Joint Spectral Distribution Modeling Using Restricted Boltzmann Machines for Voice Conversion

Ling-Hui Chen, Zhen-Hua Ling, Yan Song, Li-Rong Dai

National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R. China

chenlh@mail.ustc.edu.cn, {zhling, songy, lrdai}@ustc.edu.cn

Abstract

This paper presents a new spectral modeling and conversion method for voice conversion. In contrast to the conventional Gaussian mixture model (GMM) based methods, we use restricted Boltzmann machines (RBMs) as probability density models to model the joint distributions of source and target spectral features. The Gaussian distribution in each mixture of GMM is replaced by an RBM, which can better capture the inter-dimensional and inter-speaker correlations within the joint spectral features. Spectral conversion is performed by the maximum conditional output probability criterion. Our experimental results show that the similarity and naturalness of the proposed method are significantly improved comparing with the conventional GMM based method.

Index Terms: Voice Conversion, Restricted Boltzmann Machine, Spectral envelope

1. Introduction

Voice conversion technique aims to modify the input speech of one speaker (source) to make it sounds like uttering by another certain speaker (target). Many methods have been developed for voice conversion during the past years. Among these methods, GMM [1] and joint density GMM (JDGMM) [2] based methods have become a mainstream method because of its statistical approach. In JDGMM based method, joint spectral feature space of source speaker and target speaker is modeled by a GMM, and spectral conversion is performed by minimizing mean square error (MMSE) criterion. This method is further improved by involving dynamic features and using maximizing output probability parameter generation (MOPPG) algorithm [3].

Although the performance of JDGMM based voice conversion is quite stable, there still exist some problems. The most serious one is that the quality of the generated speech is severely degraded. This is mainly due to the over-smoothing effect in the converted speech features caused by the inadequate modeling of Gaussian based framework. In this approach, the converted spectral features are mainly determined by the target speaker part of the mean vectors of the JDGMM [4]. The model is trained by maximizing likelihood (ML) criterion, which means the mean of each mixture is a weighted average on all training samples from the training set. This averaging process removes most of the detailed characteristics in the training samples and leads to a muffled voice quality in converted speech.

Many approaches have been proposed to cope with the this problem, eg. integrating global variance (GV) in conversion with MOPPG algorithm [3], using trajectory model [5] to compensate the inadequate modeling in GMM with a whole sentence trajectory modeling, or frequency warping method to directly modify the spectral envelopes to retain their detailed characteristics [6], etc.

In this paper, we propose a new modeling method to improve the joint spectral feature modeling. We use the restricted Boltzmann machines (RBMs) to model the feature space instead of Gaussians. In this paper, the feature can be not only high-level spectral features (eg. mel-cepstra or line spectral pairs), but also the raw spectral envelopes extracted by STRAIGHT analysis [7]. An RBM is a graphic model, it has been successfully applied to hidden Markov model (HMM) based statistical parametric speech synthesis [8] as a density model to describe the distribution of spectral features at each HMM state. RBMs have better ability to capture the correlations between two speakers and can also describe the distribution of high order spectral envelopes.

The remaining of this paper is organized as follows. In section 2, we will describe the details of our proposed method. Section 3 shows our experimental results. Conclusions will be given in section 4.

2. Method

In the conventional joint space based voice conversion methods which incorporate dynamic features [3], the joint feature vector is $\boldsymbol{v}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$, where $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$, $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ denote the static features of the source and target speakers, $\Delta\boldsymbol{x}_t$ and $\Delta\boldsymbol{y}_t$ denote the corresponding dynamic features. A statistical model $\lambda^{(v)}$ is trained to model the joint feature space. When converting an input feature sequence $\tilde{\boldsymbol{X}} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T]$, where T is the length of the sequence, two steps are taken. First, a sequence of conditional distributions $P(\boldsymbol{Y}_t|\boldsymbol{X}_t, \lambda^{(v)})$ is determined frame-by-frame. Then, the trajectory of the whole sequence is generated by MOPPG algorithm from the conditional distributions:

$$\tilde{\boldsymbol{y}}^* = \arg\max_{\tilde{\boldsymbol{y}}} P(\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{X}}, \lambda^{(v)}) = \arg\max_{\tilde{\boldsymbol{y}}} \prod_{t=1}^{T} P(\boldsymbol{Y}_t|\boldsymbol{X}_t, \lambda^{(v)}) \tag{1}$$

where the sequence of converted features (including dynamic features) $\tilde{\boldsymbol{Y}}$ is a linear transformation of the static feature sequence $\tilde{\boldsymbol{y}}$.

