USING CONDITIONAL RESTRICTED BOLTZMANN MACHINES FOR SPECTRAL ENVELOPE MODELING IN SPEECH BANDWIDTH EXTENSION

Yingxue Wang[1,3]*, Shenghui Zhao*[1]*, Dan Qu*[2,3]*, Jingming Kuang*[1]

[1]School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
[2]Zhengzhou Institute of Information Science and Technology, Zhengzhou 450002, China
[3]School of Computer Science, Carnegie Mellon University, US

yxwang.bit@gmail.com, shzhao@bit.edu.cn

ABSTRACT

In this paper, we present a conditional restricted Boltzmann machine (CRBM) based speech bandwidth extension (BWE) method. A CRBM is employed to obtain time information and model deep non-linear relationships between the spectral envelope features of low frequency (LF) and high frequency (HF). Two exclusive CRBMs are adopted to model the distribution of LF's and HF's spectral envelope features. respectively. A neural network (NN) is then used to model the joint distribution of hidden variables extracted from the two CRBMs. The proposed method takes advantage of the strong ability of CRBM in discovering the temporal correlation between adjacent frames and modeling deep non-linear relationships between input and output. Both the objective and subjective evaluations indicate that our proposed method outperforms the conventional Gaussian mixture model based methods and other NN based methods.

Index Terms— Speech bandwidth expansion, conditional restricted Boltzmann machine, Gaussian mixture model

1. INTRODUCTION

Speech bandwidth expansion (BWE) is a technique that regenerates the missing high frequency parts in order to improve the quality of speech. Many approaches have been proposed for BWE during the last decades. Among these approaches, mapping method based on Gaussian Mixture Model (GMM) [1] is widely used and a number of improvements [2, 3, 4, 5] have been proposed.

However, the derived mapping function by GMM is a piece-wise linear transformation which is maybe insufficient to describe the complex non-linear relationship between the low frequency (LF) and high frequency (HF). To capture the characteristics of speech more precisely, a deeper non-linear architecture is required. One example of deeper BWE methods was imposed by [6] based on deep neural network (DNN). Previously, we also proposed to use the restricted Boltzmann machines (RBM) [7] to obtain a non-linear relationship between LF and HF. Although these approaches were reported to outperform the GMM-based approaches, they assumed the speech frames were independent of each other, ignoring the temporal information of speech. The importance of the temporal information as well as its advantage had been investigated appropriately in several literatures. In [8, 9, 10], temporal information was included in GMM-based method by utilizing the delta features of spectral envelopes. In [11], time correlation properties of speech were embedded into spectrum estimation by utilizing the Hidden Markov Model (HMM).

In this paper, we propose to use conditional restricted Boltzmann machine (CRBM) for spectral envelope modeling to avoid the frame-by-frame independence assumption in most of the speech BWE methods. We do this by combining two CRBMs and a concatenation neural network (NN). A CRBM is a non-linear probabilistic model used to capture temporal information of speech and obtain high-order feature space where speech features are converted more easily than in an original acoustic feature space. In our approach, we first train two exclusive CRBMs for the LF and HF, aiming to capture high-order features in an unsupervised manner. Then we train a NN using the projected features produced by two CRBMs. The important advantage of our proposed method is its ability to capture the temporal correlation between adjacent frames of speech and obtain the deep non-linear relationships between the spectral envelope features of LF and HF.

2. SPEECH BANDWIDTH EXTENSION BASED ON GMM

Let $\boldsymbol{x}_t = [x_1, x_2, \cdots, x_m]^{\mathrm{T}}$ and $\boldsymbol{y}_t = [y_1, y_2, \cdots, y_n]^{\mathrm{T}}$ be the m dimensional LF and n dimensional HF feature vectors at frame t, respectively. The operator $[\cdot]^{\mathrm{T}}$ denotes matrix transposition. In GMM based method, the joint probability density function of the joint feature space $\boldsymbol{z}_t = \left[\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{y}_t^{\mathrm{T}}\right]^{\mathrm{T}}$ is modeled by a GMM

