USING BIDIRECTIONAL ASSOCIATIVE MEMORIES FOR JOINT SPECTRAL ENVELOPE MODELING IN VOICE CONVERSION

Li-Juan Liu, Ling-Hui Chen, Zhen-Hua Ling, Li-Rong Dai

National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R. China
{ljliu037, chenlh}@mail.ustc.edu.cn, {zhling, lrdai}@ustc.edu.cn

ABSTRACT

The spectral envelope is the most natural representation of speech signal. But in voice conversion, it is difficult to directly model the raw spectral envelope space, which is high dimensional and strongly cross-dimensional correlated, with conventional Gaussian distributions. Bidirectional associative memory (BAM) is a two-layer feedback neural network that can better model the cross-dimensional correlations in high dimensional vectors. In this paper, we propose to reformulate BAMs as Gaussian distributions in order to model the spectral envelope space. The parameters of BAMs are estimated using the contrastive divergence algorithm. The evaluations on likelihood show that BAMs have better modeling ability than Gaussians with diagonal covariance. And the subjective tests on voice conversion indicate that the performance of the proposed method is significantly improved comparing with the conventional GMM based method.

Index Terms— Spectral envelope modeling, bidirectional associative memory, contrastive divergence, voice conversion

1. INTRODUCTION

Voice conversion (VC) is a technique that modifies speech of one speaker (source speaker) in order to make it sound like that of another certain speaker (target speaker), while keeping the linguistic information unchanged.

Many methods have been proposed to build the spectral mapping relationship between source and target speakers. Among these methods, joint density GMM (JDGMM) [1] is one of the mainstream statistical approaches for its stable performance. However, speech converted by JDGMM suffers from serious over-smoothing problem [2] [3]. This problem mainly comes from two factors. The first one is the use of high-level spectral features [4], such as line spectral pairs and mel-cepstra, which are extracted from the raw spectra of speech. Some detailed characteristics on the raw spectra are lost during the extraction process. The second one is the statistical averaging of the Gaussian distribution. In JDGMM, the mean vectors of target conditional distributions are very close to the mean vectors of the target marginal distributions [5], which are the weighted averaging of all training samples. The averaging operation removes most of the detailed characteristics in the spectra, and results in a muffled sound in the converted speech. Both factors attribute to the difficulty in estimating the covariance matrices of GMM. Full covariance matrices are difficult to learn, especially for high dimensional features. Therefore, diagonal covariance matrices are usually adopted to model feature space with weakly cross-dimensional correlations, e.g. mel-cepstrum.

There are some approaches attempting to solve this problem in JDGMM-based framework, such as considering global variance (GV) in maximum output probability parameter generation (MOPPG) [3], directly modeling the spectral feature trajectories [6]. Restricted Boltzmann machine (RBM) [7] has been adopted to replace Gaussian distribution in modeling the spectral envelope at each HMM state in the HMM-based statistical parametric speech synthesis [4]. Though this modeling method can improve the conversion performance when it is applied in VC [8], the performance is closely related to the estimation of the modes of conditional distributions.

In this paper, we propose a new method to model raw spectral envelopes. We reformulate the two-layer feedback neural network BAMs as Gaussian distributions. The cross-dimensional correlations in spectral envelopes can be well modeled by the weights of BAM that interact between the neurons from the two layers. The contrastive divergence (CD) algorithm with 1-step Gibbs sampling [9] is adopted to estimate the parameters of BAM. Experimental results show that BAMs outperform Gaussians in describing the distribution of raw spectral envelope space, and the conversion performance is significantly improved.

This paper is organized as follows: Section 2 gives a brief review on JDGMM-based VC. Our proposed method is described in section 3. In section 4, we will present the experimental results. The conclusion is given in section 5.

