INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Jointly Optimizing Activation Coefficients of Convolutive NMF Using DNN for Speech Separation

Hao Li[1], *Shuai Nie*[2], *Xueliang Zhang*[1], *Hui Zhang*[1]

[1]College of Computer Science, Inner Mongolia University
[2]National Laboratory of Patten Recognition, Institute of Automation, Chinese Academy of Sciences
lihao.0214@163.com shuai.nie@nlpr.ia.ac.cn cszxl@imu.edu.cn alzhu.san@163.com

Abstract

Convolutive non-negative matrix factorization (CNMF) and deep neural networks (DNN) are two efficient methods for monaural speech separation. Conventional DNN focuses on building the non-linear relationship between mixture and target speech. However, it ignores the prominent structure of the target speech. Conventional CNMF model concentrates on capturing prominent harmonic structures and temporal continuities of speech but it ignores the non-linear relationship between the mixture and target. Taking these two aspects into consideration at the same time may result in better performance. In this paper, we propose a joint optimization of DNN models with an extra CNMF layer for speech separation task. We also utilize an extra masking layer on the proposed model to constrain the speech reconstruction. Moreover, a discriminative training criterion is proposed to further enhance the performance of the separation. Experimental results show that the proposed model has significant improvement in PESQ, SAR, SIR and SDR compared with conventional methods.

Index Terms: speech separation, Convolutive non-negative matrix factorization (CNMF), Deep neural networks (DNN)

1. Introduction

Speech separation aims to segregate the target speech from the mixture, which is important for many realistic applications, such as speech communication and automatic speech recognition (ASR) [1, 3, 15]. Although it has been studied for years, current separation systems are still far behind human capability. Finding a good speech separation system remains an unsolved problem, especially in low signal-to-noise (SNR) and non-stationary noise conditions.

Non-negative matrix factorization (NMF) is a popular algorithm used for monaural speech separation, which can capture the prominent structure pattern of the speech. The basic principle of NMF-based method is to represent the features of the clean speech and noise via sets of basis spectra matrices and their activation coefficients. Mixture signal is then analyzed using the concatenated sets of basis spectra matrix. The clean speech and the noise can be reconstructed using its corresponding basis matrix and activation coefficients. NMF captures the frequency structure of signals. However, previous studies, such as in [6], show that the speech has prominent structure on both frequency and temporal axes. Therefore, the speech separation would be benefit from considering the temporal pattern. In [12], a more efficient approach, convolutive non-negative matrix factorization (CNMF), is proposed which involves the sharing of decompositions among a set of bases with time shift and has been shown to perform well in speech separation task. CNMF can capture the complex temporal and frequency patterns of speech, which is an extension of NMF. The basis spectra obtained by CNMF using only one speech utterance are shown in Figure 1. We can see that the basis spectra reflect the time-frequency structure of the speech very well. However, the CNMF is a linear model that limits its capability. Another problem is that CNMF-based model relies on an iterative method in separating stage which makes it inefficient.

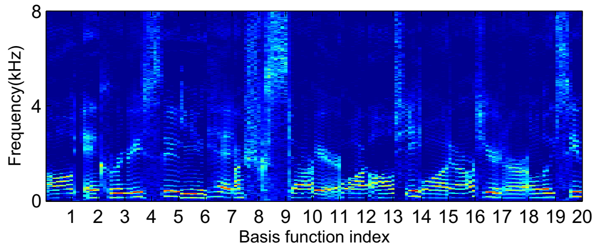

Figure 1: *The basis spectra of speech for a single female speaker (SA1) taken from the TIMIT speech database*

Recently, the deep neural network (DNN) has attracted many researchers' attention. The speech separation methods based on DNN can achieve good performance due to the powerful capabilities of DNN on building the non-linear relationships between the mixture and the target. In general, DNN-based model predicts a mask or the magnitude spectrogram of interest [8, 16, 17] directly, which has no constraint on temporal and frequency structures of the target speech. Therefore, the DNN-based speech separation system would benefit from integrating the target speech patterns to DNN.