$$f\left(z_t|\Theta^z\right) = \sum_{l=1}^{L} a_l^z \mathcal{N}\left(\boldsymbol{z}_t; \boldsymbol{\mu}_l^z; \boldsymbol{\Sigma}_l^z\right), \sum_{l=1}^{L} a_l^z = 1 \qquad (1)$$

where $\mathcal{N}(\cdot)$ denotes a Gaussian distribution of joint vector $\boldsymbol{z}_t$. Θ^z denotes a parameter set of the GMM, which consists

of weights, mean vectors and the covariance matrices for individual mixture components. The total number of mixture components is L. a_l^z, $\boldsymbol{\mu}_l^z$ and $\boldsymbol{\Sigma}_l^z$ are the weight, mean vector and covariance matrix of the l^{th} mixture component, and

$$\boldsymbol{\mu}_l^z = \begin{bmatrix} \boldsymbol{\mu}_l^x \\ \boldsymbol{\mu}_l^y \end{bmatrix} \qquad \boldsymbol{\Sigma}_l^z = \begin{bmatrix} \boldsymbol{\Sigma}_l^{xx} & \boldsymbol{\Sigma}_l^{xy} \\ \boldsymbol{\Sigma}_l^{yx} & \boldsymbol{\Sigma}_l^{yy} \end{bmatrix} \tag{2}$$

$\boldsymbol{\mu}_l^x$ and $\boldsymbol{\mu}_l^y$ are the mean vectors of the l^{th} component for the LF and HF respectively, Σ_l^{xx} and $\boldsymbol{\Sigma}_l^{yy}$ are the corresponding covariance matrices, $\boldsymbol{\Sigma}_l^{xy}$ and $\boldsymbol{\Sigma}_l^{yx}$ are the cross-covariance matrices. The GMM is trained with the expectation maximization (EM) algorithm using the joint vectors as the input feature vectors.

The idea of BWE is to determine a mapping function that can approximate HF features accurately. A mapping function to convert the LF feature vector $\boldsymbol{x}_t$ to the HF feature vector $\boldsymbol{y}_t$ is derived based on the conditional probability density of the HF feature vector, given the LF feature vector. When minimizing the mean square error (MMSE) estimation rule is adopted for parameter generation, the mapping function takes the form:

$$\begin{aligned} \widetilde{\boldsymbol{y}}_t &= \int_{\Omega_y} \boldsymbol{y}_t f_{y|x}\left(\boldsymbol{y}_t|\boldsymbol{x}_t\right) \\ &= \sum_{l=1}^{L} p_l\left(x\right)\left[\boldsymbol{\mu}_l^y + \boldsymbol{\Sigma}_l^{yx}\left(\boldsymbol{\Sigma}_l^{xx}\right)^{-1}\left(\boldsymbol{x}_t - \boldsymbol{\mu}_l^x\right)\right], \end{aligned} \tag{3}$$

where $p_l\left(x\right)$ is the probability of $\boldsymbol{x}_t$ belonging to the l^{th} component, i.e

$$p_l\left(x\right) = \frac{a_l \mathcal{N}\left(\boldsymbol{x}_t, \boldsymbol{\mu}_l^x, \boldsymbol{\Sigma}_l^{xx}\right)}{\sum_{m=1}^{L} a_m \mathcal{N}\left(\boldsymbol{x}_t, \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^{xx}\right)}. \tag{4}$$

According to Eq.3, the mapping function is a piece-wise linear transformation. The converted HF feature vector $\widetilde{\boldsymbol{y}}_t$ is determined by the current acoustic vector from LF, which means that the GMM based method ignores the time information.

3. PROPOSED SPEECH BANDWIDTH EXTENSION

Our speech bandwidth extension system uses CRBM to capture high-order features and time-related information. We briefly discuss the main idea of CRBM as well as its parameters estimation and give details of our proposed method in this section.

