^{H_x\times 1}$, $\mathbf{b}_{h_y} \in \mathcal{R}^{H_y\times 1}$ and $\mathbf{b}_y \in \mathcal{R}^{3N\times 1}$ are the bias terms of the corresponding layers. H_x and H_y are the number of units in the two hidden layers.

Unlike the conventional training algorithm for the feedforward neural networks, which is the BP algorithm using the MMSE criterion for regression tasks such as spectral conversion [17], the proposed DNN is trained layer-by-layer by generative learning. At the training stage, as illustrated in Fig. 3, each two

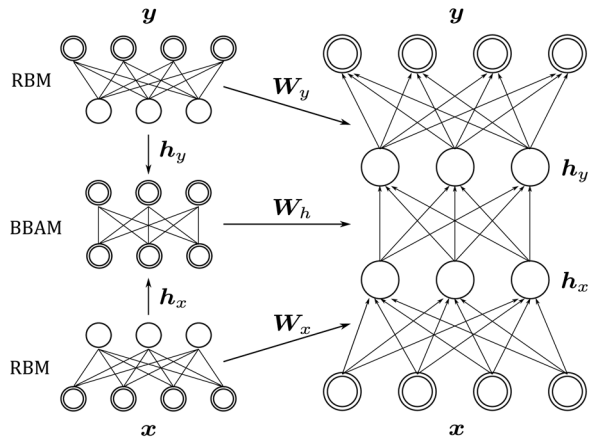

Fig. 3. Illustration of the construction of the proposed DNN. The DNN (right) is trained layer-by-layer from three stochastic two-layer neural networks (shown on the left).

directly connected layers are considered as a undirected stochastic neural network. In other words, two RBMs are adopted to model the distributions of spectral envelopes for the source and target speakers respectively. Then, a BBAM is employed to model the joint distribution of the hidden variables extracted from the two RBMs. The extracted hidden variables can be considered as the high-order representations of the spectral envelopes. It is worthwhile to note that the proposed network can be easily extended to a deeper network by replacing the RBMs with deep stochastic neural networks, such as DBNs or deep Boltzmann machines (DBMs) [38].

To be more specific, the training process of the proposed DNN is conducted in three steps as follows:

1) Train an RBM $\theta_x = \{\mathbf{W}_x, \mathbf{b}_x, \mathbf{b}_{h_x}\}$ using the training data $\mathbf{x}$ of the spectral envelopes of the source speaker. Then given the visible samples $\mathbf{x}$, draw their corresponding hidden samples $\mathbf{h}_x$ from the conditional distribution derived from θ_x, which is
$$P(h_{x,t,i} = 1|\mathbf{x}_t, \theta_x) = g\left(\mathbf{x}_t^{\mathrm{T}}\mathbf{w}_{x,*i} + b_{h_x,i}\right), \tag{35}$$
where $\mathbf{w}_{x,*i}$ is the i-th column vector in $\mathbf{W}_x$.
2) Similarly train an RBM $\theta_y = \{\mathbf{W}_y, \mathbf{b}_y, \mathbf{b}_{h_y}\}$ using the training data $\mathbf{y}$. And then draw samples $\mathbf{h}_y$ given $\mathbf{y}$ from the conditional distribution derived from θ_y
$$P(h_{y,t,j} = 1|\mathbf{y}_t, \theta_y) = g(\mathbf{y}_t^{\top}\mathbf{w}_{y,*j} + \mathbf{b}_{h_y}), \tag{36}$$
where $\mathbf{w}_{y,*j}$ is the j-th column vector in $\mathbf{W}_y$.
3) In the last step, a BBAM $\eta_h = \{\mathbf{W}_h, \mathbf{b}_{h_x}, \mathbf{b}_{h_y}\}$ is trained using the drawn data $\mathbf{h}_x$ and $\mathbf{h}_y$, which are parallel because the visible samples $\mathbf{x}$ and $\mathbf{y}$ are time aligned. Note that the bias terms are fixed to those estimated in 1) and 2), and are not updated in this step.

The learning algorithm for the BBAM in step 3) is similar to that of the GBAM as introduced in Section III-D. In the BBAM, the conditional distributions for sampling data on the Gibbs chain when performing CD algorithm are Bernoulli distributions

$$P(h_{y,j} = 1|\mathbf{h}_x, \eta_h) = g(\mathbf{h}_x^{\top}\mathbf{w}_{h,*j} + b_{h_y,j}), \tag{37}$$
$$P(h_{x,i} = 1|\mathbf{h}_y, \eta_h) = g(\mathbf{w}_{h,i*}\mathbf{h}_y + b_{h_x,i}), \tag{38}$$

where $\mathbf{w}_{h,i*}$ and $\mathbf{w}_{h,*j}$ are the i-th row vector and j-th column vector in $\mathbf{W}_h$.

The conventional training process of a DNN consists of two steps: the layer-by-layer unsupervised generative pre-training and the joint fine-tuning of all layers. However, the MMSE criterion may not be suitable for training the DNN for the purpose of generating speech spectra. Firstly, we observed in our experiments that the spectral distortion, which is the cost function of the MMSE criterion, is not consistent with human perception. Similar phenomenon have also been reported in other research works [11], [24]. Secondly, one assumption of the MMSE criterion is that the predictive conditional distribution is a single Gaussian distribution, which is not consistent with the multimodal nature of training data. In this paper, only layer-by-layer generative training is adopted for training the proposed DNN. Therefore the estimated DNN is a stochastic neural network, which can theoretically predict a conditional distributions with up to 2^{H_y} Gaussian components. Unsupervised training is employed for training the two RBMs. Then supervised training is employed for training the intermediate BBAM using the parallel high-order representations extracted by the RBMs. Therefore, even without further joint fine-tuning of all layers, the proposed GTDNN models the mapping relationship between the spectral envelopes of the two speakers by the joint modeling of the hidden variables using the BBAM. The patterns embedded in the high-order binary representations are considered to be relatively simpler than those in the spectral envelopes. Therefore one BBAM is relatively sufficient to model the mapping relationship. On the other hand, the GTDNN includes multiple non-linear layers, thus it has the potential of better describing the highly non-linear mapping relationship compared to the piece-wise linear transformations. Therefore, it is not necessary to construct a mixture of GTDNNs like we have done in the MoRBM and MoGBAM based methods. A global DNN is constructed for the spectral envelope conversion between two speakers.