In this paper, we propose a novel DNN integrating CNMF to predict magnitude spectrograms of speech and noise simultaneously. First, we use CNMF to obtain the basis spectra from clean speech and noise. Second, we combine activation coefficients computed by DNN with basis spectra to reconstruct the magnitude spectrograms of the target speech and the noise. Finally, the errors between the target speech and estimation are used to update the weights of DNN. In addition, to enforce a reconstruction constraint, we add an extra masking layer on the model. A discriminative training objective is also explored.

The rest of this paper is organized as follows. In Section 2, we introduce the problem formulation of the separation task by CNMF algorithm. In Section 3, we present the proposed model. The method of optimizing the proposed model is also put

This research was partly supported by the China National Nature Science Foundation (No.61365006, No.61273267, No.61403370, No.61573357, No.61503382 and No.91120303).

http://dx.doi.org/10.21437/Interspeech.2016-120

forward. Experiments and evaluation are provided in Section 4. Finally, we summarize our work in Section 5.

2. Problem formulation

Given a noisy signal, we assume that the magnitude spectrograms of speech and noise are additive. The magnitude spectrogram of mixture can be approximated as:

$$\mathbf{X}(t,f) \approx \mathbf{Y}_s(t,f) + \mathbf{Y}_n(t,f) \tag{1}$$

where $\mathbf{X}$, $\mathbf{Y}_s$ and $\mathbf{Y}_n$ are the magnitude spectrum of the mixture, clean speech and noise, respectively. t presents the frame index. f presents the frequency index. Previous study has shown that this approximate yields suitable results [10].

Applying CNMF to $\mathbf{Y}_s$ and $\mathbf{Y}_n$, respectively, we can obtain the approximate factorization of $\mathbf{Y}_s$ and $\mathbf{Y}_n$ via sets of basis spectra and corresponding activation coefficients as follows:

$$\begin{aligned} \mathbf{Y}_s &\approx \sum_{t=0}^{T-1} \mathbf{W}_s(t) \cdot \overset{t\rightarrow}{\mathbf{H}_s} \\ \mathbf{Y}_n &\approx \sum_{t=0}^{T-1} \mathbf{W}_n(t) \cdot \overset{t\rightarrow}{\mathbf{H}_n} \end{aligned} \tag{2}$$

where $\mathbf{W}_s(t) \in \mathbb{R}_+^{F\times L_s}$ and $\mathbf{W}_n(t) \in \mathbb{R}_+^{F\times L_n}$ are the basis spectra matrices of speech and noise, respectively. L_s and L_n are the numbers of basis vectors to represent speech and noise. $\mathbf{H}_s \in \mathbb{R}_+^{L_s\times N}$ and $\mathbf{H}_n \in \mathbb{R}_+^{L_n\times N}$ are the activation coefficients. T denotes the temporal extent in frames of the CNMF bases. The $\overset{i\rightarrow}{(\cdot)}$ denotes a column-shift operator that moves the columns of its argument by i spots to the right. As each column is shifted off to the right, the leftmost column are filled by zero. Conversely, the $\overset{i\leftarrow}{(\cdot)}$ operators shifts columns off to the left, with zero filling on the right. Here is an example.

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} & \overset{0\rightarrow}{\mathbf{A}} &= \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} \\ \overset{1\rightarrow}{\mathbf{A}} &= \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{bmatrix} & \overset{2\leftarrow}{\mathbf{A}} &= \begin{bmatrix} 3 & 4 & 0 & 0 \\ 7 & 8 & 0 & 0 \end{bmatrix} \end{aligned} \tag{3}$$

In general, the basis spectra matrices $\mathbf{W}_s$ and $\mathbf{W}_n$ are learned using clean speech and noise in advance. Given the $\mathbf{W}_s$ and $\mathbf{W}_n$, the magnitude spectrogram of the noisy $\mathbf{X}$ can be rewritten as (4), where $\hat{\mathbf{H}}_s$ and $\hat{\mathbf{H}}_n$ are computed iteratively by CNMF using minimum square error (MSE) criterion. More details could be found in [12].

$$\mathbf{X} \approx \hat{\mathbf{Y}}_s + \hat{\mathbf{Y}}_n \approx \sum_{t=0}^{T-1} \left(\begin{bmatrix} \mathbf{W}_s(t)\ \mathbf{W}_n(t) \end{bmatrix} \begin{bmatrix} \overset{t\rightarrow}{\hat{\mathbf{H}}_s} \\ \overset{t\rightarrow}{\hat{\mathbf{H}}_n} \end{bmatrix} \right) \tag{4}$$