3.1. CRBM

CRBM is the conditional form of RBM proposed by Taylor et al [12]. In this model, short-term temporal structures can be captured by making the visible and hidden units receive additional input from past and future states of visible units dynamically. Given a hidden vector $\boldsymbol{h}^t = [h_1^t, h_2^t, \cdots h_J^t]^{\mathrm{T}}$, $h_j^t \in \{0,1\}$ a visible vector $\boldsymbol{v}^t = [v_1^t, v_2^t, \cdots v_I^t]^{\mathrm{T}}$, $v_i^t \in \{0,1\}$ and a conditional vector $\boldsymbol{v}^{t-r}$(r is the number of previous frames from the current frame taken in account, here we choose $r = 1$ for simplicity) at the current frame t, the conditional probability could be defined as follows:

$$p(\boldsymbol{v}^t \mid \boldsymbol{v}^{t-1}) = \frac{1}{Z}\sum_{\boldsymbol{h}^t} \exp(-E(\boldsymbol{v}^t, \boldsymbol{h}^t \mid \boldsymbol{v}^{t-1})) \tag{5}$$

$$\begin{aligned} E(\boldsymbol{v}^t, \boldsymbol{h}^t \mid \boldsymbol{v}^{t-1}) = &-\boldsymbol{b}^{\mathrm{T}}\boldsymbol{v}^t - \boldsymbol{c}^{\mathrm{T}}\boldsymbol{h}^t - (\boldsymbol{v}^t)^{\mathrm{T}}\boldsymbol{W}^{v^t h^t}\boldsymbol{h}^t \\ &- (\boldsymbol{v}^{t-1})^{\mathrm{T}}\boldsymbol{W}^{v^{t-1}v^t}\boldsymbol{v}^t \\ &- (\boldsymbol{v}^{t-1})^{\mathrm{T}}\boldsymbol{W}^{v^{t-1}h^t}\boldsymbol{h}^t \end{aligned} \tag{6}$$

$$Z = \sum_{\boldsymbol{v}^t, \boldsymbol{h}^t} \exp(-E(\boldsymbol{v}^t, \boldsymbol{h}^t \mid \boldsymbol{v}^{t-1})) \tag{7}$$

where $\boldsymbol{b}$ and $\boldsymbol{c}$ are a bias vector of visible units and a bias vector of hidden units respectively. $\boldsymbol{W}^{v^t h^t}$, $\boldsymbol{W}^{v^{t-1}v^t}$ and $\boldsymbol{W}^{v^{t-1}v^t}$ are the weight matrices between $\boldsymbol{v}^t$ and $\boldsymbol{h}^t$, $\boldsymbol{v}^{t-1}$ and $\boldsymbol{v}^t$, $\boldsymbol{v}^{t-1}$ and $\boldsymbol{h}^t$ respectively.

In this model, there are five parameters to be estimated: $\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{W}^{v^{t-1}h^t}, \boldsymbol{W}^{v^t h^t}, \boldsymbol{W}^{v^{t-1}v^t}$. These parameters are estimated by maximizing the log-likelihood $\mathcal{L} = \log\prod_t p\left(\boldsymbol{v}^t \mid \boldsymbol{v}^{t-1}\right)$. Differentiating partially with respect to each parameter, we obtain

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{ij}^{v^t h^t}} = \left\langle v_i^t h_j^t \right\rangle_{data} - \left\langle v_i^t h_j^t \right\rangle_{model} \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{i'i}^{v^{t-1}v^t}} = \left\langle v_i^t v_{i'}^{t-1} \right\rangle_{data} - \left\langle v_i^t v_{i'}^{t-1} \right\rangle_{model} \tag{9}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{i'j}^{v^{t-1}h^t}} = \left\langle v_{i'}^{t-1} h_j^t \right\rangle_{data} - \left\langle v_{i'}^{t-1} h_j^t \right\rangle_{model} \tag{10}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_i} = \left\langle v_i^t \right\rangle_{data} - \left\langle v_i^t \right\rangle_{model} \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{c}_j} = \left\langle h_j^t \right\rangle_{data} - \left\langle h_j^t \right\rangle_{model} \tag{12}$$

where $\langle\cdot\rangle_{data}$ and $\langle\cdot\rangle_{model}$ indicate the expectations of the input data and the inner model. Because $\langle\cdot\rangle_{model}$ is extremely expensive to compute exactly, the contrastive divergence approximation to the gradient is used, where $\langle\cdot\rangle_{model}$ is replaced by running the Gibbs sampler initialized at the data for one full step [13].