2.1. Conventional GMM based method

In conventional JDGMM based methods [3], the density model is a GMM $\lambda^{(v)} = \{\alpha_m, \boldsymbol{\mu}_m^{(v)}, \boldsymbol{\Sigma}_m^{(v)}; m = 1, \cdots, M\}$, in which

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

M is the total number of mixtures, α_m, $\boldsymbol{\mu}_m^{(v)}$ and $\boldsymbol{\Sigma}_m^{(v)}$ are the weight, mean vector and covariance matrix of the m-th mixture, and

$$\boldsymbol{\mu}_m^{(v)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(v)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}. \quad (2)$$

In order to reduce the number of model parameters and computational cost, diagonal cross covariance matrices are often used in conversion phase, which means diagonal matrices are adopted for $\boldsymbol{\Sigma}_m^{(xx)}$, $\boldsymbol{\Sigma}_m^{(yy)}$, $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$.

In conversion phase, the output distribution given an input $\boldsymbol{X}_t$ is still an M mixture GMM $\lambda_t^{(y|x)}$, whose parameters of the m-th mixture are

$$\boldsymbol{\mu}_m^{y|x} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)-1}(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(x)}), \quad (3)$$

$$\boldsymbol{\Sigma}_m^{y|x} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)-1}\boldsymbol{\Sigma}_m^{(xy)}, \quad (4)$$

$$\alpha_m^{y|x} = P(m|\boldsymbol{X}_t, \lambda^{(v)}). \quad (5)$$

For simplicity, only a single Gaussian with the maximum posterior probability is often adopted for each frame. Then, a close-form solution to (1) can be achieved. The detailed formula can be found in [3].

However, the voice quality of converted speech is degraded due to an over-smoothing effect. In JDGMM the inter-speaker correlations are modeled by $\boldsymbol{\Sigma}_m^{(xy)}$, but in the diagonal case, only the correlations between two speakers in the same dimension are considered. We observed in our experiments that the same dimension in two speaker's features is often not the most correlated. As a result, the linear transformation term $\boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)-1}$ in (3) often tend to be very small. Therefore the main component in converted feature is the target means $\boldsymbol{\mu}_m^{(y)}$ of the Gaussian distributions. The estimated means of GMM are weighted averages of all training samples. This is a smoothing process which wipes most of the detailed characteristics in original spectral features and leads to the over-smoothing effect in converted spectral features.

2.2. Proposed method

2.2.1. RBM as a density model

An RBM [1] is an undirected graphical model [9]. This model has a two-layer architecture: a visible layer and a hidden layer. The visible stochastic units $\boldsymbol{v} = [v_1, \cdots, v_V]^\top$ corresponding to the acoustic features are connected to the hidden stochastic units $\boldsymbol{h} = [h_1, \cdots h_H]^\top$, where V and H denote the numbers of the visible and hidden units.

In spectral feature modeling, features are real-valued. In this case, Gaussian-Bernoulli RBMs (GBRBMs) are adopted. The GBRBM is composed by Gaussian stochastic visible units and binary stochastic hidden units, and its energy function is defined as:

$$E(\boldsymbol{v}, \boldsymbol{h}) = \sum_{i=1}^{V} \frac{(a_i - v_i)^2}{2\sigma_i^2} - \sum_{j=1}^{H} b_j h_j - \sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij} h_j \frac{v_i}{\sigma_i}, \quad (6)$$

where $\boldsymbol{a} = [a_i, \cdots, a_V]^\top$, $\boldsymbol{b} = [b_1, \ldots, b_H]^\top$ and $\boldsymbol{W} = \{w_{ij}\}_{V\times H}$ are the parameters of this model, the parameter $\{\sigma_i\}$ is commonly fixed to the standard deviation of the training data, and is usually omitted for notational simplicity.