During the conversion phase, for an input frame $\mathbf{x}_t$, the conditional distribution of the output $\mathbf{y}_t$ derived from the proposed model as

$$P(\mathbf{y}_t|\mathbf{x}_t,\psi) \simeq P(\mathbf{y}_t|\tilde{\mathbf{h}}_{y,t},\theta_y), \tag{39}$$

where

$$\tilde{\mathbf{h}}_{y,t} \sim P(\mathbf{h}_{y,t}|\tilde{\mathbf{h}}_{x,t},\eta_h), \tag{40}$$

$$\tilde{\mathbf{h}}_{x,t} \sim P(\mathbf{h}_{x,t}|\mathbf{x}_t,\theta_x), \tag{41}$$

are the hidden samples drawn from the corresponding conditional distributions. Two strategies are adopted and compared in this paper. The first one is mean-field approximation, which considers this model exactly as a conventional feedforward DNN. Another one samples data according to the maximal likelihoods of the model generating them, e.g., $\tilde{\mathbf{h}}_{x,t}$ is sampled as

$$\tilde{\mathbf{h}}_{x,t} = \arg\max_{\mathbf{h}_{x,t}} P(\mathbf{h}_{x,t}|\mathbf{x}_t,\theta_x), \tag{42}$$

to be more specific, the sampling procedure runs as follows

$$\tilde{h}_{x,t,i} = \begin{cases} 1 & P(h_{x,t,i}=1|\mathbf{x}_t,\theta_x) \geq 0.5, \\ 0 & o.w. \end{cases} \tag{43}$$

$\tilde{\mathbf{h}}_{y,t}$ is obtained in the same manner.

Therefore, the conditional distribution can be approximated by a single Gaussian distribution with diagonal covariance matrix, which is

$$P(\mathbf{y}_t|\mathbf{x}_t,\psi) \propto \mathcal{N}(\mathbf{y}_t;\boldsymbol{\Sigma}_y^{\frac{1}{2}}(\mathbf{W}_y^\top \tilde{\mathbf{h}}_{y,t}+\mathbf{b}_y),\boldsymbol{\Sigma}_y), \tag{44}$$

where $\boldsymbol{\Sigma}_y = \mathrm{diag}\{\sigma_{y,1}^2,\cdots,\sigma_{y,3N}^2\}$ is the global shared diagonal covariance matrix. The parameter generation algorithm in the conventional methods can be applied to generate the converted spectral envelopes without any modification.

IV. Experiments

A. Experimental Conditions

A Chinese speech database was used in our experiments. This database contains four professional speakers, including two male speakers (M1 and M2) and two female speakers (F1 and F2). Four pairs of conversions were used in our experiments to evaluate the performance of the proposed method, including two inter-gender conversions (F1-M1 and M2-F2) and two intra-gender conversions (M1-M2 and F2-F1), 100 parallel sentences were adopted for each speaker, from which 80 sentences were selected randomly as the training set and the remaining 20 sentences were used as the test set. The speech waveforms were recorded in 16 kHz/16 bit format.

The baseline system in our experiment is the conventional JDGMM-based system. The 40-order mel-cepstra (not including the 0-th coefficients for frame powers) were adopted as the spectral features for the baseline system. The mel-cepstra were extracted from the spectral envelopes calculated by the STRAIGHT vocoder [39]. The spectral envelopes were also directly used as the spectral features in the proposed method and some of the other methods chosen for comparison. The FFT length of the STRAIGHT analysis was set to 1024, which leads to a 513 dimensional spectral envelope vector for each frame. Therefore, the dimensionality of the mel-cepstra and spectral envelopes, including the static, velocity and acceleration components, were $40 \times 3 = 120$ and $513 \times 3 = 1539$ respectively. The frame shift for calculating spectral envelopes was set to 5 ms. The training samples for joint space were constructed by performing time alignment between the parallel mel-cepstral sequences of source and target speakers using the dynamic time warping (DTW) algorithm. Joint spectral envelopes were constructed using the DTW paths derived from their corresponding mel-cepstra.

In the spectral envelope modeling, the spectral amplitudes of all frames were normalized to the same energy. And logarithmic spectral amplitudes were adopted as the spectral features for modeling. We observed in our experiments that the energy normalization can slightly improve the performance. At conversion time, the energies of input spectral envelopes were retained to recover energies in converted spectral envelopes. The CD learning with 1-step Gibbs sampling (CD-1) was employed to train RBMs and BAMs. A GPU (NVIDIA GeForce GTX TITAN) was used to accelerate the training procedure. The stochastic batch gradient descent algorithm was adopted to update the model parameters. The size of each mini-batch was set to 10 and the learning rate was set to 0.0001. The hyperparameter of momentum term is set to 0.5 in beginning 5 epochs and

fixed to 0.9 after the 5-th epoch. 200 epochs were executed for estimating the parameters of each RBM and 50 epochs were executed for those of each BAM.

Since this paper focuses on the spectral conversion, the fundamental frequencies (F_0) were converted using the conventional method, which is simply a linear transformation in the log-scale of F_0s to equalize the mean and variance of converted and target $\log F_0$s [11].

B. Effectiveness of RBMs and BAMs as Probability Density Functions

At first, the effectiveness of RBMs and GBAMs as probability density functions to model the distribution of joint spectral envelope space was examined by evaluating their average log-probabilities when generating the samples in the training and test set. Both the effectivenesses on mel-cepstral and spectral envelope modeling were evaluated. In our experiments, the number of hidden units of an RBM was set to 100 for mel-cepstral modeling and 300 for spectral envelope modeling. Note that the number of hidden units in the RBMs was empirically optimized for mel-cepstral modeling, but not tuned for spectral envelope modeling. The average log-probability of the RBM on spectral envelopes was still growing when we increased the number of hidden units. 300 was adopted here considering the computational efficiency at both the training and conversion stage.

The comparison was taken among the GMM, MoRBM and MoGBAM. The first experiment was taken to verify the effectiveness of these models on different numbers of mixture components. The cross-diagonal covariance matrices were employed in the GMMs. This experiment was conducted on the conversion pair F1-M1. We varied the number of mixture components from 1 to 128. The average log-probabilities of these models on both the training and test sets are shown in Fig. 4. Examining the results of mel-cepstral modeling in Fig. 4(a), we see that the MoGBAM doesn't show clear superiority over the GMM, while the MoRBM shows some, but not significant advantages. In the case of modeling spectral envelopes, we see from Fig. 4(b) that both MoRBM and MoGBAM show significant improvements in average log-probabilities over the GMM in test set. The MoRBM still gives the highest average log-probabilities among the three models.