Once the parameters are estimated, the conditional probability of $\boldsymbol{h}^t$ given $\boldsymbol{v}^t$ and $\boldsymbol{v}^{t-1}$ and the conditional probability of $\boldsymbol{v}^t$ given $\boldsymbol{h}^t$ and $\boldsymbol{v}^{t-1}$ are respectively written as:

$$p\left(\boldsymbol{h}_j^t = 1 \mid \boldsymbol{v}^t, \boldsymbol{v}^{t-1}\right) = \sigma\left(c_j + (\boldsymbol{v}^t)^{\mathrm{T}}\boldsymbol{W}_{:j}^{v^t h^t} + (\boldsymbol{v}^{t-1})^{\mathrm{T}}\boldsymbol{W}_{:j}^{v^{t-1}h^t}\right) \tag{13}$$

$$p\left(\boldsymbol{v}_i^t = 1 \mid \boldsymbol{h}^t, \boldsymbol{v}^{t-1}\right) = \sigma\left(b_i + (\boldsymbol{h}^t)^{\mathrm{T}}(\boldsymbol{W}_{i:}^{v^t h^t})^{\mathrm{T}} + (\boldsymbol{v}^{t-1})^{\mathrm{T}}\boldsymbol{W}_{:j}^{v^{t-1}v^t}\right) \tag{14}$$

where $\boldsymbol{W_{i:}}$ and $\boldsymbol{W_{:j}}$ denote the column vector and the row vector in $\boldsymbol{W}$ respectively, and σ indicates a sigmoid function; i.e. $\sigma\left(x\right) = \frac{1}{1+e^x}$.

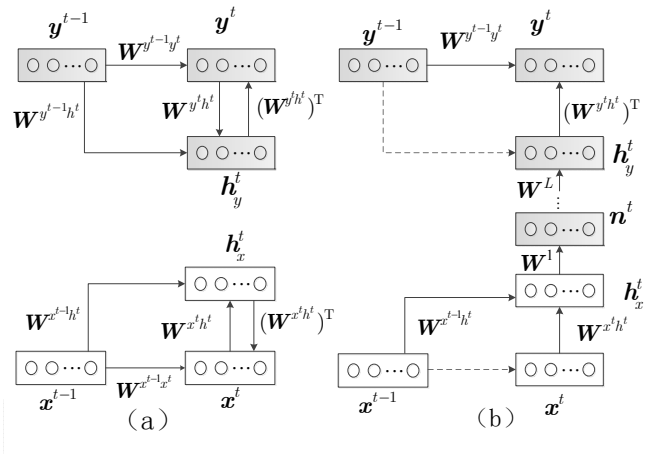

Fig. 1. *(a) CRBMs for the low frequency (below) and high frequency (above) (b)our proposed speech bandwidth extension architecture combining two CRBMs and a NN*

3.2. Speech bandwidth extension using CRBM

Figure 1 shows an overview of our proposed speech BWE system. In our approach, two CRBMs (one for the low frequency and the other for the high frequency) are adopted to describe the distribution of $\boldsymbol{x}^t$, $\boldsymbol{x}^{t-1}$ and $\boldsymbol{y}^t$, $\boldsymbol{y}^{t-1}$ respectively. $\boldsymbol{x}^t$ and $\boldsymbol{x}^{t-1}$ are spectral envelop features of the LF at frame t and $t-1$ respectively. $\boldsymbol{y}^t$ and $\boldsymbol{y}^{t-1}$ are spectral envelope features of the HF at frame t and $t-1$ respectively. Then a NN is employed to model the joint distribution of $\boldsymbol{h}_x^t$ and $\boldsymbol{h}_y^t$. $\boldsymbol{h}_x^t$ and $\boldsymbol{h}_y^t$ are the hidden variables extracted from two CRBMs respectively. The extracted hidden variables can be considered as the high-order binary representation of the raw spectral envelopes. Therefore, the parameter set of our proposed method is given by