Moreover, we also define the soft time-frequency mask $\mathbf{M}$ to restrict the sum of $\mathbf{Y}_s$ and $\mathbf{Y}_n$ to $\mathbf{X}$, which is shown as (5).

$$\begin{aligned} \mathbf{M}_s &= \frac{\hat{\mathbf{Y}}_s}{\hat{\mathbf{Y}}_s + \hat{\mathbf{Y}}_n} \\ \mathbf{M}_n &= \frac{\hat{\mathbf{Y}}_n}{\hat{\mathbf{Y}}_s + \hat{\mathbf{Y}}_n} \end{aligned} \tag{5}$$

where, $\mathbf{M}_s$ and $\mathbf{M}_n$ are the soft masks of the target speech and the noise, respectively. Then, the estimated separation spectra $\widetilde{Y}_s$ and $\widetilde{Y}_n$ can be computed as follows:

$$\begin{aligned} \widetilde{\mathbf{Y}}_s &= \mathbf{M}_s \otimes \mathbf{X} \\ &= \frac{\sum_{t=0}^{T-1} \mathbf{W}_s(t) \cdot \overset{t\rightarrow}{\hat{\mathbf{H}}_s}}{\sum_{t=0}^{T-1} \left(\begin{bmatrix} \mathbf{W}_s(t)\ \mathbf{W}_n(t) \end{bmatrix} \begin{bmatrix} \overset{t\rightarrow}{\hat{\mathbf{H}}_s} \\ \overset{t\rightarrow}{\hat{\mathbf{H}}_n} \end{bmatrix} \right)} \otimes \mathbf{X} \\ \widetilde{\mathbf{Y}}_n &= \mathbf{M}_n \otimes \mathbf{X} \\ &= \frac{\sum_{t=0}^{T-1} \mathbf{W}_n(t) \cdot \overset{t\rightarrow}{\hat{\mathbf{H}}_n}}{\sum_{t=0}^{T-1} \left(\begin{bmatrix} \mathbf{W}_s(t)\ \mathbf{W}_n(t) \end{bmatrix} \begin{bmatrix} \overset{t\rightarrow}{\hat{\mathbf{H}}_s} \\ \overset{t\rightarrow}{\hat{\mathbf{H}}_n} \end{bmatrix} \right)} \otimes \mathbf{X} \end{aligned} \tag{6}$$

where the division is performed element-wise, and $\otimes$ denotes an element-wise multiplication. Finally, with the separated spectra of the speech and the noise, we can convert the estimation from frequency domain to time domain using the noisy phase and inverse short time Fourier transform (STFT).

3. Proposed method

From the above, we can see that the key to solve the speech separation problem are to obtain the basis spectra $\mathbf{W}_s$ and $\mathbf{W}_n$ and estimate the activation coefficients $\hat{\mathbf{H}}_s$ and $\hat{\mathbf{H}}_n$. We use CNMF algorithm to generate $\mathbf{W}_s$ and $\mathbf{W}_n$ on a set of clean speeches and different kinds of noises. With fixed $\mathbf{W}_s$ and $\mathbf{W}_n$, we use a DNN model to learn the $\hat{\mathbf{H}}_s$ and $\hat{\mathbf{H}}_n$ from the mixture. Then, we can obtain the estimation use (6). Finally, we use the error between the estimation and the target to update the DNN.

3.1. Model Architecture

To estimate the $\widetilde{\mathbf{Y}}_s$ and $\widetilde{\mathbf{Y}}_n$ from mixture $\mathbf{X}$, we integrate CNMF into DNN model. The architecture of the model is shown in Figure 2. The DNN model consists of one input layer, several hidden layers, one output layer, and two extra layers: CNMF layer and masking layer. Extra layers are deterministic layers without any update during the DNN training.