[1] In the rest of this paper, GBRBM is written as RBM for short.

An RBM can also be treated as a probability density model, whose probability of generating a sample $\boldsymbol{v}$ is

$$P(\boldsymbol{v}) = \frac{1}{\mathcal{Z}} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h})), \quad (7)$$

where $\mathcal{Z} = \int \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}))d\boldsymbol{v}$ is the partition function. Since the calculation of the partition function is intractable, the learning of model parameters $\{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$ is normally carried out by using the contrastive divergence (CD) algorithm [10] under ML criterion and the partition function can be approximated by using the annealed importance sampling method [11].

2.2.2. System construction

In this paper, we divide the acoustic space into several sub-spaces, and model each of them with an RBM. Therefore, our RBM-based voice conversion system is built based on the conventional GMM-based system. A GMM is trained on acoustic feature space (eg. mel-cepstra) in advance to divide the acoustic space. Training samples are assigned to each of sub-spaces according to the maximum posterior probabilities of generating the samples from corresponding Gaussian mixtures. Then, the RBM for each sub-space $\lambda_m^{(v)}$ is trained.

In this way, the proposed model can be viewed as a mixture of RBMs. In case of spectral conversion, the visible layer is composed by two parts, each part corresponding to one speaker. As a result, the model parameters for m-th sub-space can be written as $\boldsymbol{a}_m = [\boldsymbol{a}_m^{(x)^\top}, \boldsymbol{a}_m^{(y)^\top}]^\top$ and $\boldsymbol{W}_m = [\boldsymbol{W}_m^{(x)^\top}, \boldsymbol{W}_m^{(y)^\top}]^\top$, where $(\cdot)^{(x)}$ and $(\cdot)^{(y)}$ denote to the model parameters for the source part and target part.

In our early experiments on static feature conversion, the quality of converted speech is degraded due to a discontinuous problem in the features generated directly by RBMs. In order to address this problem, dynamic features are used to smooth the converted feature sequence. At conversion time, the conditional distribution $P(\boldsymbol{Y}_t|\boldsymbol{X}_t, \lambda^{(v)})$ is approximated with a single Gaussians, whose mean vectors are generated from the RBMs by

$$\boldsymbol{Y}_t^* = \arg\max_{\boldsymbol{Y}_t} \sum_{m=1}^{M} P(m|\boldsymbol{X}_t, \lambda^{(v)}) P(\boldsymbol{Y}_t|\boldsymbol{X}_t, \lambda_m^{(v)}), \quad (8)$$

$$\doteq \arg\max_{\boldsymbol{Y}_t} P(\boldsymbol{Y}_t|\boldsymbol{X}_t, \lambda_{m^*}^{(v)}), \quad (9)$$

where m^* is the mixture with maximum posterior probability determined by the JDGMM. and the covariance matrix is approximated with the variances of all training samples in the corresponding sub-spaces. Similarly, only the mixture with maximum posterior probability is used in conversion. Once the output distribution sequence is determined, the converted static spectral feature sequence is calculated in the same manner with GMM-based voice conversion by MOPPG algorithm (1).

2.2.3. Spectral modeling using RBMs

As mentioned in section 2.1, diagonal cross covariance matrix is often adopted for each mixture of JDGMM. This is because full precision matrices for high-order features are difficult to get due to the over-fitting caused by large number of parameters and the cost calculation of inverting high order matrix. But this problem doesn't exist in the RBM modeling. The inter-dimensional correlations can be modeled by the weight parameter $\boldsymbol{W}$ and can be better learned by CD algorithm with Gibbs sampling. Thus,

RBMs can be trained on not only high-level spectral features but also directly on the raw spectral envelopes.