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_x, \boldsymbol{\theta}_y, \boldsymbol{\theta}_n\} \tag{15}$$

where $\boldsymbol{\theta}_x = \left\{\boldsymbol{W}^{x^t x^{t-1}}, \boldsymbol{W}^{x^{t-1} h^t}, \boldsymbol{W}^{x^t h^t}, \boldsymbol{b_x}, \boldsymbol{c}_x\right\}$ are the parameters of the CRBM using the training spectral envelopes of LF, $\boldsymbol{\theta}_y = \left\{\boldsymbol{W}^{y^t y^{t-1}}, \boldsymbol{W}^{y^{t-1} h^t}, \boldsymbol{W}^{y^t h^t}, \boldsymbol{b_y}, \boldsymbol{c}_y\right\}$ are the parameter of the CRBM using the training spectral envelopes of HF, $\boldsymbol{\theta}_n = \{\boldsymbol{W}^1, \cdots, \boldsymbol{W}^L, \boldsymbol{d}^1, \cdots, \boldsymbol{d}^L\}$ ($L-1 = \{0, 1, 2\}$ is the number of the hidden layers, $\boldsymbol{d}^l$ is the bias vector of the l^{th} layer, $\boldsymbol{W}^l$ is the weight matrix from the $(l-1)^{th}$ layer to the l^{th} layer) are the parameters of the NN trained using the extracted hidden variables. The weight matrices $\boldsymbol{W}^{x^t x^{t-1}}, \boldsymbol{W}^{x^{t-1} h^t}, \boldsymbol{W}^{y^t y^{t-1}}, \boldsymbol{W}^{y^{t-1} h^t}$ can absorb time-related information.

At the training phase of the NN, the input vectors and the output vectors are written as

$$\boldsymbol{h}_x^t = \sigma\left(c_x + \boldsymbol{W}^{x^t h^t}\boldsymbol{x}^t + \boldsymbol{W}^{x^{t-1} h^t}\boldsymbol{x}^{t-1}\right) \tag{16}$$

$$\boldsymbol{h}_y^t = \sigma\left(c_y + \boldsymbol{W}^{y^t h^t}\boldsymbol{y}^t + \boldsymbol{W}^{y^{t-1} h^t}\boldsymbol{y}^{t-1}\right) \tag{17}$$

Weight parameters $\boldsymbol{\theta}_n$ of NN are estimated by minimizing the error between the output $\widetilde{\boldsymbol{h}}_y^t$ of NN and the target vector $\boldsymbol{h}_y^t$. Once the parameters of NN are obtained, the input vector $\boldsymbol{h}_x^t$ can be converted to:

$$\widetilde{\boldsymbol{h}}_y^t = \boldsymbol{o}^l = \sigma\left(\boldsymbol{s}^l\right), l \geq 1 \tag{18}$$

where $\boldsymbol{s}^l = \boldsymbol{W}^l \boldsymbol{o}^{l-1} + \boldsymbol{d}^l$ and $\boldsymbol{o}^{l-1} = \sigma\left(\boldsymbol{s}^{l-1}\right), \boldsymbol{o}^0 = \boldsymbol{h}_x^t$.

To map the output of the NN to the spectral envelope features of the HF, we can use backward inference of CRBM using Eq.14,

$$p\left(\boldsymbol{y}^t \mid \widetilde{\boldsymbol{h}}_y^t, \boldsymbol{y}^{t-1}\right) = \sigma\left(b_y + (\boldsymbol{W}^{y^t h^t})^{\mathrm{T}}\widetilde{\boldsymbol{h}}_y^t + \boldsymbol{W}^{y^{t-1} y^t}\boldsymbol{y}^{t-1}\right) \tag{19}$$

According to Eq.18 and Eq.19, the mapping function of our method from a LF's spectral envelope feature vector $\boldsymbol{x}^t$ to a HF's spectral envelope feature vector $\boldsymbol{y}^t$ at frame t, given the previous feature vectors $\boldsymbol{x}^{t-1}$ and $\boldsymbol{y}^{t-1}$ is given as:

$$\boldsymbol{y}^t = \sigma\left(b_y + (\boldsymbol{W}^{y^t h^t})^{\mathrm{T}}\sigma\left(\boldsymbol{s}^l\right) + \boldsymbol{W}^{y^{t-1} y^t}\boldsymbol{y}^{t-1}\right) \tag{20}$$

As Eq.20 indicates, the mapping function is a composite function of multiple different non-linear functions. We need a current acoustic vector from LF, and previous vectors from both LF and HF to reconstruct the HF's current acoustic vector. Thus, additional information from adjacent signal frames can be captured and deep non-linear relationships between spectral envelope features of LF and HF can be discovered by our proposed method.