As a probability density model, an RBM is a product of experts (PoE) model [40] and it can be considered to be equivalent to a GMM with 2^H structured components. Therefore it has the ability of analyzing the latent patterns embedded in the spectral envelopes. RBMs can thus generate the training and test samples with higher probabilities. On the other hand, although the GBAM is trained using the CD algorithm, which is borrowed from the training of the RBM, it is still equivalent to a single Gaussian distribution that can only represent one pattern in spectral envelope space. The superiority of the MoGBAM over GMM attributes to its full covariance modeling determined by the weight matrices.

In addition, Table I shows the average log-probabilities of GMM modeling with full covariance matrix (FGMM). We see from the results of mel-cepstra that the average log-probabilities of FGMM are higher than those of other models in the test set, although performance starts descending after 16 mixtures. In spectral envelope modeling, although the average log-probability starts decreasing from 2 mixtures, it is much higher than those of other models when the number of components is no more than 4. This indicates that the modeling abilities of MoRBMs and MoGBAMs are still insufficient in capturing the inter-dimensional correlations. However, FGMMs can only be well learned with a very small number of components, which is inadequate to effectively model the multi-modality of the phonetic structure in the spectral feature distribution. Besides, the conditional distribution of FGMM also employs full covariance matrices, which may greatly increase the computational complexity of parameter generation. Therefore diagonal covariance and cross-covariance matrices are adopted for conversion [11].

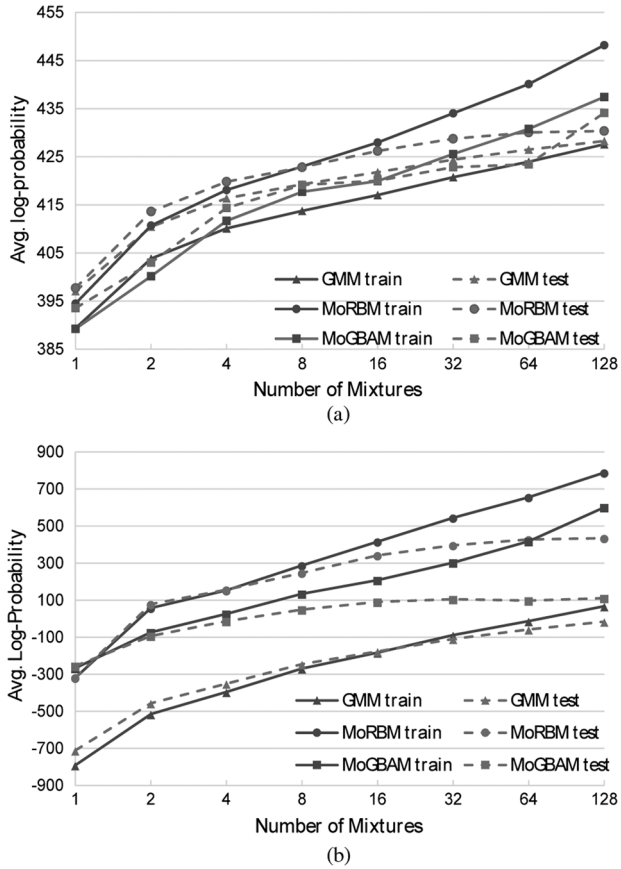

Fig. 4. The average log-probabilities of the GMM, MoRBMs and MoGBAM when generating (a) mel-cepstra and (b) spectral envelopes for the training and test sets of conversion pair F1-M1.

TABLE I
THE AVERAGE LOG-PROBABILITIES OF THE FULL COVARIANCE GMM GENERATING MEL-CEPSTRA AND SPECTRAL ENVELOPES OF THE TRAINING AND TEST SETS OF CONVERSION PAIR F1-M1

# of mix.	Mel cepstra		Spectral envelopes	
	train	test	train	test
1	421.99	426.92	9.95×10^3	9.35×10^3
2	441.01	443.27	1.02×10^4	8.74×10^3
4	451.94	451.22	1.10×10^4	4.83×10^3
8	459.28	454.20	1.18×10^4	-1.14×10^4
16	469.20	459.58	1.29×10^4	-9.81×10^4
32	480.50	458.88	–	–
64	492.58	453.15	–	–

TABLE II
THE AVERAGE LOG-PROBABILITIES OF THE GMM, MORBM AND MOGBAM WHEN GENERATING MEL-CEPSTRA FOR THE TRAINING AND TEST SETS OF ALL FOUR CONVERSION PAIRS

	GMM		MoGBAM		MoRBM	
	train	test	train	test	train	test
F1-M1	427.57	428.24	437.40	434.10	448.17	430.33
M2-F2	426.31	428.21	440.99	430.40	452.72	438.22
M1-F1	447.55	442.23	462.91	444.74	471.48	450.10
F2-F1	419.27	418.54	432.53	421.43	445.06	429.87
Avg.	430.17	429.31	443.46	432.67	454.36	437.13

TABLE III
THE AVERAGE LOG-PROBABILITIES OF THE GMM, MORBM AND MOGBAM WHEN GENERATING SPECTRAL ENVELOPES FOR THE TRAINING AND TEST SETS OF ALL FOUR CONVERSION PAIRS

	GMM		MoGBAM		MoRBM	
	train	test	train	test	train	test
F1-M1	62.29	-20.91	596.02	105.86	783.27	429.55
M2-F2	111.30	47.58	615.03	190.09	842.34	530.90
M1-F1	422.35	302.61	976.20	454.86	1144.28	710.48
F2-F1	-21.92	-54.54	465.96	140.80	725.50	441.57
Avg.	143.50	68.68	663.30	222.90	873.85	528.13

Although a GBAM is equivalent to a full covariance Gaussian, its covariance coefficients are constrained by the weights of the GBAM, and the covariance matrix of its predictive conditional distribution is diagonal. Therefore, the MoGBAM, which provides full covariance modeling with a larger number of mixture components, is a trade-off between modeling inter-dimensional correlation and multi-modality.

A second experiment was conducted to further verify the effectiveness of these models on more conversion pairs. In this experiment, the number of mixtures was set to 128, and the average log-probabilities were calculated on all four conversion pairs. Table II and Table III show the results for the mel-cepstral and spectral envelope modeling respectively. Note that the average log-probabilities of F1-M1 are the same as those shown in Fig. 4. Examining the differences among the models in all four conversion pairs, we see that the performance of the models is consistent among different conversion pairs for both mel-cepstral and spectral envelope modeling. The results are similar to those observed in the first experiment.