2. VOICE CONVERSION BASED ON JDGMM

Let $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top, \Delta^2\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top, \Delta^2\boldsymbol{y}_t^\top]^\top$ be the $3D$ dimensional source and target feature vectors at frame t, respectively. The operator $[\cdot]^\top$ denotes matrix transposition, $\boldsymbol{x}_t$, $\Delta\boldsymbol{x}_t$ and $\Delta^2\boldsymbol{x}_t$ represent the static, dynamic and acceleration feature components of source spectra, $\boldsymbol{y}_t$, $\Delta\boldsymbol{y}_t$ and $\Delta^2\boldsymbol{y}_t$ represent those of target spectra, D is the dimension of static feature vector.

In JDGMM based method, the probability density function of the joint feature space $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ is modeled by a GMM

$$P(\boldsymbol{Z}_t; \boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} \beta_m N(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \ \sum_{m=1}^{M} \beta_m = 1, \quad (1)$$

where M denotes the number of mixture components, β_m,$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}$, $\boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}$ are the weight, mean

This work was partially funded by the National Nature Science Foundation of China (Grant No. 61273032, Grant No. 61273264) and the National 973 program of China (Grant No. 2012CB326405).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

vector and covariance matrix of the m-th mixture component, and $\boldsymbol{\Sigma}_m^{(xx)}$, $\boldsymbol{\Sigma}_m^{(xy)}$, $\boldsymbol{\Sigma}_m^{(yx)}$, $\boldsymbol{\Sigma}_m^{(yy)}$ are usually set as diagonal matrices.

At conversion stage, for an input sentence, the feature sequence is $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, ..., \boldsymbol{X}_T^\top]^\top$, T is the total number of frames, the converted static feature sequence $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, ..., \boldsymbol{y}_T^\top]^\top$ can be generated by the maximum output probability criterion as

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(z)}), \tag{2}$$

$$s.t. \qquad \boldsymbol{Y} = \boldsymbol{M}\boldsymbol{y}, \tag{3}$$

where $\boldsymbol{M}$ is a $3DT \times DT$ dimensional matrix that is used to generate the feature sequence consisting of static, dynamic and acceleration components from static feature sequence.

As the conditional probability distribution of target feature is a GMM, the sub-optimal mixture sequence $\hat{\boldsymbol{m}} = \{\hat{m}_1, \cdots \hat{m}_T\}$ is usually adopted to reduce the computational complexity [3]. Then the conditional distribution can be approximated by a single Gaussian distribution whose mean vector and covariance matrix are

$$\boldsymbol{\mu}_{\hat{m}_t}^{(y|x)} = \boldsymbol{\mu}_{\hat{m}_t}^{(y)} + \boldsymbol{\Sigma}_{\hat{m}_t}^{(yx)} {\boldsymbol{\Sigma}_{\hat{m}_t}^{(xx)}}^{-1} (\boldsymbol{X}_t - \boldsymbol{\mu}_{\hat{m}_t}^{(x)}), \tag{4}$$

$$\boldsymbol{\Sigma}_{\hat{m}_t}^{(y|x)} = \boldsymbol{\Sigma}_{\hat{m}_t}^{(yy)} - \boldsymbol{\Sigma}_{\hat{m}_t}^{(yx)} {\boldsymbol{\Sigma}_{\hat{m}_t}^{(xx)}}^{-1} \boldsymbol{\Sigma}_{\hat{m}_t}^{(xy)}, \tag{5}$$

where $\hat{m}_t = \arg\max_m \ P(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(z)})$. Thus, the converted feature sequence can be generated with a closed-form solution as

$$\boldsymbol{y}^* = (\boldsymbol{M}^\top \boldsymbol{U}_{\hat{\boldsymbol{m}}}^{-1} \boldsymbol{M})^{-1} \boldsymbol{M}^\top \boldsymbol{U}_{\hat{\boldsymbol{m}}}^{-1} \boldsymbol{E}_{\hat{\boldsymbol{m}}}, \tag{6}$$

where $\boldsymbol{E}_{\hat{\boldsymbol{m}}}$ and $\boldsymbol{U}_{\hat{\boldsymbol{m}}}$ are the concatenated conditional mean vector and covariance matrices in the sub-optimal sequence $\hat{\boldsymbol{m}}$.