4. EXPERIMENTS

4.1. setup

We conducted speech bandwidth extension experiments using two Chinese speech databases, comparing our method (CRBM) with the traditional GMM-based method (GMM) and our previous work (RBM). The first Chinese speech database is from the NTT Advanced Technology Corporation (NTT-AT). The second database is from Ericsson and Beijing Institute of Technology (EBIT). The data in the two databases is sampled at a 16-kHz sampling rate with 16-bits resolution. Each utterance in the two databases lasts 8s. A high-pass filtering supplied the high frequency signal. The low frequency signal resulted from a 0.3 to 3.4 kHz band-pass filtering followed by a down-sampling and up-sampling with a factor 2. We used 64 utterances randomly selected from all speech sound classes in NTT and 64 utterances in EBIT as our training set. The test set consisted of the 32 utterances in NTT and the 32 utterances in EBIT that were not included in the training data.

A GMM with 128 components was trained for the baseline system. The 16-order and 10-order line spectral frequencies (LSFs) [14] were adopted as the spectral envelope features for the LF and HF respectively. LSFs are first normalized to have zero mean and unit variance, and then converted to binary using a sigmoid function before feeding them

into CRBMs. The frame size and the frame shift for calculating spectral envelopes was set to 20ms and 10ms respectively. We investigated on three neural network architectures (arc.1: no hidden layer, arc.2: 1 hidden layer and 128 hidden units in the hidden layer, arc.3: 2 hidden layers and 128 hidden units in each hidden layer) for the following experiments. For each architecture, the contrastive divergence (CD) learning with 10-step Gibbs sampling was employed to train two CRBMs. The stochastic batch gradient descent algorithm was adopted to update the model parameters. The size of each mini-batch was set to 64. The learning rate and momentum were set to 0.0001 and 0.9. The number of epochs of CRBMs and NNs were set to 500 and 300 respectively. The number of hidden units of two CRBMs was fixed to 64. We used the LF excitation signal as HF excitation signal. In order to adjust the power of the extended HF excitation signal, we builded a one-to-one codebook from the LSFs of LF to gains between the HF signal and the synthesized signal that the LF excitation signal filtered through the high frequency synthesis filter.

To assess the overall quality of reconstructed speech, we conducted objective evaluation and subjective evaluation. For the objective evaluation, we used root mean square log-spectral distortion (RMS-LSD) to measure how close the reconstructed speech is to the original one. We calculated the RMS-LSD for each frame and averaged the RMS-LSD values. For the subjective evaluation, mean opinion score (MOS) listening tests were conducted.

4.2. objective evaluation

We measured the RMS-LSD in the missing high frequency (4-8 kHz). The definition of RMS-LSD is as follows,

$$D = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{1}{w_2 - w_1}\int_{w_1}^{w_2}\left[20\log\left(\mathrm{G}_c\frac{\mathrm{A}_n(w)}{\hat{\mathrm{A}}_n(w)}\right)\right]^2 dw} \quad (21)$$

$$\mathrm{G}_c = \frac{1}{w_2 - w_1}\int_{w_1}^{w_2}\left[20\log\left(\frac{\hat{\mathrm{A}}_n(w)}{\mathrm{A}_n(w)}\right)\right] dw, \quad (22)$$

where $A_n(w)$ and $\widehat{A}_n(w)$ denote the original and the reconstructed power spectrum of n^{th} frame of high frequency respectively, w_1, w_2 are the lower and higher bound of the missing high frequency, compensating gain factor G_c has the effect of removing the mean difference between the two log envelopes. The smaller the value of RMS-LSD is, the closer the reconstructed high frequency is to the original high frequency, the better the speech quality is. The RMS-LSD results are shown in Figure 2. Experimental results show that CRBM based method has lower log spectral distortion than GMM and RBM based method.