2.2.4. *Spectral conversion*

In conversion phase, given an input $\boldsymbol{X}_t$, the mean of the approximate Gaussian distribution $P(\boldsymbol{Y}_t|\boldsymbol{X}_t, \lambda_{m^*}^{(v)})$ is estimated by

$$\boldsymbol{Y}_t^* \doteq \arg\max_{\boldsymbol{Y}_t} P(\boldsymbol{Y}_t|\boldsymbol{X}_t, \lambda_{m^*}^{(v)}) = \arg\max_{\boldsymbol{Y}_t} P(\boldsymbol{X}_t, \boldsymbol{Y}_t). \tag{10}$$

where $P(\boldsymbol{X}_t, \boldsymbol{Y}_t)$ is the joint distribution of the visible units given by the RBM corresponding to the m^*-th sub-space. The index m^* is ignored to simplify the notation. The last term is get by applying Bayesian rules because $\boldsymbol{Y}_t$ is irrelevant to $P(\boldsymbol{X}_t)$. There is no closed-form solution for (10), the gradient descent algorithm can be adopted:

$$\boldsymbol{Y}_t^{(i+1)} = \boldsymbol{Y}_t^{(i)} + \gamma \cdot \frac{\partial \log P(\boldsymbol{X}_t, \boldsymbol{Y}_t)}{\partial \boldsymbol{Y}_t}\bigg|_{\boldsymbol{Y}_t = \boldsymbol{Y}_t^{(i)}}, \tag{11}$$

where i denotes the iteration number, γ is the step size and

$$\frac{\partial \log P(\boldsymbol{X}_t, \boldsymbol{Y}_t)}{\partial \boldsymbol{Y}_t} = -(\boldsymbol{Y}_t - \boldsymbol{a}^{(y)}) + \sum_{j=1}^{H} \frac{\exp(b_j + \boldsymbol{v}_t^\top \boldsymbol{w}_j)}{(1 + \exp(b_j + \boldsymbol{v}_t^\top \boldsymbol{w}_j))} \boldsymbol{w}_j^{(y)}. \tag{12}$$

Where $\boldsymbol{w}_j$ is the j-th column vector of $\boldsymbol{W}$. In contrast to the GMM-based conversion, $\boldsymbol{Y}_t^*$ isn't the average of training samples any more.

However, unlike that in HMM-based speech synthesis systems [8], the gradient based searching process must be performed frame-by-frame in real-time conversion. This is very time consuming and is nearly unbearable attribute to the nonlinear term in calculation of log probability derived from (7):

$$\log P(\boldsymbol{v}) = -\sum_{i=1}^{V}(a_i - v_i)^2 + \sum_{j=1}^{H} f(b_j + \boldsymbol{v}^\top \boldsymbol{w}_j) - \log \mathcal{Z}, \tag{13}$$

where $f(x) = \log(1 + \exp(x))$. But we can approximate $f(x)$ with a simple function:

$$g(x) = \begin{cases} x & x \geq 0, \\ 0 & x < 0. \end{cases} \tag{14}$$

Figure 1 indicates that $g(x)$ is a good approximation for $f(x)$ when $|x| > 4$, and we observed in our experiments that most of the activation term of hidden units $s_j = b_j + \boldsymbol{v}^\top \boldsymbol{w}_j$ is either very big or very small (eg. as we counted in the training set, $p(|s| > 4) = 0.94$), so this approximation should be reasonable. Therefore, we got an approximated solution for mean generation:

$$\boldsymbol{Y}_t^* \doteq \boldsymbol{a}^{(y)} + \sum_{j \in \{s_j > 0\}} \boldsymbol{w}_j^{(y)}. \tag{15}$$

An initial value of $\boldsymbol{Y}_t$ should be set to determinate which units are activated for the input frame. In this paper, the initial value is set to the target part of the mode of the RBM [8].

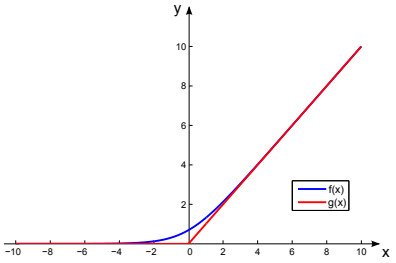

Figure 1: Approximation of function $f(x) = \log(1 + \exp(x))$.

