Speech Spectrum Restoration Based on Conditional Restricted Boltzmann Machine

Xugang Lu, Shigeki Matsuda, Chiori Hori

National Institute of Information and Communications Technology, Japan

Abstract

Many speech enhancement algorithms have been proposed for speech restoration from distorted speech. However, if some components of the signal are completely missed or distorted, there is no way for those algorithms to restore the clean speech. Considering that the restricted Boltzmann machine (RBM) is a stochastic version of the Hopfield network which can be used as an associative memory, we propose to use its "recall" ability for speech spectrum restoration when some parts of the speech spectrum are completely missed or distorted. Traditionally, in training the RBM, speech spectral patches are randomly selected as input. There is no consideration of the temporal correlation between different input spectral patches. In this study, we further propose to model this temporal correlation by using a conditional RBM (CRBM). The inference on the CRBM is almost the same as that of on the RBM by only modifying the biases as conditional dynamic biases. We did experiments for clean speech reconstruction and distorted speech restoration based on the trained models. Our experimental results showed that both the RBM and CRBM worked well in restoration task. By incorporating temporal correlation in the CRBM, a further improvement on reconstruction and restoration accuracy was achieved.

Index Terms: Restricted Boltzmann machine, conditional restricted Boltzmann machine, speech restoration.

10.21437/Interspeech.2013-722

1. Introduction

Reconstructing clean speech from noisy or distorted ones is one of the most important tasks in speech technology. Many noise reduction and speech enhancement algorithms have been proposed within these decades for this task [1]. Most of the algorithms try to design a gain function for signal filtering. The gain function is usually estimated based on tracking the signal to noise ratio (SNR) or something related to statistical information. However, if some components of the signal is completely missed or distorted, there is no way to restore the clean speech by using those algorithms.

Speech has well organized structures, such as phonemes, syllables, and words. Correspondingly, in speech spectrum, time-frequency patterns also have strong regular structures. The regular structures are distributed in different time-frequency bands with strong dependency. We argue that if these dependency structures are modeled in a generative model, the missed or distorted parts can be restored from the model based on the dependency modeling. Although traditional neural network can be used to learn the statistical regularity of speech for noise reduction [2], it is difficult to use the learned neural network to generate speech spectrum by only given parts of spectrum as observed inputs. Considering that the restricted Boltzmann machine (RBM) is a stochastic version of the Hopfield network which can be used as an associative memory, we propose to use its "recall" ability for speech spectrum restoration when some parts of the spectrum are completely missed or distorted.

The restricted Boltzmann machine (RBM) has already been widely used for data modeling and pattern recognition [3, 4]. Recently, it was successfully used in speech feature extraction and acoustic modeling for automatic speech recognition (ASR) in building a deep neural network [5]. In this study, we take a different application for using the RBM for speech restoration. The basic principle is that in training the RBM, data feature dependency is encoded in the model. When the distorted feature is given as input to the RBM, the hidden states of the RBM will response with learned parameters which try to generate the corresponding given data with learned patterns. In order to train the RBM as an associative memory for speech restoration, the RBM must be pretrained with a clean speech data set. Traditionally, in training, the visible input vectors are constructed from over-lapped speech spectral patches, and each vector is randomly selected from a training data set. There is no consideration of the local temporal correlation between the training patches. The learned parameters only reflect the static structure of the training data. However, the local temporal correlation structure is one of the most important information for speech signal (prior knowledge). We argue that if this prior knowledge is incorporated in modeling, the model will be more powerful than only modeling the static structure using the RBM.

Temporal restricted Boltzmann machine (TRBM) and conditional restricted Boltzmann machine (CRBM) have been applied in modeling human motion patterns [6, 7]. They share the same idea as incorporating sequential observations in training an RBM. We borrow the same idea to incorporate temporal correlation structure in speech spectrum modeling, and expect that the trained model will give accurate restoration even when input speech is distorted. The remainder of this paper is organized as follows. Section 2 introduces the CRBM learning framework for speech spectrum modeling by taking temporal correlation structures between spectral patches into consideration. In Section 3, we carry out experiments to evaluate the framework on speech spectrum reconstruction for clean speech and restoration for distorted speech. Discussions and conclusion are given in section 4.