C. System Construction

Six systems were constructed to evaluate the spectral conversion performance of the proposed method[1]. Two systems were constructed using mel-cepstra as the spectral features:

- *GMM-MCEP*: the baseline system, which used a conventional JDGMM with 128 mixture components;
- *GMM-GV*: an extension to the baseline system, an additional GV model [11] was trained to alleviate the over-smoothing effect in mel-cepstra converted by *GMM-MCEP*;

[1]Some speech examples converted by these systems can be found at http://staff.ustc.edu.cn/ chenlh/GTDNNVC/demo.html.

The remaining four systems were constructed to directly use spectral envelopes as the spectral features:

- *GMM-SPE*: a system constructed by replacing each Gaussian component in JDGMM with a Gaussian distribution with cross-diagonal covariance matrix in spectral envelope space;
- *MoRBM*: constructed using the method introduced in Section III-C, with the number of hidden units set to 300 for each RBM;
- *MoGBAM*: constructed using the method introduced in Section III-D;
- *GTDNN*: constructed using the proposed method, with the number of units in each hidden layer set to 2048 in our experiments.

Because of the strong inter-dimensional correlations in spectral envelopes, it is infeasible to directly estimate a JDGMM on spectral envelope space using the conventional training methods. Besides, in order to be fair to compare the Gaussian distribution, RBM and GBAM, the same mixture structure was shared among the *GMM-SPE*, *MoRBM* and *MoGBAM* systems. In other words, the *JDGMM* is adopted to partition the joint space into several sub-spaces, then the distribution of spectral envelopes in each sub-space is modeled by a Gaussian distribution, an RBM and a GBAM in the *GMM-SPE*, *MoRBM* and *MoGBAM* systems respectively.

Note that in our experiments, the system using the MoRBM, MoGBAM and GTDNN for mel-cepstral conversion were not included in our subjective evaluations. This is because our informal listening tests indicated that these methods did not show significant superiority over the GMM method in mel-cepstral conversion. This can be explained by the average log-probability results in the previous section. Unlike in spectral envelope modeling, the superiority of RBMs and GBAMs over GMMs in mel-cepstral modeling are not clear as shown in Fig. 4(a). The inter-dimensional correlations of mel-cepstra are much weaker than those of spectral envelopes. Therefore, the mel-cepstral conversion cannot benefit from the advantages of the GBAMs. Likewise, the advantage of RBMs in modeling with large numbers of Gaussian components failed to achieve better performance on mel-cepstral as well. Neural networks are good at modeling the inter-dimensional correlations. The proposed GTDNN did not work better in mel-cepstra because its performance depends on the modeling of speaker dependent spaces of RBMs.

D. Comparison Between GTDNN and Fine-Tuned DNN

Before comparing with the other methods, we firstly conducted several experiments to find the best configurations for the proposed method on the F1-M1 conversion pair. Two strategies for the conversion using the GTDNN were compared, including

1) *GTDNN-mf*: the mean-field sampling was adopted in sampling data for the hidden variables in (40) and (41);
2) *GTDNN*: the sampling method described by (43) was adopted for sampling data for the hidden variables.

Secondly, the *GTDNN* is also compared with the fine-tuned DNN (*FTDNN*), which was fine-tuned using the MMSE criterion. The initialization of the network parameters was given by the GTDNN. Because the number of training samples was much

TABLE IV
LOG-SPECTRAL DISTORTIONS BETWEEN THE SPECTRAL ENVELOPES OF NATURAL SPEECH AND THOSE CONVERTED BY THE *GTDNN*, *GTDNN-mf*, *FTDNN* AND *GMM-SPE*

method	LSD (dB)
GMM-SPE	4.61
GTDNN	5.38
GTDNN-mf	4.92
FTDNN	**4.55**

TABLE V
THE RESULTS OF PREFERENCE TESTS (%) THAT COMPARE THE SIMILARITY (SIM.) AND NATURALNESS (NAT.) BETWEEN DIFFERENT SYSTEMS BASED ON DNN. N/P IS SHORT FOR "NO PREFERENCE". p IS THE p-VALUE GIVEN BY THE t-TEST FOR EXAMINING THE SIGNIFICANCE BETWEEN THE COMPARED SYSTEMS

	GTDNN	*GTDNN-mf*	*FTDNN*	N/P	p
Sim.	39.29	37.86	–	22.85	0.822
	53.57	–	32.14	14.29	0.006
	-	**42.14**	26.43	31.43	0.012
Nat.	**60.00**	27.14	–	12.86	0.000
	77.14	–	14.28	8.58	0.000
	–	**62.86**	16.43	20.71	0.000

smaller than that of the parameters in the *FTDNN*, the "dropout" [41] and weight decay [36] techniques were employed to prevent the BP algorithm from over-fitting.

At first, the average log-spectral distortions (LSDs) between the converted and target reference spectral envelopes in the test set were calculated and compared. For a reference, the average LSD of the *GMM-SPE* system was also calculated. All the spectral envelopes were normalized to the same power before calculating the LSDs. Due to the difference between the durations of the converted and the reference natural spectral envelope sequences, the DTW algorithm was employed to perform the alignment between mel-cepstra converted by *GMM-MCEP* and those extracted from reference spectral envelopes. The same alignments were used in the calculation of different systems. The results are listed in Table IV. As we can see, the average LSDs of the *GTDNN* and *GTDNN-mf* are much larger than that of the *GMM-SPE* system. After fine-tuning, the average LSD of the *FTDNN* system decreases to become slightly smaller than the *GMM-SPE* system. Because the cost function (mean square error) for the fine-tuning is directly related to the average LSD, the complex non-linear mapping function described by the DNN can reasonably generate spectral envelopes with lower average LSDs.

Next, we conducted several preference evaluations to compare the subjective performances among the *GTDNN*, *GTDNN-mf* and *FTDNN* systems. In these evaluations, all 20 sentences in the test set of the conversion pair F1-M1 were converted by the three systems. Two paired comparison listening tests were conducted. Seven Chinese-native listeners were involved in these tests. Both the similarity and naturalness were evaluated. Note that the naturalness relates mostly to the quality of the speech because the natural prosodic information, such as durations, intonations and so on, of the input speech were directly copied into the converted speech. The preference test results are shown in Table V. We see from the comparisons in similarity that the *GTDNN* and *GTDNN-mf* methods outperformed the *FTDNN*, and there is no clear preference between *GTDNN* and *GTDNN-mf*. In the naturalness test, the *GTDNN* method outperformed both the *FTDNN* and *GTDNN-mf* methods. This is quite interesting because the results of subjective tests are exactly the opposite of the results derived from the objective LSDs. These results indicate that the LSD measurement is inconsistently related to human auditory perception. On the other hand, the mean-field sampling in *GTDNN-mf* may result in an averaging effect in the generated spectral envelopes, which results in worse performance compared to *GTDNN*. Based on these results, the *GTDNN* system is used in the remaining experiments to compare with the other methods.