3. SPECTRAL ENVELOPE MODELING WITH BIDIRECTIONAL ASSOCIATIVE MEMORIES

3.1. Bidirectional Associative Memory

A BAM [10] is a special feedback neural network that can store the patterns of binary vector pairs. Once one pattern is given, the network can recall the other one. It has been applied in the area of pattern recognition, signal processing, etc.

For this energy-based model, the patterns of input vectors are stored in the interaction matrix $\boldsymbol{W}$ at a local minimum of the system energy. As demonstrated in Fig.1, the BAM is made up of two symmetrically connected layers that has no interconnections among neurons in the same layer, thus the energy function of this network is defined as

$$\boldsymbol{E}(\boldsymbol{a}, \boldsymbol{b}) = -\boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{b}, \tag{7}$$

where $\boldsymbol{a} = [a_1, a_2, ..., a_N]^\top$ and $\boldsymbol{b} = [b_1, b_2, ..., b_P]^\top$ are binary stochastic variables corresponding to the neurons in the two layers respectively. $\boldsymbol{W}$ is learned under Hebbian law.

3.2. Reformulating BAM as a Gaussian Probability Model

In this paper, instead of using BAM by the conventional way, we develop a probability model based on it and extend it from the binary form to the Gaussian form in order to model the real-valued data.

When the neurons in the BAM correspond to Gaussian stochastic variables with zero mean, the energy function of this BAM is written as

$$\boldsymbol{E}(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{N} \frac{a_i^2}{2\sigma_{a,i}^2} + \sum_{j=1}^{P} \frac{b_j^2}{2\sigma_{b,j}^2} - \sum_{i=1}^{N} \sum_{j=1}^{P} w_{i,j} \frac{a_i}{\sigma_{a,i}} \frac{b_j}{\sigma_{b,j}}, \tag{8}$$

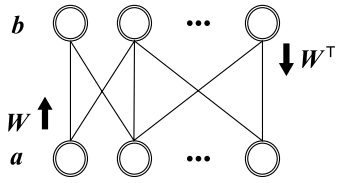

Fig. 1. The structure of bidirectional associative memory. When it is applied in voice conversion, the numbers of neurons in the down-layer and top-layer are equal to the dimension of feature vectors of each speaker.

where $\sigma_{a,i}$ and $\sigma_{b,j}$ are the parameters of the model, in this paper, we fix them to the standard deviation of corresponding random variables. We name this model as Gaussian BAM (GBAM).

The joint distribution over the two layer neurons is defined as

$$P(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{\mathcal{Z}} \exp\{-\boldsymbol{E}(\boldsymbol{a}, \boldsymbol{b})\}, \tag{9}$$

and $\mathcal{Z} = \int\int \exp\{-E(\boldsymbol{a}, \boldsymbol{b})\} d\boldsymbol{a} d\boldsymbol{b}$ is the partition function. For the joint variables vector $\boldsymbol{v} = [\boldsymbol{a}^\top, \boldsymbol{b}^\top]^\top$, this p.d.f can be written in a Gaussian-form [11] with zero mean and precision matrix

$$\boldsymbol{\Lambda} = \boldsymbol{\Gamma}^{-1} \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{W} \\ -\boldsymbol{W}^\top & \boldsymbol{I} \end{bmatrix} \boldsymbol{\Gamma}^{-1}, \tag{10}$$

where $\boldsymbol{\Gamma} = diag\{\boldsymbol{\Gamma}^{(a)}, \boldsymbol{\Gamma}^{(b)}\}$ and $\boldsymbol{\Gamma}^{(a)} = diag\{\sigma_{a,1}, ..., \sigma_{a,N}\}$, $\boldsymbol{\Gamma}^{(b)} = diag\{\sigma_{b,1}, ..., \sigma_{b,P}\}$. Therefore, as a probability density model, if the precision matrix $\boldsymbol{\Lambda}$ is positive definite, the GBAM is equivalent to a Gaussian distribution.