2. Conditional restricted Boltzmann machine

In speech spectrum modeling based on the RBM, each spectral patch is randomly selected from a training data set. In order to incorporate local temporal correlation structure between spectral patches in modeling, we adopted the conditional restricted Boltzmann machine (CRBM) which was originally proposed in modeling human movement [7]. The CRBM can be used to capture the temporal dependency of the input vectors. The basic framework of the CRBM for speech processing is illustrated

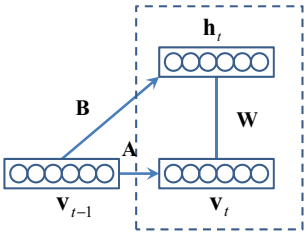

Figure 1: Conditional restricted Boltzmann machine for temporal correlation modeling.

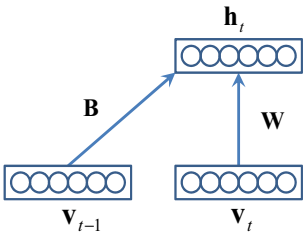

Figure 2: Estimation of hidden variable from visible variable.

in Fig. 1. In this figure, the dot-lined box is the traditional RBM. Two directed links are added to the RBM. The information flow only can be propagated alone the arrowed directions, while the information can be propagated in bi-directions in the RBM (undirected link). $\mathbf{v}_t$ is the current input visible variable (vector), $\mathbf{v}_{t-1}$ is the time delayed (with delay order 1) vector. These visible vectors are continuous spectral patches made from speech spectrum [8]. Compared with traditional modeling by the RBM, two transform matrices, $\mathbf{A}$ and $\mathbf{B}$ as shown in figure 1, are added corresponding to the two arrowed links. For simplicity, the linking matrix in visible layer is an autoregressive matrix that is used to model the temporal correlation between input vectors. Because the inference of the RBM is constrained by the previous visible variable, the inference of the RBM in Fig. 1 can be carried out as a conditional inference.

Two steps are applied for the inference in the CRBM. The first step is forward inference (given current and previous visible inputs to infer hidden state of neurons). Fig .2 illustrates the inference. The previous visible variable ($\mathbf{v}_{t-1}$) is used to adjust the bias of the hidden layer neurons as (dynamic hidden bias):

$$\hat{b}_{j,t} = b_j + B_{:,j}\mathbf{v}_{t-1}, \tag{1}$$

where b_j is the static hidden bias which is the same as used in traditional training of the RBM, $B_{:,j}$ is the jth column of matrix $\mathbf{B}$. The sigmoid function is used as the output of hidden unit as (suppose the input data is with mean and variance normalization):

$$p\left(h_{j,t}=1|\mathbf{v}_t,\mathbf{v}_{t-1}\right) = \sigma\left(\hat{b}_{j,t} + W_{:,j}\mathbf{v}_t\right), \tag{2}$$

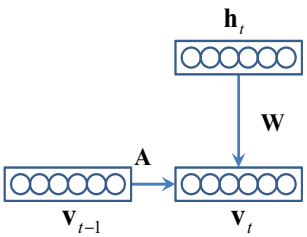

Figure 3: Estimation of visible variable from hidden variable.

where $\sigma(.)$ is the logistic sigmoid function, and $W_{:,j}$ is the jth column of matrix $\mathbf{W}$.

The second step is backward inference (given the current hidden state of neurons and previous visible input to infer current visible input). The inference is shown in Fig. 3. The previous visible input is used to adjust the bias of the visible layer neurons as (dynamic visible bias):

$$\hat{a}_{i,t} = a_i + A_{i,:}\mathbf{v}_{t-1}, \tag{3}$$

where a_i is the static visible bias, $A_{i,:}$ is the ith row of matrix $\mathbf{A}$. The estimation of the current visible variable is obtained as (Gaussian distribution with mean zero, and variance 1):

$$p\left(v_{i,t}|\mathbf{h}_t,\mathbf{v}_{t-1}\right) = \text{Gaussian}\left(\hat{a}_{i,t} + W_{i,:}\mathbf{h}_t, 1\right), \tag{4}$$

where $W_{i,:}$ is the ith row of matrix $\mathbf{W}$. For reconstruction, the visible input is estimated as in Eq. 5 which is the mean of the Gaussian distribution as:

$$\hat{v}_{i,t} = \hat{a}_{i,t} + W_{i,:}\mathbf{h}_t \tag{5}$$

Similar as in the RBM learning, parameter updating for the CRBM is given in the following (refer to [7] for details):