3. Experiments

3.1. Experimental conditions

We used a Chinese speech corpus database with two speakers in our experiments. The source speaker is female and the target is male. 100 parallel sentences were used. We randomly selected 80 sentences for model training and used the remaining 20 sentences for test.

Two kinds of spectral features were used in our experiments: mel-cepstrum and original spectral envelope. 513-order spectral envelopes were calculated by STRAIGHT analysis at 5ms frame shift with 1024 FFT length, then 40-order mel-cepstra (not including the 0-th frame power coefficients) were extracted from the spectral envelopes. In spectral envelope modeling, log spectral envelopes were adopted. Dynamic time warping (DTW) algorithm was adopted to align the training feature sequences of two speakers. Spectral envelopes were aligned using the time warp information of corresponding mel-cepstra. An 128-mixture GMM was trained on mel-cepstra, and the feature space was divided into 128 sub-spaces by this GMM.

The RBMs were trained by CD learning with 1-step Gibbs sampling (CD1) and the learning rate is 0.0001. Batch size for batch gradient descent learning was set to 10 and 200 epochs were executed in model parameter estimation. The training samples in each sub-space were normalized to zero mean and unit variance before RBM training. For computational efficiency, the number of hidden units was set to 20 for mel-cepstra and 100 for spectral envelopes.

3.2. Subjective evaluation

In order to evaluate the performance of the proposed method. Five systems were built for comparison[2]:

- **MCEP-GMM** The baseline GMM-based method in section 2.1, mel-cepstra were used as spectral features.
- **MCEP-RBM** The proposed method for mel-cepstrum modeling.
- **SPE-RBM** The proposed method for spectral envelope modeling.
- **SPE-GMM** Built in the similar manner with SPE-RBM, the difference is that a single Gaussian distribution was adopted for spectral envelope modeling in each sub-space.
- **SPE-RBM-AP** The only difference between this system and SPE-RBM is that the approximated conversion function (15) is used.

[2]Some speech examples converted by these systems can be found at `http://home.ustc.edu.cn/~chenlh/demo/IS2013_SMCRBM.html`.

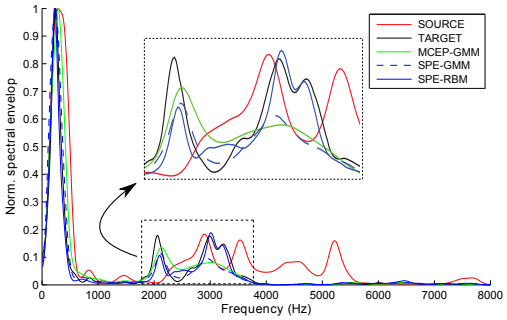

Figure 2: Examples of source, target and converted spectral envelopes, spectral envelopes were normalized by their maximum value.

The latter 3 systems were trained on sub-spaces divided by MCEP-GMM. Similarly, in conversion stage of spectral envelope systems, mel-cepstra were also extracted to determine which sub-space each frame belongs to. Twenty sentences in test set were converted by each system for the evaluations. Time cost in SPE-RBM-AP was about less than 1 minute for each sentence. It was greatly decreased comparing with SPE-RBM, in which each sentence took about 1 hour with 20 iterations.

Figure 2 presents some representative examples of source, target and converted spectral envelopes. In MCEP-GMM system, converted spectral envelopes were recovered from converted mel-cepstra. The formants of SPE-RBM spectral envelope is much closer to that of target. Moreover, comparing with the other 2 converted spectral envelopes, the SPE-RBM envelope has sharper formants and retained more detailed characteristics, which indicates that the over-smoothing effect in the proposed method was decreased.

Table 1: Subjective preference scores (%) between MCEP-GMM and MCEP-RBM systems, where N/P denotes "No Preference", p means the p-value of t-test between two systems.