TABLE VI
PREFERENCE TEST RESULTS (%) OF SIMILARITY (SIM.) AND NATURALNESS (NAT.) BETWEEN THE *GMM-MCEP* SYSTEM THE *GMM-SPE* SYSTEM

	GMM-MCEP	*GMM-SPE*	N/P	p
Sim.	12.08	22.08	65.84	0.007
Nat.	12.08	**26.25**	61.67	0.000

E. Subjective Comparisons in Spectral Envelope Conversion

In this section, we describe several subjective evaluations to compare the proposed method with the conventional methods. The first subjective evaluation was conducted to compare the *GMM-MCEP* and *GMM-SPE* systems. Ten sentences were selected from the test set of each conversion pair to comprise a set of forty sentences for the subjective comparison. The preference listening test conditions were the same as those in section IV-D. Table VI shows the results. It can be seen from the table that the *GMM-SPE* system performs slightly better than the *GMM-MCEP* system. Although the p-values given by the t-tests show some significance, the listeners had no preference for most of the sentence pairs. The mean vectors of each Gaussian component in *GMM-MCEP* and *GMM-SPE* systems are physically very close to each other because the mel-cepstra can be considered to be linearly transformed from their corresponding logarithmic spectral envelopes. The spectral envelopes can represent the speech signal more accurately, which leads to a slight superiority in the *GMM-SPE* system.

The second subjective evaluation was conducted to compare performance among *GMM-SPE*, *MoRBM*, *MoGBAM* and *GTDNN* in terms of mean opinion score (MOS). In this evaluation, twenty sentences were converted by each system for all four conversion pairs. Eight Chinese-native listeners participated in this evaluation. The listeners were asked to give a 5-scale opinion score (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) on both similarity and naturalness for each sentence. Fig. 5(a) shows the similarity MOS of each system and each conversion pair. Fig. 5(b) shows the naturalness MOS results. It can be seen from the figures that the performance of each method was consistent among different conversion pairs. *GTDNN* shows the best performance among these systems, followed by *MoGBAM*. *MoRBM* system is better than *GMM-SPE* system, but the improvement is not as significant as for the *MoGBAM* and *GTDNN* systems. The results also indicate that

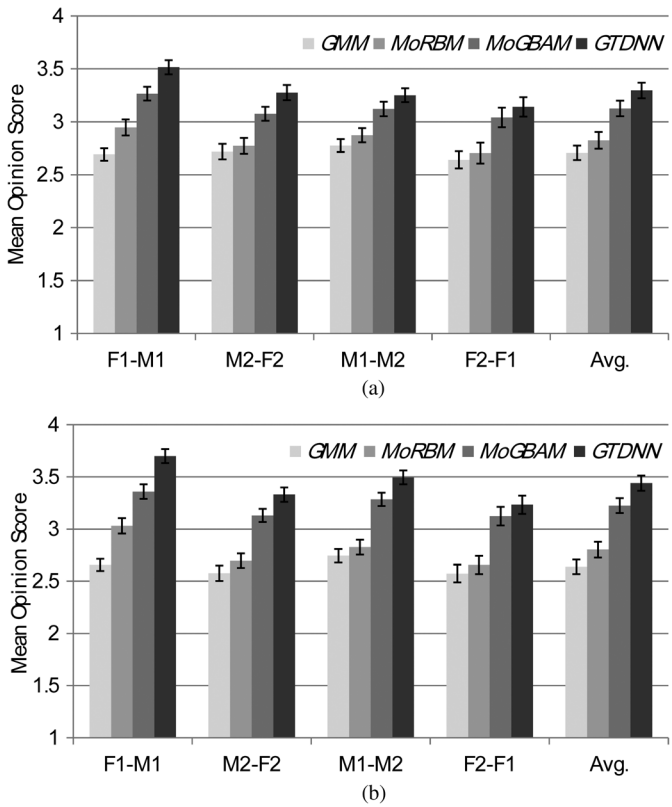

Fig. 5. MOS results for the *GMM-SPE*, *MoRBM*, *MoGBAM* and *GTDNN* methods on each conversion pair in similarity (a) and naturalness (b). The error bars show 95% confidence intervals.

the improvement of the proposed method in similarity is not as much as that for naturalness, but is still considerable. Note that the listeners were affected by the quality of the speech when giving the similarity scores because speech with higher quality tends to be considered as being more similar to natural speech.

Although *GTDNN* is the only system among the four that was not constructed using mixture models, its performance is the best. The strong ability of RBMs in modeling and generating is confirmed by the performance of *GTDNN* because all the converted spectral envelopes are generated from the RBM of the target speaker. In *MoRBM*, although the converted spectral envelopes are also generated from the RBMs, its performance is much worse than that of the *GTDNN*. This can be attributed to the difference between the utilization of RBMs in the two systems. In *GTDNN*, RBMs are only used to model the spectral envelope distributions of speaker dependent spaces. But in *MoRBM*, RBMs are adopted to model the joint distributions, which means that they are used not only for the distribution modeling but also for capturing the correlations between the spectral envelopes of the source and target speakers. At the conversion stage of the MoRBM-based method, the target spectral envelopes are missing, and the RBMs are employed to predicted the complete joint spectral envelopes from the incomplete input spectral envelopes. Our solution to this problem is to employ the gradient descent algorithm with an initialization of the complete joint spectral envelopes. The initialization is critical because RBMs are PoE models that describe very sharp distributions. An inappropriate initialization of the *MoRBM* system results in a poor utilization of the strong generating ability of the RBMs. Such problems does not exist in the *GTDNN* system because the intermediate BBAM can successfully predict the hidden variables for generating the converted spectral envelopes.

F. Comparing with GMM-GV

We also compared the performance of the systems using spectral envelopes with the *GMM-GV* system. The JDGMM based method, considering GV in parameter generation, is one of the best state of the art methods. The GV technique is well-known for its ability to alleviate the over-smoothing effect in GMM-based voice conversion [11] and HMM-based speech synthesis [42].