$$\begin{aligned}
\Delta W_{ij} &\propto \sum_t \left(\langle v_{i,t}h_{j,t}\rangle_{\text{data}} - \langle v_{i,t}h_{j,t}\rangle_{\text{CD}}\right)\\
\Delta A_{ki} &\propto \sum_t \left(\langle v_{i,t}v_{k,t-1}\rangle_{\text{data}} - \langle v_{i,t}v_{k,t-1}\rangle_{\text{CD}}\right)\\
\Delta B_{kj} &\propto \sum_t \left(\langle h_{j,t}v_{k,t-1}\rangle_{\text{data}} - \langle h_{j,t}v_{k,t-1}\rangle_{\text{CD}}\right)\\
\Delta a_i &\propto \sum_t \left(\langle v_{i,t}\rangle_{\text{data}} - \langle v_{i,t}\rangle_{\text{CD}}\right)\\
\Delta b_j &\propto \sum_t \left(\langle h_{j,t}\rangle_{\text{data}} - \langle h_{j,t}\rangle_{\text{CD}}\right),
\end{aligned} \tag{6}$$

where $\langle\cdot\rangle_{\text{data}}$ means average on data, while $\langle\cdot\rangle_{\text{CD}}$ denotes average on model reconstructed data based on contrastive divergence (CD) training algorithm [3]. In our study, only one step CD algorithm was used in data model estimation and parameter updating.

After the CRBM is trained, spectral pattern can be generated or predicted based on model parameters with given initial inputs. In addition, since the model encodes the dependency of data features, it is possible to use the model to restore input data when some components of the input are missed or distorted.

3. Experiments and evaluations

In this section, we evaluate the performance of the RBM and CRBM on speech spectrum reconstruction and restoration. A continuous English speech data set was used in our experiments. The raw speech feature was Mel filter band power spectrum (40 filter bands) extracted from windowed speech (20 ms frame size with 10 ms frame shift). Based on the Mel power spectrum, spectral patches were extracted from several continuous frames. Totally, 77450 spectral patches were used in training. Another 73460 spectral patches were used in testing. In the CRBM learning, a constant learning rate 0.005 and momentum 0.9 were used. The minibatch size was set as 128.

3.1. Clean speech spectrum reconstruction

We investigate the reconstruction accuracy for clean speech based on the RBM and CRBM. In the CRBM training, if the links between the previous visible vector and the RBM are cut out (refer to Fig. 1), the CRBM will degenerate to the traditional RBM. In this case, the input only takes each spectral patch

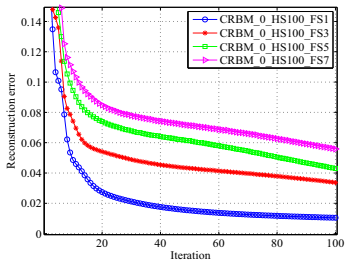

Figure 4: Reconstruction error curve in the RBM learning with different temporal window size for making spectral patches.

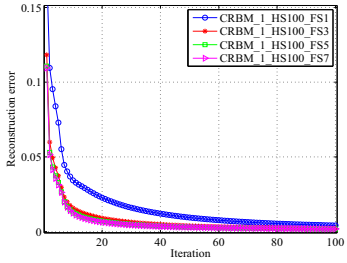

Figure 5: Reconstruction error curve in the CRBM learning conditioned on one previous visible input with different temporal window size for making spectral patches.

without considering their temporal correlations. But for taking temporal dependency between spectral patches in training the CRBM, we extracted the spectral patches in a longer window for each input training case (a continuous spectral patches segmented from continuous speech).

3.1.1. Selection of spectral patch size

Large spectral patches encode much more temporal correlation information than small ones. We want to examine the effect of patch size on reconstruction accuracy. An CRBM with hidden layer size of 100 was used. In extracting spectral patches, the temporal window size was set as 1, 3, 5, and 7 for experiment, respectively. The reconstruction error curves are shown in Figs. 4, and 5. In Figs. 4 and 5, "CRBM_#delay_HS#_FS#" means the learning of the CRBM conditioned on previous temporal input with delay number "#delay", hidden layer size of "HS#", and frame size of "HS#". For example, "CRBM_1_HS100_FS5" represents the CRBM learning conditioned on one previous temporal visible input with hidden layer size of 100 and spectral patch size as 5 frames of Mel spectrum. From Figs. 4 and 5, we can see that in the RBM training, large size spectral patch results in large reconstruction error. Therefore, we can not expect to use large spectral size to incorporate long temporal correlation information in reconstruction. But in the CRBM, the temporal correlation information is well captured with a consistently decrease in reconstruction error when large spectral patch was used. Considering that large spectral patch size results in large number of model parameters, in the following experiments, 5-frame spectral patches are used to make visible input vectors for the CRBM.