The Gaussian reformulated by GBAM can model the correlations between variables from different layers by $\boldsymbol{W}$. Although the variables in the same layer are independent of each other once the other layer is given, the correlations in the same layer can be captured by $\boldsymbol{W}$ during the information flowing between the two layers. Therefore, GBAM is capable of modeling the probability density of spectral feature vectors that contain strongly correlated coefficients.

3.3. Parameter Estimation

Since we evaluate the GBAM as a probability model, different from the standard learning algorithm for BAM, we estimate the parameters under the maximum likelihood (ML) criterion. Because of the structure relation between $\boldsymbol{W}$ and $\boldsymbol{\Lambda}$ in (10), the derivative over $\boldsymbol{W}$ is calculated with the likelihood in (9) instead of the standard form of Gaussian. So the derivative over $\boldsymbol{W}$ is written as

$$\partial \boldsymbol{W} = {\boldsymbol{\Gamma}^{(a)}}^{-1} (E_d[\boldsymbol{a}\boldsymbol{b}^\top] - E_m[\boldsymbol{a}\boldsymbol{b}^\top]) {\boldsymbol{\Gamma}^{(b)}}^{-1}, \tag{11}$$

where $E_d[\cdot]$ denotes the expectation with respect to the data distribution and $E_m[\cdot]$ denotes the expectation with respect to the distribution defined by the model. CD algorithm [9] is adopted to approximate the expectation over model distribution. Since the variables in one layer are independent of each other if the variables in the other layer are given, a Gibbs chain is run to iteratively sample data from $P(\boldsymbol{a}^k|\boldsymbol{b}^{k-1})$ and $P(\boldsymbol{b}^k|\boldsymbol{a}^{k-1})$, which are both Gaussian distributions, at the k-th step. The Gibbs chain starts from the initial states $\boldsymbol{a}^0$ and $\boldsymbol{b}^0$ given by the training samples.

Parameter matrix $\boldsymbol{W}$ is estimated by the gradient descent algorithm as

$$\boldsymbol{W}^{(k)} = \boldsymbol{W}^{(k-1)} + \Delta\boldsymbol{W}^{(k)}, \tag{12}$$

where the update at the k-th step $\Delta\boldsymbol{W}^{(k)}$ consists of the derivative over $\boldsymbol{W}^{(k)}$ as well as a momentum term and a regularization term. The momentum term is added in order to accelerate the learning speed [12], the regularization term is adopted to guarantee the positive definite of $\boldsymbol{\Lambda}$. Therefore, $\Delta\boldsymbol{W}^{(k)}$ is written as

$$\Delta\boldsymbol{W}^{(k)} = \alpha\Delta\boldsymbol{W}^{(k-1)} + \rho(\partial\boldsymbol{W}^{(k)} - \epsilon\boldsymbol{W}^{(k-1)}), \tag{13}$$

where α, ρ, ϵ are the coefficients of the momentum item, learning rate and the regularization item separately, $\partial\boldsymbol{W}^{(k)}$ is the derivative calculated at the k-th step.

3.4. System Construction

We propose to directly model the distribution of joint spectral envelope space by using GBAMs instead of Gaussians. The variables in the down-layer and up-layer are corresponding to the source part and target part of the joint space. The training of GBAM is very time consuming, so it is infeasible to directly train a mixture of GBAMs with the expectation maximization (EM) algorithm. Therefore, the joint space is divided into several sub-spaces, and each sub-space is modeled by a GBAM. In this paper, we use a GMM to divide the joint space. Since it is difficult to train a GMM directly on the spectral envelope space, a JDGMM is estimated on the joint high-level spectral feature space, and the joint spectral envelope is assigned to the sub-space that has the maximum posterior probability of generating the corresponding high-level spectral feature. The training data in each sub-space are normalized to zero mean before they are used to estimate the parameters of GBAM. So the parameter set for the m-th sub-space is given by