3.2. Discriminative Training

When $\widetilde{\mathbf{Y}}_s$ and $\widetilde{\mathbf{Y}}_n$, the estimates of $\mathbf{Y}_s$ and $\mathbf{Y}_n$, are obtained from the proposed hybrid model, we can optimize the DNN parameters using the discriminative objective function as in (7).

$$\begin{aligned} J = &\frac{1}{2} \left(\left\| \mathbf{Y}_s - \widetilde{\mathbf{Y}}_s \right\|_2^2 + \left\| \mathbf{Y}_s - \widetilde{\mathbf{Y}}_s \right\|_2^2 \right) \\ &- \frac{\lambda}{2} \left(\left\| \mathbf{Y}_s - \widetilde{\mathbf{Y}}_n \right\|_2^2 + \left\| \mathbf{Y}_n - \widetilde{\mathbf{Y}}_s \right\|_2^2 \right) \end{aligned} \tag{7}$$

where $\|\cdot\|_2$ is the ℓ_2 norm between the two matrices. λ is a constant, chosen by the performance of the experiment. In [5], the author has shown that the discriminative objective function can obtain a better performance compared with conventional MSE based objective function.

3.3. Optimization

The optimization of the proposed model includes two stages: forward propagation stage and back propagation stage. From the architecture of the proposed model, we can see that the

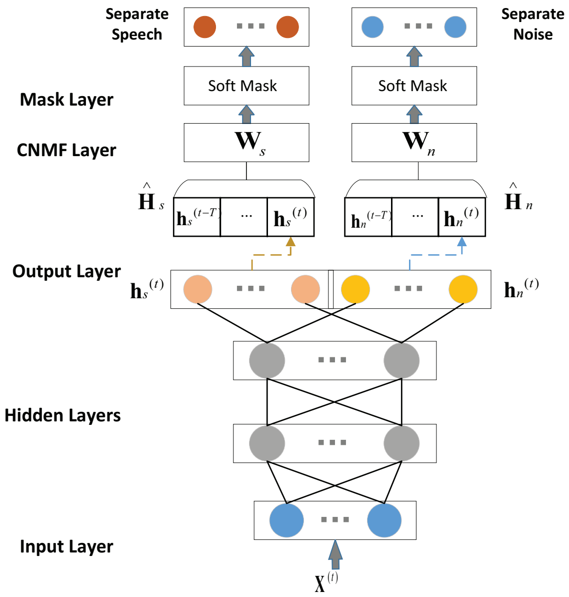

Figure 2: *Proposed model architecture*

weight share strategy should be employed when updating the weights of the DNN.

3.3.1. *Forward propagation*

Suppose that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_m]$ is the input sequences, where $\mathbf{x}_m$ is the *m-th* frame of input. The activations of hidden and output layers can be computed recursively using (8) as follows:

$$\mathbf{Z}_{l+1} = \mathbf{W}_l \times \mathbf{A}_l + \mathbf{b}_l;\ \mathbf{A}_{l+1} = f(\mathbf{Z}_{l+1}) \tag{8}$$

where, $\mathbf{W}_l$ and $\mathbf{b}_l$ are the weights and bias. $\mathbf{A}_l$ is the activations of the layer l. $\mathbf{A}_0$ is the input $\mathbf{X}$. $f(\cdot)$ is the activation function.

After getting the $\hat{\mathbf{H}}_s$ and $\hat{\mathbf{H}}_n$ from the output layer of DNN, we can use $\mathbf{W}_s$ and $\mathbf{W}_n$ to compute the $\hat{\mathbf{Y}}_s$ and $\hat{\mathbf{Y}}_n$ by (2), which are the outputs of CNMF layer in Figure 2. The final outputs of the proposed model, $\widetilde{\mathbf{Y}}_s$ and $\widetilde{\mathbf{Y}}_n$, are computed according to (6).

3.3.2. *Back propagation*

The weights of DNN are updated by computing the gradients of the objective function (7) using chain rules:

$$\nabla \mathbf{W}_l = \frac{\partial J}{\partial \mathbf{Z}_{l+1}} \frac{\partial \mathbf{Z}_{l+1}}{\partial \mathbf{W}_l} = \frac{\partial J}{\partial \mathbf{Z}_{l+1}} (\mathbf{A}_l)^{\mathbf{T}} \tag{9}$$