	MCEP-GMM	MCEP-RBM	N/P	p
Similarity	34	43	23	0.00
Naturalness	15	71	14	0.00

We took several subjective listening tests to evaluate these systems. Seven listeners took apart in these tests. Figure 3 shows the mean opinion scores (MOS) (5 scale: 1-worst, 5-best) for MCEP-GMM, SPE-GMM and SPE-RBM. We see that the proposed SPE-RBM system is significantly better than the conventional GMM systems. SPE-GMM is also better than MCEP-GMM, especially in naturalness.

A comparison between MCEP-RBM and MCEP-GMM was also taken. Results shown in Table 1 indicate that MCEP-RBM is superior to MCEP-GMM. The feedback from listeners was that they can tell the difference but the improvement is not significant as that between SPE-RBM and SPE-GMM. This can be attribute to that spectral envelopes have stronger inter-dimensional correlations which can be well modeled by RBMs. Although mel-cepstra have weaker inter-dimensional correlations, the inter-speaker correlations are still better modeled by RBMs. That's why MCEP-RBM outperformed MCEP-GMM.

Then, a preference test between SPE-RBM and SPE-RBM-AP was taken. The preference scores were shown in Table 2.

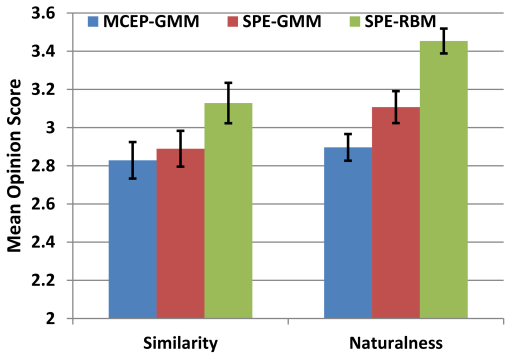

Figure 3: Mean opinion score (MOS) of similarity and naturalness. Error bars shows the 95% confidence interval.

Table 2: Subjective preference scores (%) between SPE-RBM and SPE-RBM-AP systems.

	SPE-RBM	SPE-RBM-AP	N/P	p
Similarity	23	26	51	0.66
Naturalness	17	57	26	0.00

We see that there isn't significant difference between the two systems in similarity. But in naturalness, SPE-RBM-AP is significantly better. We can see from Table 3 that SPE-RBM-AP generated spectral envelopes with higher probability, and the percentage of activations that have absolute value larger than 4 ($p(|s| > 4)$) in SPE-RBM-AP is higher than that in SPE-RBM, which means the hidden units of SPE-RBM have higher determinacy. And the conversion by the approximated solution is more stable than the gradient based searching.

Table 3: Average log output probability and $p(|s| > 4)$ in the test set. The joint features are joint of input and converted features.

	Avg. log output prob.	$p(\|s\| > 4)$
SPE-RBM	-2557.78	89.24%
SPE-RBM-AP	-2439.65	95.49%

Our proposed conversion method for spectral envelopes was superior to the conventional over-smoothing methods. Some informal listening tests show that its performance can be further improved by combing GV based parameter generation, especially in MCEP-RBM.

4. Conclusions

We have proposed an RBM-base spectral modeling and conversion method in this paper. The joint acoustic space are divided into several sub-spaces, each of which is modeled by an RBM. The RBM can be trained directly on STRAIGHT spectral envelope. At conversion stage, acoustic features are converted by maximizing the conditional output probability criterion. Experiments show that the proposed method significantly outperformed the conventional GMM based methods. Even so, the proposed method still needs to be built based on GMM systems, therefore, using mixture of RBM to directly model the entire acoustic space will be a future work.

5. References

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

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

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

[4] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed gmm and map adaptation," in *Eurospeech*, 2003, pp. 2413–2416.

[5] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory hmms," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 417–430, 2011.

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

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

[8] Z. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines for statistical parametric speech synthesis," in *Proc. ICASSP*, 2013(Accepted).

[9] 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.

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

[11] R. Salakhutdinov, "Learning deep generative models," Ph.D. dissertation, University of Toronto, 2009.