$$\theta_m = \{\boldsymbol{\eta}_m, \boldsymbol{\Gamma}_m, \boldsymbol{W}_m\},\ m = 1, \cdots M, \tag{14}$$

where $\boldsymbol{\eta}_m$ and $\boldsymbol{\Gamma}_m^2$ are the mean vector and diagonal covariance matrix of the training data in the m-th sub-space, and $\boldsymbol{W}_m$ is the weight matrix of the GBAM in the same space.

At conversion stage, the GBAM is treated exactly as a Gaussian distribution. For the m-th sub-space, given a source spectral envelope feature vector $\boldsymbol{X}_t$, the conditional distribution of target spectral envelope is $N(\boldsymbol{Y}_t; \boldsymbol{\eta}_m^{(y|x)}, {\boldsymbol{\Gamma}_m^{(y)}}^2)$, where

$$\boldsymbol{\eta}_m^{(y|x)} = \boldsymbol{\eta}_m^{(y)} + \boldsymbol{\Gamma}_m^{(y)}\boldsymbol{W}_m^{\top}{\boldsymbol{\Gamma}_m^{(x)}}^{-1}(\boldsymbol{X}_t - \boldsymbol{\eta}_m^{(x)}). \tag{15}$$

Then the static spectral envelope feature sequence is generated in a way similar to the procedure in the GMM-based method.

Though RBMs can also model the joint spectral envelope for VC [8], the relationship between the source and the target spectral envelope feature vectors are captured through the hidden variables. As the states of hidden variables are unknown at the conversion stage, the conversion performance closely relies on the initialization of target nodes. But in the proposed method, there is no hidden layer and the converted spectral vectors can be generated directly .

4. EXPERIMENTS

4.1. Experimental Conditions

A Chinese speech corpus with a female and a male speakers was adopted to build a female-to-male conversion. Waveforms are recorded in 16kHz/16bit format. 100 parallel utterances were used in our experiments. 80 utterances were randomly chosen as the training set, and the remaining 20 utterances were used for testing.

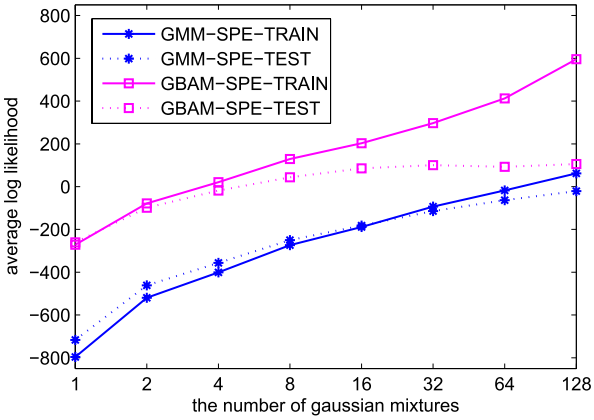

Fig. 2. Average log likelihoods on the training and test sets when modeling spectral envelope using GMM-based model and GBAM-based model.

Spectral envelope features were extracted by STRAIGHT [13] analysis. The number of FFT points was 1024, which lead to a 513-dimensional spectral envelope. 40-order mel-cepstra (excluding the 0-th coefficient) were extracted from spectral envelopes as spectral features for GMM modeling. The dynamic time align (DTW) algorithm was used to align the mel-cepstrum sequences of source and target speakers. Spectral envelopes were aligned using the aligning information of the corresponding mel-cepstra.

In the training of GBAMs, one-step Gibbs sampling was applied in the CD algorithm. The batch size was set to 10. The initial value of $\boldsymbol{W}$ was zero. The learning rate ρ and regularization coefficients ϵ were set to 0.00005 and 0.03 respectively. The learning procedure started with the momentum of 0.5 and changed to 0.9 after 5 epoches.