To simplify the notations, we introduce a variable δ defined as $\delta_l = \frac{\partial J}{\partial z_l}$. First, for the output layer $(l = n_l)$, we have:

$$\delta_{n_l} = \frac{\partial J}{\partial \hat{\mathbf{H}}_{n_l}} \otimes \frac{\hat{\mathbf{H}}_{n_l}}{\partial \mathbf{Z}_{n_l}} = \left[\frac{\partial J}{\partial \hat{\mathbf{H}}_s}, \frac{\partial J}{\partial \hat{\mathbf{H}}_n} \right] \otimes f'\left(\mathbf{Z}_{n_l}\right) \tag{10}$$

where,

$$\begin{aligned} \frac{\partial J}{\partial \hat{\mathbf{H}}_s} &= \sum_{t=1}^{T-1} \left(\mathbf{W}_s\left(t\right)\right)^{\mathbf{T}} \overset{t\leftarrow}{\mathbf{B}_s} \\ \frac{\partial J}{\partial \hat{\mathbf{H}}_n} &= \sum_{t=1}^{T-1} \left(\mathbf{W}_n\left(t\right)\right)^{\mathbf{T}} \overset{t\leftarrow}{\mathbf{B}_n} \end{aligned} \tag{11}$$

and,

$$\begin{aligned} \mathbf{B}_s &= [-(\mathbf{Y}_s - \widetilde{\mathbf{Y}}_s) + \lambda(\mathbf{Y}_n - \widetilde{\mathbf{Y}}_s) \\ &\quad + (\mathbf{Y}_n - \widetilde{\mathbf{Y}}_n) - \lambda(\mathbf{Y}_s - \widetilde{\mathbf{Y}}_n)] \otimes \frac{\widetilde{\mathbf{Y}}_n}{\widetilde{\mathbf{Y}}_s + \widetilde{\mathbf{Y}}_n} \\ \mathbf{B}_n &= [(\mathbf{Y}_s - \widetilde{\mathbf{Y}}_s) - \lambda(\mathbf{Y}_n - \widetilde{\mathbf{Y}}_s) \\ &\quad - (\mathbf{Y}_n - \widetilde{\mathbf{Y}}_n) + \lambda(\mathbf{Y}_s - \widetilde{\mathbf{Y}}_n)] \otimes \frac{\widetilde{\mathbf{Y}}_s}{\widetilde{\mathbf{Y}}_s + \widetilde{\mathbf{Y}}_n} \end{aligned} \tag{12}$$

Second, for the *l-th* layer $(l = n_l - 1, n_l - 2, \ldots, 1)$, we have:

$$\delta_l = \left(\left(\mathbf{W}_l\right)^T \delta_{l+1} \right) \otimes f'(\mathbf{Z}_l) \tag{13}$$

Then, we can compute the partial derivatives of the discriminative objective function with respects to the DNN weights by (14)

$$\nabla \mathbf{W} = \delta_l \left(\mathbf{A}_{l-1}\right)^T \tag{14}$$

Finally, with partial derivatives, we can update the weights of the DNN by Limited-memory Broyden-Fletcher-Goldfarb-Shanno Algorithm (LBFGS) [7].

4. Experiments

4.1. Experiment setup

In order to evaluate the performance of the proposed separation model, the TIMIT [2] corpus and the NOISEX-92 [13] corpus are used in experiments. The TIMIT corpus are used as the clean database and the NOISEX-92 corpus are used as interference, respectively. The TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English. The NOISEX-92 contains 15 common types of noise in real-world environment. Each type noise is about 4 minutes long. The sampling rate is 16kHz for all utterances.

A fixed 32-ms frame size was used with 50% overlap between frames. The discrete Fourier transform (DFT) is applied on each frame. And the length of the DFT is 512.

Table 1: *The average value of SDR in different* λ

λ	0.005	0.007	0.01	0.03	0.05	0.1
SDR	9.6482	9.6609	9.6702	**9.7364**	9.6516	9.5924

In order to train the proposed model, we chose 50 speakers from the TIMIT randomly and select 2 utterances randomly for each speaker. 100 selected utterances are randomly mixed with 10 types of noise to generate 3000 mixtures as training set. The SNR of the mixture are distributed from -7dB to 7dB continuously. At test stage, we chose 200 clean utterances, unseen in training set, from the TIMIT corpus randomly. Then, we use clean speech mixed with all 15 types of noises randomly to generate 400 test utterances. Among the 15 type noises, 5 types are unseen in training set, which are considered as the unmatched noises to evaluate the robustness of the proposed model. The SNR of the test set are distributed in $\{$*-10,-7,-5,-2,0,2,5,7,10*$\}$dB randomly. Meanwhile, we also generate 500 validation utterances. In validation set, we use all 15 types noises. And the clean speech used differ from those used in training set.