4.3. subjective evaluation

For MOS tests, 12 participants were asked to listen to the original speech signals in the test set and the reconstructed speech signals for each method, and to select how close the reconstructed speech is to the original one on a 5-point scale (1-bad,2-poor,3-fair,4-good,5-excellent). Figure 2 and figure 3 summarize the objective and subjective experimental results. As shown in these figures, "CRBM" based method outperforms other conversional methods (GMM, RBM) in both criteria. The reason for the improvement is attributed to the fact that CRBM can capture time-related information and nonlinear relationship between spectral envelope features of LF and HF.

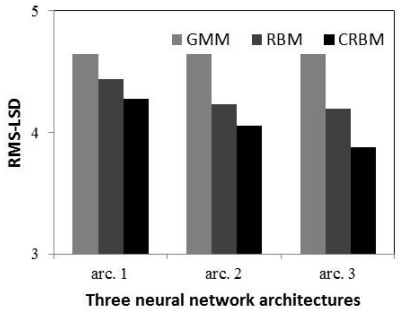

Fig. 2. *Results of Objective evaluation*

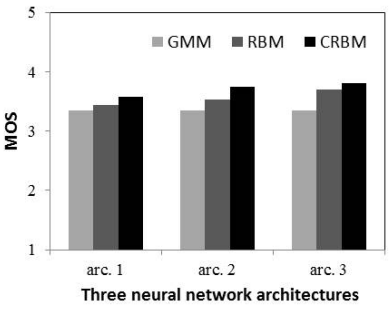

Fig. 3. *Results of Subjective evaluation*

5. CONCLUSION

In this paper, we proposed a new speech bandwidth extension method that combined two conditional restricted Boltzmann machines and a neural network to construct a non-linear relationships between the low frequency's and high frequency's spectral envelope features and capture the time-related information. The objective evaluation and subjective evaluation showed improvement of the proposed method when compared with the conventional GMM based method and other network based method.

6. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (Grant No: 61175067).

7. REFERENCES

[1] Kun-Youl Park and Hyung Soon Kim, "Narrowband to wideband conversion of speech using gmm based transformation," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 3, pp. 1843–1846.

[2] Xin Liu, Chang-Chung Bao, Mao-shen Jia, and Yongtao Sha, "A harmonic bandwidth extension based on gaussian mixture model," in *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 474–477.

[3] Hannu Pulakka, Ulpu Remes, Kalle Palomäki, Mikko Kurimo, and Paavo Alku, "Speech bandwidth extension using gaussian mixture model-based estimation of the highband mel spectrum," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5100–5103.

[4] Murali D Mohan, Dileep B Karpur, Manoj Narayan, and J Kishore, "Artificial bandwidth extension of narrowband speech using gaussian mixture model," in *Communications and Signal Processing (ICCSP), 2011 International Conference on*. IEEE, 2011, pp. 410–412.

[5] Hannu Pulakka, Ulpu Remes, Santeri Yrttiaho, Kalle Palomäki, Mikko Kurimo, and Paavo Alku, "Bandwidth extension of telephone speech to low frequencies using sinusoidal synthesis and a gaussian mixture model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2219–2231, 2012.

[6] Kehuang Li and Chin-Hui Lee, "A deep neural network approach to speech bandwidth expansion," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4395–4399.

[7] Yingxue Wang, Shenghui Zhao, Wenbo Liu, Ming Li, and Jingming Kuang, "Speech bandwidth expansion based on deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Amr H Nour-Eldin, Turaj Z Shabestary, and Peter Kabal, "The effect of memory inclusion on mutual information between speech frequency bands," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 3, pp. III–III.

[9] Amr H Nour-Eldin and Peter Kabal, "Combining frontend-based memory with mfcc features for bandwidth extension of narrowband speech," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4001–4004.

[10] Amr H Nour-Eldin and Peter Kabal, "Memory-based approximation of the gaussian mixture model framework for bandwidth extension of narrowband speech.," in *INTERSPEECH*, 2011, pp. 1185–1188.

[11] Peter Jax and Peter Vary, "Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 1, pp. I–680.

[12] Graham W Taylor, Geoffrey E Hinton, and Sam T Roweis, "Modeling human motion using binary latent variables," in *Advances in neural information processing systems*, 2006, pp. 1345–1352.

[13] Geoffrey E Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[14] Samir Chennoukh, A Gerrits, G Miet, and R Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. IEEE, 2001, vol. 1, pp. 665–668.