4.2. Results and Analysis

We built the following systems to evaluate the performance of the proposed method[1]:

a) *GMM-MCEP*: The conventional JDGMM-based method. Mel-cepstra were used as the spectral features;

b) *GMM-MCEP-GV*: Considering GV at parameter generation in GMM-MCEP system. GV is the well-known method to address the over-smoothing problem in VC;

c) *GBAM-SPE*: The proposed method in this paper;

d) *GMM-SPE*: A single Gaussian distribution with diagonal covariance matrix was adopted to model the spectral envelope distribution in each sub-space.

First, the average log likelihoods with different number of sub-spaces were evaluated on both the training and test sets to compare the modeling ability between GBAMs and Gaussians. From Fig.2, we can see that the likelihoods of GBAMs are significantly larger than those of Gaussians with diagonal covariance matrices. This verifies that GBAM does better in modeling spectral envelope.

[1]Some speech examples converted by these systems can be found at `http://home.ustc.edu.cn/~ljliu037/ICASSP2014_GBAMVC.html`.

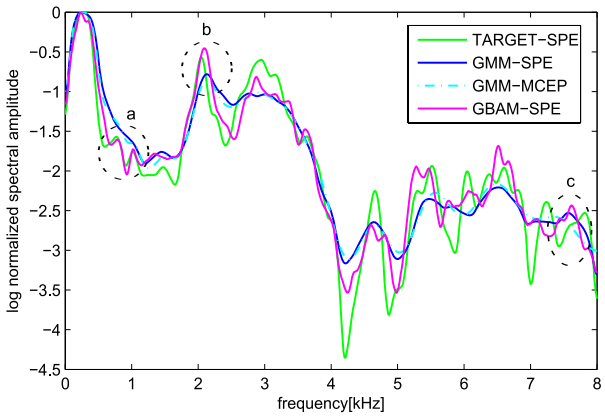

Fig. 3. An example of converted spectral envelopes. The spectral envelope of GMM-MCEP system was obtained by recovering from the converted mel-cepstrum.

Fig.3 shows an example of target spectral envelope as well as the corresponding ones converted by GMM-MCEP, GMM-SPE, and GBAM-SPE systems. The maximal values of those spectral envelopes are normalized to 1, and the envelopes are presented in log-scale. We can see that the envelopes converted by GMM-SPE and GMM-MCEP are almost the same and both are over smoothed, while GBAM-SPE generates the local spectral characteristics of the target, e.g.circle (a), (b), (c), benefiting from the ability of modeling the full dimensional relationship between source and target spectral envelope feature vectors. However, this leads to a larger log spectral distortion(LSD) compared with GMM-based systems as shown in Table1. This is reasonable because the local characteristics of the spectra reconstructed by GBAM-SPE may not always be close to those of the nature ones while the spectra converted using GMM-based approaches stay around the mean vectors of the Gaussians representing marginal probability density functions of the target speaker.

When the mixture number got larger than 128, the training data in some sub-spaces became sparse that GBAMs couldn't be trained. But the informal listening test showed that speech converted by GMM-based systems got no significantly improvement, we fix the mixture number to 128 for all the systems in the subjective tests.

We conducted mean opinion score (MOS) tests to evaluate speech naturalness and similarity among GBAM-SPE system, GMM-MCEP system and GMM-SPE system. All the twenty sentences in the test set were used as the evaluation set. The number of listeners was seven. The evaluation results presented in Fig.4 demonstrate that the conversion performance of GBAM-SPE is better than that of GMM-MCEP and GMM-SPE.

Table 1. Average log spectral distortions(dB) for different systems. The spectral envelopes of GMM-MCEP, GMM-MCEP-GV systems were recovered from the converted mel-cepstra before LSDs were calculated.