For CNMF, we use all utterances of the TIMIT to train the basis spectra of clean speech. We train the basis spectra of the noise using 9 types of noises which is same as the noises added in DNN training set.

We evaluate the performance using Source to Interference Ratio(SIR), Source to Distortion Ratio (SDR), Source to Artifacts Ratio (SAR) [14] and Perceptual Evaluation of Speech Quality (PESQ) [11]. SAR, SIR, and SDR reflect the artifacts introduced by separation process, the suppression of interference and the overall performance, respectively. PESQ reflects the quality of the objective speech. Higher value means better performance for all evaluation criteria.

Table 2: *Speech separation performance with various source algorithms in matched noise*

SNR	DNN			CNMF			DNN-CNMF		
(dB)	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
-10	4.03	7.57	3.1	-1.32	0.17	**6.02**	**4.6**	**8.46**	3.04
-7	7.15	12.16	6.71	2.37	4.34	6.06	**8.23**	**14.62**	**7.68**
-5	8.04	13.15	7.81	3.82	5.81	6.8	**9.44**	**16.45**	**9.11**
-2	8.37	12.71	8.64	4.66	6.37	8.01	**9.37**	**15.18**	**9.46**
0	8.78	13.06	9.19	5.55	7.19	8.76	**9.83**	**15.3**	**10.14**
2	10.39	15.57	10.81	7.06	8.9	9.72	**11.92**	**18.32**	**12.4**
5	11.05	16.23	11.46	7.94	9.94	10.06	**12.91**	**18.84**	**13.68**
7	12.49	18.56	12.91	8.51	10.61	10.49	**15.01**	**21.63**	**15.84**
10	13.67	20.64	13.97	9.52	12.01	11.03	**17.56**	**24.03**	**18.59**

Table 3: *Speech separation performance with various source algorithms in unmatched noise*

SNR	DNN			CNMF			DNN-CNMF		
(dB)	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
-10	**2.01**	6.13	1.95	-1.49	0.85	**3.35**	1.59	**7.01**	1.65
-7	2.52	5.57	4.09	-0.18	1.41	**5.65**	**2.76**	**6.63**	3.8
-5	3.73	6.76	5.11	1.04	2.46	**6.41**	**4.05**	**7.92**	5.2
-2	5.93	9.23	7.28	3.58	5.19	**7.7**	**6.12**	**9.98**	7.58
0	8.4	12.6	9.17	5.26	6.97	8.42	**8.85**	**13.7**	**9.91**
2	8.09	11.22	9.42	5.69	7.19	9.13	**8.69**	**12.4**	**10.21**
5	10.15	14.48	11.11	7.44	9.28	9.97	**11.12**	**15.66**	**12.56**
7	11.23	15.64	12.23	7.8	9.61	10.2	**12.77**	**16.84**	**14.64**
10	12.74	18.53	13.34	8.84	11.08	10.57	**14.98**	**19.18**	**16.9**

4.2. Baseline Model and Parameter Selection

We compare the performance of the proposed model (denoted as 'DNN-CNMF') with conventional DNN [17] and CNMF [9] models. In all DNN models, we use a window (5 frames) of combined magnitude spectrograms as input features. All DNN models have two hidden layers of 1000 units. The conventional DNN model predicts the current frame of magnitude spectrograms of clean speech and noise directly. The weights of all DNN models are initialized randomly. All models iterate 500 times by standard back propagation algorithm. In CNMF training stage, we learn basis spectra matrices of speech and noise from training set using a convolutional of 8 frames. Besides, to capture the basis spectra sufficiently, we set 256 as the numbers of bases vectors, and iterate 200 times. Finally, we should note that, the output of the DNN-CNMF is used as activation coefficient of the CNMF which is nonnegative. Therefore, ReLu activation function [4] $(f(x) = max(0, x))$ is used for the output layer of all DNN models.

4.3. Experimental results

Before analysing the results of the experiments, we should explore the value of λ. We use the average result of SDR on test set to select λ. As showed in Table 1, we fixed $\lambda = 0.03$ in following experiments.