3.1.2. Effect of hidden layer size

The RBM with a large hidden layer size has much powerful modeling capacity than with a small one. However, the additional links to model temporal correlation in the CRBM may

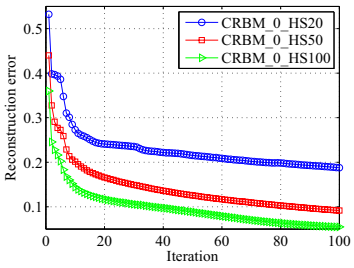

Figure 6: Reconstruction error curve in the RBM learning.

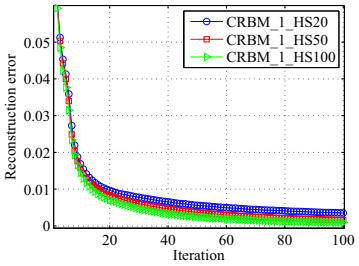

Figure 7: Reconstruction error curve in the CRBM learning conditioned on one previous visible input.

have different effects when hidden layer size is increased. We carried out experiments by setting hidden layer size as 20, 50 and 100. The reconstruction error curves were shown in Figs. 6 and 7. In these two figures, the size of the input vector is 200 (5 frames spectral patches). Comparing Figs. 6 and 7, we can see that if no temporal correlation structure between spectral patches is taken into consideration, large hidden layer size results in a significant reduction in reconstruct error. However, when the temporal correlation is explicitly modeled, no large change in reconstruction error with increasing of the hidden layer size. For further comparison, we draw the reconstruction error curves for CRBM_0_HS100, CRBM_1_HS100, and CRBM_2_HS100 in Fig. 8. From this figure, we can see that if temporal correlation structure between spectral patches is explicitly modeled in the CRBM, significant reconstruction error is reduced, but adding previous visible observation beyond 1 helps less.

After each spectral patch is reconstructed, we averaged on overlapped frames and reshaped the data to be Mel power spectrum. An example is shown in Fig. 9 to illustrate the reconstruction of clean speech spectrum. In this figure, top panel is the clean speech spectrum, the middle and bottom panels are the reconstructed (with CD-1 Gibbs sampling for reconstruction) from the CRBM_0_HS100 and CRBM_1_HS100, respectively. From this figure, we can see that the spectrum is reconstructed

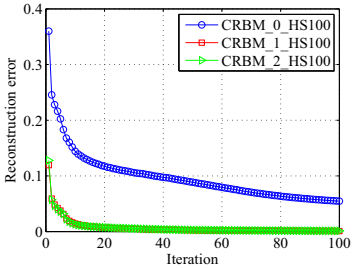

Figure 8: Comparison of reconstruction curves for the CRBM learning conditioned on one and two previous visible inputs.

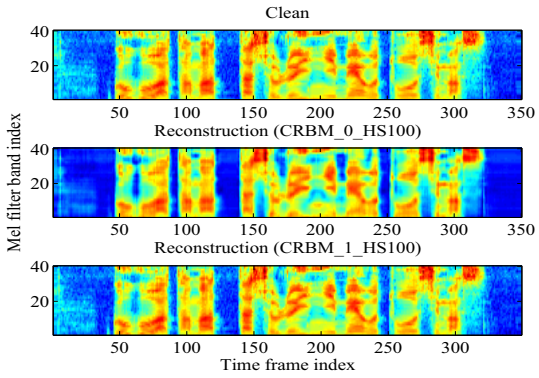

Figure 9: Mel spectrum reconstruction based on CRBM_0_HS100 and CRBM_1_HS100.

well by both the RBM and CRBM modeling. But a better reconstruction of the spectrum was achieved when temporal correlation structure was modeled by the CRBM. Quantitatively, we measure the distortion as average of the absolute difference between clean and reconstructed speech spectrum as (in dB since the Mel spectrum is in dB):

$$\text{Dist} = \frac{1}{\text{sum(t)}} \sum_t |\hat{\mathbf{v}}_t - \mathbf{v}_t|, \tag{7}$$

where $\hat{\mathbf{v}}_t$ is the reconstructed visible input (with Gibbs sampling for reconstruction). The reconstruction errors (average on each dimension) for the training and testing data sets are shown in Tab. 1. From this table, we can see that adding temporal correlation in modeling improves the reconstruction accuracy.

Table 1: Distortion between clean speech and model reconstruction from clean input (dB)

CRBM	Training set	Testing set
CRBM_0_HS100	0.3248	0.3292
CRBM_1_HS100	0.1260	0.1288

3.2. Speech restoration from distorted input

Because the RBM and CRBM learns the input feature dependency, when some components of the input feature vector are missed or distorted, the missed or distorted components can be restored by clamping the correct components to the visible output while doing Gibbs sampling on the learned models. Speech has strong correlated spectral structure which encodes information of phonemes, syllables, and words. Incorporating longer temporal window should improve the restoration accuracy. However, in real applications, it needs very large number of parameters in modeling. If the parameters are not trained well because of training algorithm and training data set size, the performance is possibly degraded. We have discussed the effect of input spectral patch size on the clean speech reconstruction (refer to 3.1.1), nevertheless, the effect for distorted speech may be different. Therefore, we carried out experiments to investigate the restoration accuracy based on models trained by using different spectral patch sizes.