	GMM-MCEP	GMM-SPE	GMM-MCEP-GV	GBAM-SPE
LSD(dB)	4.57	4.61	5.16	5.05

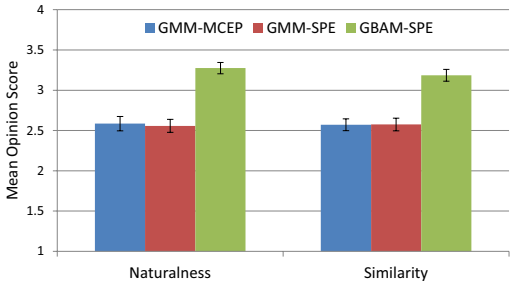

Fig. 4. Mean opinion scores of speech quality and similarity. Error bars show 95% confidence interval.

In order to investigate the performance on alleviating the over-smoothing problem by the proposed method, a preference test was taken to compare GBAM-SPE with GMM-MCEP-GV. GV was not considered in the GBAM-SPE system because we found it had no promoting effect on spectral envelope modeling. Results in Fig.5 indicate that listeners tend to prefer the voice converted by GBAM-SPE system. The p-values show an statistically significant difference between the systems on both speech naturalness and similarity and confirm the superiority of GBAM-SPE in voice conversion. Through Table.1 we can see that GBAM-SPE outperforms GMM-MCEP-GV consistently with 0.11dB lower spectral distortion though they both cause a larger distortion than GMM-SPE.

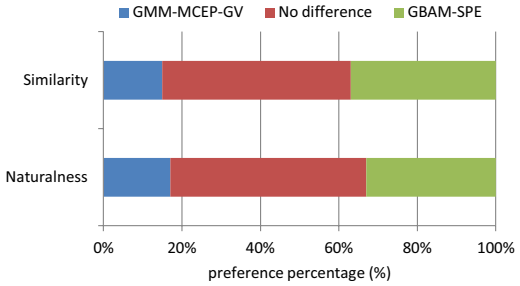

Fig. 5. Preference scores of GMM-MCEP-GV and GBAM-SPE on speech naturalness and similarity. The p-value of the t-tests are 8.3×10^{-7} and 3.2×10^{-7} on naturalness and similarity respectively.

5. CONCLUSION

In this paper, we proposed a new approach to model the spectral envelope. Two-layer feedback neural networks, BAMs, were reformulated to represent the Gaussian distributions in GMM, in order to capture the strongly cross-dimensional correlations in the high order spectral envelopes. CD algorithm with 1-step Gibbs sampling was used to estimate the parameters of BAMs. The likelihood comparision between GBAMs and Gaussians with diagonal covariances indicated that GBAMs outperformed in describing the probability distribution of the joint spectral envelope space. Subjective evaluations showed that the speech naturalness and similarity could be significantly improved. As only a female-to-male conversion was conducted in this paper, we will examine the performance of GBAMs on other speaker pairs in the future.

6. REFERENCES

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

[2] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Acoustics, Speech, and Signal Processing*, 2001, vol. 2, pp. 841–844.

[3] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, and Lang. Process*, vol. 15, no. 8, pp. 2222–2235, nov. 2007.

[4] Z.H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2129–2139, 2013.

[5] Yining Chen, Min Chu, Eric Chang, Jia Liu, and Runsheng Liu, "Voice conversion with smoothed GMM and MAP adaptation.," in *Eurospeech*, 2003, pp. 2413–2416.

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

[7] R. Salakhutdinov, *Learning deep generative models*, Ph.D. thesis, University of Toronto, 2009.

[8] 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.

[9] G.E Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 12, no. 14, pp. 1711–1800, 2002.

[10] B. Kosko, "Bidirectional associative memories," *IEEE Trans. Systems, Man and Cybernetics*, vol. 18, no. 1, pp. 49–60, 1988.

[11] T. Stafylakis, P. Kenny, M. Senoussaoui, and et al, "PLDA using Gaussian restricted Boltzmann machines with application to speaker verification," in *INTERSPEECH*, 2012.

[12] G.E Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, 2010.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–208, 1999.