Fig. 6 plots the GV values of each order of the mel-cepstra converted by the systems involved in our experiments. The GV values were averaged over the twenty sentences in the test set of conversion pair F1-M1. For the systems directly constructed on spectral envelopes, the corresponding GV values were calculated using the mel-cepstra extracted from the converted spectral envelopes. It can be seen from the figure that the GV values of the *GMM-GV* system are almost the same as those of the target natural speech, especially in the higher orders, because the GV model is directly integrated into the criterion of generating the mel-cepstra. The GV values of the systems using spectral envelopes are clearly larger than those of the baseline system *GMM-MCEP*. The differences among the GV values of systems using spectral envelopes are consistent with those in the MOS evaluation results of Section IV-E. The GV values of the *GTDNN* system are close to those of natural speech, although they are not considered in any part of the training and conversion processes of the system. Therefore, it is not necessary to integrate GV into the parameter generation process of the *GTDNN* system.

The GV was also performed in JDGMM-based spectral envelope conversion (*GMM-SPE-GV*). Although *GMM-SPE-GV* can enhance the GV of each frequency bin of spectral envelopes, the improvement of GV in the corresponding mel-cepstra is very limited as shown in Fig. 6(a). A subjective preference listening test was conducted to see the effectiveness of *GMM-SPE-GV*. This test was conducted under the same condition as the previous preference tests. The results in Table VII show that although *GMM-SPE-GV* can improve the similarity and naturalness, the improvement is not as big as that in mel-cepstra. A reason could be that each frequency bin in the spectral envelope is generated independently in *GMM-SPE-GV* and the inter-frequency dependencies are ignored in the spectral generation. Although the mel-cepstra are also generated dimension-by-dimension, each dimension contains information from all frequency bins. Based on these results, the *GMM-GV* system, which was built on mel-cepstra, was used for further subjective comparison with the proposed method.

A subjective evaluation was also conducted to compare the performance of these systems with *GMM-GV*. The listening test conditions were the same as those in the preference tests in Section IV-D. Table VIII shows the results. The effectiveness of the GV technique is confirmed by the comparison between *GMM-GV* and *GMM-SPE* systems. *GTDNN* and *MoGBAM* significantly outperform the *GMM-GV* system, although *GMM-GV* generates the mel-cepstra with larger GV

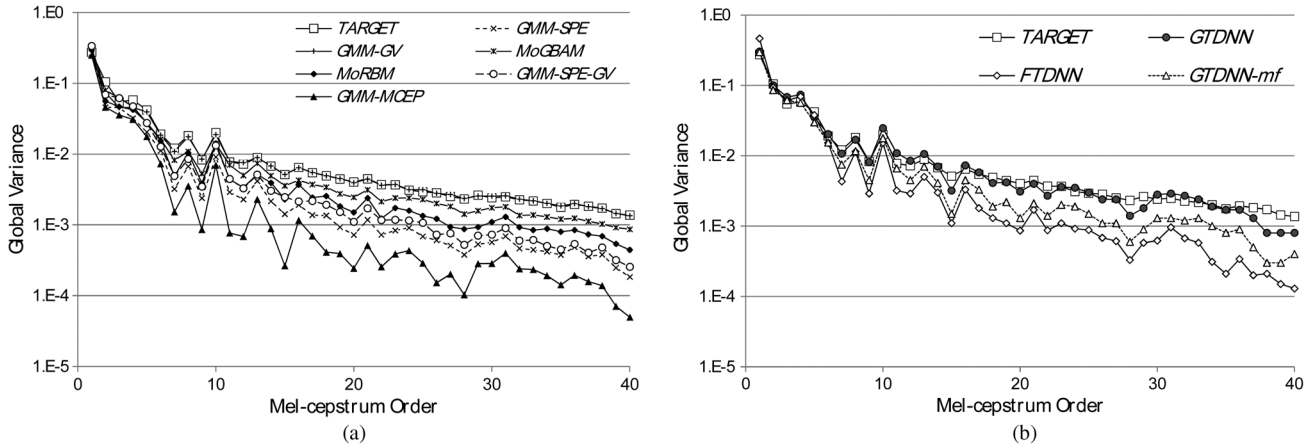

Fig. 6. The global variance (GV) values of each order of the mel-cepstra converted by (a) conventional methods and (b) DNNs along with those of the target natural speech.

TABLE VII
PREFERENCE SCORES OF COMPARISON AMONG *GMM-SPE*, *GMM-SPE-GV* AND *GMM-GV*

	GMM-SPE	*GMM-SPE-GV*	*GMM-GV*	N/P	*p*
Sim.	7.14	**48.57**	–	44.29	0.000
	–	10.71	**61.43**	27.86	0.000
Nat.	5.71	**61.43**	–	32.86	0.000
	–	9.29	**70.71**	20.00	0.000

TABLE VIII
THE RESULTS OF PREFERENCE TESTS (%) THAT COMPARE THE PERFORMANCE BETWEEN THE *GMM-GV* SYSTEM AND THE SYSTEMS USING SPECTRAL ENVELOPES

	GMM -GV	*GMM -SPE*	*MoRBM*	*MoGBAM*	*GTDNN*	N/P	*p*
Sim.	**44.17**	3.33	–	–	–	52.50	0.000
	50.83	–	7.08	–	–	42.08	0.000
	8.33	–	–	**32.92**	–	58.75	0.000
	13.75	–	–	-	**40.83**	45.42	0.000
Nat.	**77.92**	3.33	–	–	–	18.75	0.000
	63.33	–	8.75	–	–	27.92	0.000
	12.50	–	–	**45.42**	–	42.08	0.000
	12.08	–	–	–	**55.83**	32.09	0.000

values. However, GV is only a measurement in the mel-cepstral domain, the *MoGBAM* and *GTDNN* systems provide better conversion methods directly on the spectral envelopes, which therefore retain more detailed spectral characteristics so as to improve the speech quality. The performance of the *MoRBM* system is worse than the *GMM-GV* system due to the inappropriate approximations made at its conversion stage.

G. Objective Evaluations

Beside the subjective evaluations in the previous sections, we also calculated the LSDs between the spectral envelopes converted by the systems evaluated in our experiments and those extracted from the corresponding reference natural target recordings in the test set. For the mel-cepstra based systems such as *GMM-MCEP* and *GMM-GV*, the generated mel-cepstra were converted into spectral envelopes for calculating the LSDs. The same alignments as used in Section IV-D were applied prior to the LSD computation of all the systems. The average LSDs are listed in Table IX. It can be seen that there is no consistency between these objective results and the subjective evaluation results obtained in the previous experiments. The average LSD of the *GMM-MCEP* system is the lowest, but its performance is the worst in the subjective evaluations. The *GTDNN* system generates speech that performs the best in the subjective evaluations although the average LSD is only the second highest. These results are similar to those observed in Section IV-D. Again, the LSD measurement is inconsistent with human auditory perception. Similar results have also been observed in [11], [24].