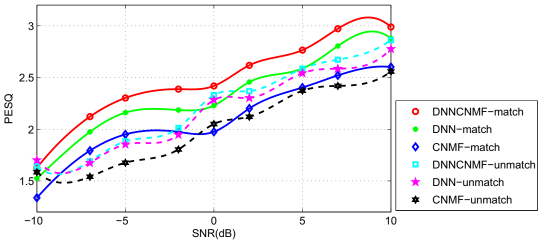

Figure 3: *PESQ with various source algorithms in matched noise and unmatched noise*

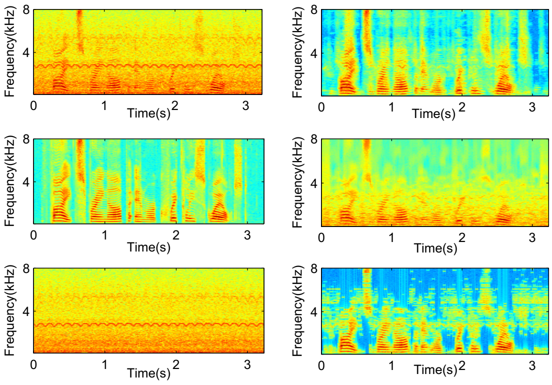

Figure 4: *Top left: The spectrogram of the mixture speech; Centre left: Clean speech spectrogram; Bottom left: Clean noise spectrogram; Top right: Speech separation using our method; Centre right: Speech separation using CNMF; Bottom right: Speech separation using DNN*

Table 2, Table 3 and Figure 3 show the results of different speech separation methods (DNN-CNMF, DNN and CNMF) with matched noise and unmatched noise conditions, respectively. We can observe that the DNN-CNMF model achieves better results on most evaluation criterions. We think it caused by two possible reasons: First, the CNMF ignores the non-linear relationship between the mixture and the target. Second, although both the DNN-CNMF and conventional DNN model can predict magnitude spectrograms, the DNN-CNMF model can reconstruct the target speech well. Due to the information of the temporal and frequency patterns of the target is employed in the model.

Finally, an example of reconstructed the clean spectrogram is shown in Figure 4. We can find that the proposed model is much better than the CNMF model and DNN model. The proposed model can suppress more interferences than conventional CNMF in all frequency bands. Meanwhile, the proposed model causes less speech distortion than DNN model in middle and high frequency bands.

5. Conclusion

In this paper, we propose a novel DNN-CNMF model for monaural speech separation. Through a series of experiments, we have proved that the performance of the proposed model outperforms the conventional CNMF-based model and DNN model. For future work, we will explore the feasibility of DNN-CNMF model in reverberation environment.

6. References

[1] C. Demir, M. Saraclar, and A. T. Cemgil, "Single-channel speech-music separation for robust asr with mixture models," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 4, pp. 725–736, 2013.

[2] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus cdrom," 1993.

[3] K. Gibak, L. Yang, H. Yi, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners." *Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[4] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks." in *AISTATS*, ser. JMLR Proceedings, G. J. Gordon, D. B. Dunson, and M. Dudłk, Eds., vol. 15. JMLR.org, 2011, pp. 315–323.

[5] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1562–1566.

[6] S. Liang, W. Liu, and W. Jiang, "A new bayesian method incorporating with local correlation for ibm estimation." *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, no. 3, pp. 476–487, 2013.

[7] D. C. LIU and J. NOCEDAL, "On the limited memory BFGS method for large scale optimization," *Math. Programming*, vol. 45, no. 3, (Ser. B), pp. 503–528, 1989.

[8] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition." in *ICASSP*. IEEE, 2013, pp. 7092–7096.

[9] P. D. O'Grady and B. A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint." *Neurocomputing*, vol. 72, no. 1-3, pp. 88–101, 2008.

[10] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation." *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 6, pp. 1766–1776, 2007.

[11] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs." in *ICASSP*. IEEE, 2001, pp. 749–752.

[12] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio Speech Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[13] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems." *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[14] E. Vincent, R. Gribonval, and C. Fvotte, "Performance measurement in blind audio source separation." *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[15] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust asr." in *ICASSP*. IEEE, 2012, pp. 4085–4088.

[16] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *Audio Speech and Language Processing IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 1849–1858, 2014.

[17] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.