In Mel filter band spectrum, the dimensions from 10 to 20 was replaced with random noise. In this condition, all information in those bands are completely lost, traditional speech enhancement algorithms can not be applied (no speech and noise statistical information can be tracked from those frequency bands). Because of the associative function of the RBM and CRBM, we could restore the spectrum based on the trained models. An example of the restoration is shown in Fig. 10. The restoration is done with 30 times Gibbs sampling on the trained CRBM. From this figure, we can see that the restoration based on CRBM_0 and CRBM_1 both work well. For a better understanding of the restoration, we quantify the restoration accuracy by using the definition in Eq. 7. The restoration error is shown in table 2. From this table, we can see that modeling local temporal correlation in the CRBM improves restoration accuracy than the RBM modeling.

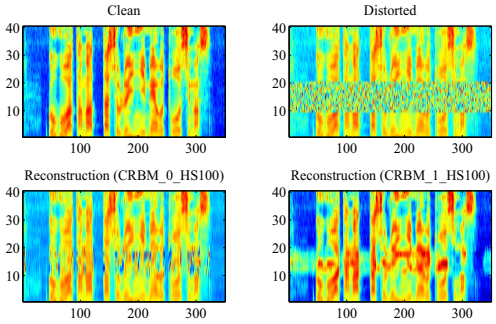

Figure 10: Restoration of the missed or distorted input. Horizontal axis: time frame index, vertical axis: Mel filter band index.

Table 2: Distortion between clean speech and model restoration from distorted input (dB)

PatchSize	1	3	5	7
CRBM_0_HS100	0.62	0.56	0.52	0.55
CRBM_1_HS100	0.58	0.46	0.44	0.46

4. Conclusion and discussions

In this study, we regarded the RBM as an associative memory for speech spectrum modeling. After it was trained, it was used to recall the original speech spectrum when some parts of the input were distorted or missed. In the RBM, there is no consideration of the local temporal correlation between training spectral patches. Since local temporal correlation is an important property of speech, we further propose to model this correlation in a CRBM. Our results showed that, compared with modeling with the RBM, the CRBM not only improved reconstruction accuracy for clean speech, but also improved restoration accuracy from distorted speech.

Several problems need to be further investigated. First of all, stacking many RBMs to be a deep belief network (DBN) has been proved to be more powerful in data modeling and pattern classification than a shallow network. The CRBM has also been made deep for dynamic movement generation [7]. In the future, we will extend this work for speech restoration by stacking many CRBMs. In this sense, the deep CRBM may be used to model the long temporal hierarchical structure of speech. The second, in our restoration experiments, we had supposed that the distorted dimensions in visible input vectors were known. Therefore, we could clamp the CRBM to the known visible features, and using Gibbs sampling to alternatively generate the missed or distorted components. However, in real applications, this should not be assumed as a prior. In the future, we will investigate the restoration without given the exact distorted or missed components.

5. References

[1] Loizou, P. C., Speech Enhancement: Theory and Practice, CRC Press, 2007.

[2] Lu, X., Tsao, Y., Matsuda, S., Hori, C., "Speech Enhancement Based on Deep Denoising Autoencoder," INTERSPEECH, Lyon, France, Aug., 2013.

[3] Hinton, G. E., and Salakhutdinov, R., "Reducing the Dimensionality of Data with Neural Networks," Science, 313: 504-507, 2006.

[4] Bengio, Y., "Learning deep architectures for AI," Foundations and Trends in Machine Learning, 2(1): 1-127, 2009.

[5] Dahl, G., Yu, D., Deng, L., Acero, A., "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, 20 (1): 30-42, 2011.

[6] Sutskever, I., Hinton, G., "Learning Multilevel Distributed Representations for High Dimensional Sequences," AISTATS, 2007.

[7] Taylor, G., Hinton, G. Roweis, S., "Modeling human motion using binary latent variables," NIPS, 1345-1352, 2007.

[8] Lu, X., Matsuda, S., Hori, C., Kashioka, H., "Speech restoration based on deep learning autoencoder with layer-wised learning," INTERSPEECH, Portland, Oregon, Sept., 2012.