TABLE IX
THE LOG-SPECTRAL DISTORTION BETWEEN THE SPECTRAL ENVELOPES OF NATURAL SPEECH AND THOSE CONVERTED BY EACH SYSTEM

	mel-cepstrum		spectral envelope			
	GMM -MCEP	*GMM -GV*	*GMM -SPE*	*MoRBM*	*MoGBAM*	*GTDNN*
F1-M1	4.57	5.16	4.61	5.25	5.05	5.38
M2-F2	4.33	4.86	4.35	5.23	4.67	5.13
M1-M2	4.00	4.21	4.00	4.73	4.24	4.53
F2-F1	4.34	4.64	4.34	5.19	4.57	4.74
Avg.	**4.31**	4.72	4.33	5.10	4.63	4.94

Fig. 7 shows the spectrograms of one sentence in the test set of the conversion pair F1-M1 converted by the *GMM-MCEP*, *GMM-GV* and *GTDNN* systems respectively. We can see that the formants of the spectra converted by the *GTDNN* system are clearly sharper than those converted by the *GMM-MCEP* and *GMM-GV* systems. The quality of speech converted by the *GMM-MCEP* system is degraded due to the fact that most of the spectral details are smoothed. The spectral details in the higher frequency area ($4 \sim 8$ kHz) are seriously smoothed by the conversion in the *GMM-GV* and *GMM-MCEP* systems. On the contrary, lots of detailed spectral characteristics are preserved by the conversion using *GTDNN*. This benefits from the effective modeling of RBMs on raw spectral envelopes directly. These observations from the spectrograms are consistent with the subjective evaluation results presented in previous sections.

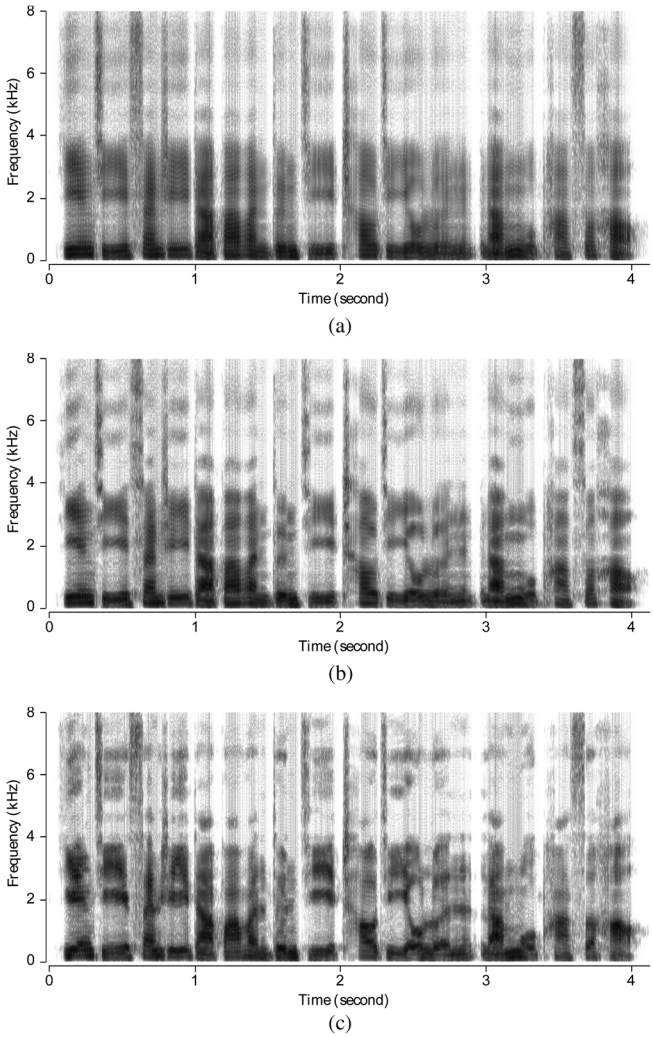

Fig. 7. Spectrograms drawn directly using the same segment of spectral envelopes of the target natural speech and those converted by each system (a) *GMM-MCEP* (b) *GMM-GV* (c) *GTDNN*.

V. Conclusion and Future Work

In this paper, we have proposed a new method for direct conversion of spectral envelopes extracted by the STRAIGHT vocoder for statistical voice conversion. A DNN was employed for the spectral envelope conversion. Unlike a conventional DNN, the proposed DNN was trained by layer-wise generative training. Two RBMs were trained to model the spectral envelope distributions of source and target speakers respectively. A BBAM was them employed to model the joint distribution of the hidden variables extracted from the two RBMs. The two RBMs were joined using a BBAM. The proposed GTDNN can benefit from the strong modeling and generating ability of RBMs and the superiority of the BBAM in deriving the conditional distribution. During conversion, the conditional distribution of the output spectral envelope, given an input spectral envelope, can be derived layer-by-layer. The derived conditional distribution was approximated by a single Gaussian distribution. Therefore, the conventional MOPPG algorithm in the JDGMM-based method was applied to generate the converted spectral envelopes in the proposed method. Our experimental results showed significant improvement achieved by the proposed method. In addition, the over-smoothing effect in the converted speech was effectively alleviated.

In this paper, a DNN with only two hidden layers was studied. Deeper neural networks are known to be more powerful in describing non-linear mapping relationships. Therefore, future work will include using deeper networks for spectral envelope conversion. Specifically, DBNs or DBMs, which are deeper generative models, can be employed to model the spectral envelope distributions instead of RBMs. Further more, a cascade of multiple BBAMs can be employed to model the mapping relationship between the hidden variables extracted from the RBMs, DBNs or DBMs. Finding a better criterion than MMSE for joint fine-tuning of the parameters in all layers is also necessary in the future. On the other hand, based on the experiences of DNN-based speech recognition [43], the DNN can be trained in a speaker-independent way using sufficient training data from multiple source speakers. In this way we intend to construct an any-to-one conversion model using the proposed framework in the future.

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1988, pp. 655–658.

[2] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 172–183, Jan. 2014.

[3] T. L. New, M. Dong, P. Chan, X. Wang, B. Ma, and H. Li, "Voice conversion: From spoken vowels to singing vowels," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2010, pp. 1421–1426.

[4] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 9, pp. 2505–2517, Nov.. 2012.

[5] Z. W. Shuang, R. Bakis, S. Shechtman, D. Chazan, and Y. Qin, "Frequency warping based on mapping formant parameters," in *Proc. Interspeech*, 2006, pp. 2290–2293.

[6] D. Erro and A. Moreno, "Weighted frequency warping for voice conversion," in *Proc. Interspeech*, 2007, pp. 1965–1968.

[7] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

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

[9] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1301–1312, Jul. 2006.

[10] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," *Eurospeech*, pp. 2413–2416, 2003.

[11] T. Toda, A. 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.

[12] 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.

[13] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *Proc. ICASSP*, 2007, vol. 4, pp. IV–513–IV–516.

[14] Z. Wu, T. Virtanen, T. Kinnunen, E. Chng, and H. Li, "Exemplar-based unit selection for voice conversion utilizing temporal information," in *Proc. Interspeech*, 2013.

[15] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 806–817, Mar. 2012.

[16] N. Pilkington, H. Zen, and M. J. F. Gales, "Gaussian process experts for voice conversion," in *Proc. Interspeech*, 2011, pp. 2761–2764.

[17] S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 954–964, Jul. 2010.
[18] T. Nakashika, T. Takashima, R. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *Proc. Interspeech*, 2013, pp. 369–372.
[19] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted Boltzmann machine for voice conversion," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, 2013, pp. 104–108.
[20] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion," in *Proc. Interspeech*, 2013, pp. 3052–3056.
[21] L.-J. Liu, L.-H. Chen, Z.-H. Ling, and L.-R. Dai, "Using bidirectional associative memories for joint spectral envelope modeling in voice conversion," in *Proc. ICASSP*, 2014, pp. 7884–7888.
[22] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis," in *Proc. ICASSP*, 2013, pp. 7825–7829.
[23] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 12, no. 14, pp. 1711–1800, 2002.
[24] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2129–2139, Oct. 2013.
[25] Y. Tang and R. Salakhutdinov, "Learning stochastic feedforward neural networks," in *Advances in Neural Information Processing Systems 26*. Cambridge, MA, USA: MIT Press, 2013, pp. 530–538.
[26] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *Proc. ICASSP*, 2013, pp. 8012–8016.
[27] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel distributed processing: explorations in the microstructure of cognition*, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA, USA: MIT Press, 1986, vol. 1, ch. 6, pp. 194–281.
[28] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[29] R. Salakhutdinov, "Learning deep generative models," Ph.D. dissertation, Univ. of Toronto, Toronto, ON, Canada, 2009.
[30] B. Kosko, "Bidirectional associative memories," *IEEE Trans. Systems, Man, Cybern.*, vol. 18, no. 1, pp. 49–60, Jan. 1988.
[31] H. Oh and S. C. Kothari, "A pseudo-relaxation learning algorithm for bidirectional associative memory," in *Proc. Int. Joint Conf. Neural Networks (IJCNN'92)*, 1992, vol. 2, pp. 208–213.
[32] M. Hattori, M. Hagiwara, and M. Nakagawa, "Quick learning for bidirectional associative memory," *IEICE Trans. Inf. Syst.*, vol. 77, no. 4, pp. 385–392, 1994.
[33] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Comput. Speech Lang.*, vol. 21, no. 1, pp. 153–173, 2007.
[34] T. Stafylakis, P. Kenny, M. Senoussaoui, and P. Dumouchel, "PLDA using Gaussian restricted Boltzmann machines with application to speaker verification," in *Proc. Interspeech*, 2012.
[35] Y. Bengio, "Learning deep architectures for AI," *Foundat. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009.
[36] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade*. New York, NY, USA: Springer, 2012, vol. 7700, pp. 599–619.
[37] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart, "Distributed representations," in *Parallel distributed processing: explorations in the microstructure of cognition*, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA, USA: MIT Press, 1986, vol. 1, ch. 6, pp. 77–109.
[38] R. Salakhutdinov and G. E. Hinton, "Deep Boltzmann machines," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2009, pp. 448–455.
[39] 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 Commun.*, vol. 27, no. 3, pp. 187–208, 1999.
[40] G. E. Hinton, "Products of experts," in *Proc. 9th Int. Conf. Artif. Neural Netw. (ICANN '99) (Conf. Publ. No. 470)*, 1999, vol. 1, pp. 1–6, IET.
[41] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580 2012.
[42] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. 90, no. 5, pp. 816–824, 2007.
[43] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

Ling-Hui Chen received the B.E. degree in electronic information engineering, and Ph.D. degree in signal and information processing from University of Science and Technology of China, Hefei, China, in 2008 and 2013, respectively. From April 2010 to October 2010, he was a visiting student at Nagoya Institute of Technology, Japan. He is currently a joint postdoctoral researcher at University of Science and Technology of China and iFLYTEK Co., Ltd., China. His research interests include voice conversion, speech synthesis and machine learning.

Zhen-Hua Ling (M'10) received the B.E. degree in electronic information engineering, M.S. and Ph.D. degree in signal and information processing from University of Science and Technology of China, Hefei, China, in 2002, 2005, and 2008, respectively.

From October 2007 to March 2008, he was a Marie Curie Fellow at the Centre for Speech Technology Research (CSTR), University of Edinburgh, U.K. From July 2008 to February 2011, he was a joint postdoctoral researcher at University of Science and Technology of China and iFLYTEK Co., Ltd., China. He is currently an associate professor at University of Science and Technology of China. He also worked at the University of Washington, USA, as a visiting scholar from August 2012 to August 2013. His research interests include speech processing, speech synthesis, voice conversion, speech analysis, and speech coding. He was awarded IEEE Signal Processing Society Young Author Best Paper Award in 2010.

Li-Juan Liu received the B.E. degree in communication engineering from HeFei University of Technology, China, in 2012. She is currently a graduate student in the Department of Electronic Engineering and Information Science at the University of Science and Technology of China. Her research interests include voice conversion, speech synthesis, and deep learning.

Li-Rong Dai was born in China in 1962. He received the B.S. degree in electrical engineering from Xidian University, Xian, China, in 1983 and the M.S. degree from Hefei University of Technology, Hefei, China, in 1986, and the Ph.D. degree in signal and information processing from the University of Science and Technology of China, Hefei, in 1997.

He joined University of Science and Technology of China in 1993. He is currently a Professor of the School of Information Science and Technology, University of Science and Technology of China. His current research interests include speech synthesis, speaker and language recognition, speech recognition, digital signal processing, voice search technology, machine learning, and pattern recognition. He has published more than 50 papers in these areas.