

EXPLOITING SPECTRO-TEMPORAL STRUCTURES USING NMF FOR DNN-BASED SUPERVISED SPEECH SEPARATION

Shuai Nie¹, Shan Liang¹, Hao Li², XueLiang Zhang², ZhanLei Yang¹, WenJu Liu¹, LiKe Dong³

¹National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

²College of Computer Science, Inner Mongolia University

³Electric Power Research Institute of ShanXi Electric Power Company, China State Grid Corp

{shuai.nie, sliang, zhanlei.yang, lwj}@nlpr.ia.ac.cn {cszx1, cslh}@imu.edu.cn

ABSTRACT

The targets of speech separation, whether ideal masks or magnitude spectrograms of interest, have prominent spectro-temporal structures. These characteristics are very worthy to be exploited for speech separation, however, they are usually ignored in previous works. In this paper, we use nonnegative matrix factorization (NMF) to exploit the spectro-temporal structures of magnitude spectrograms. With nonnegative constraints, NMF can capture the basis spectra patterns of speech and noise. Then the learned basis spectra are integrated into a deep neural network (DNN) to reconstruct the magnitude spectrograms of speech and noise with their nonnegative linear combination. Using the reconstructed spectrograms, we further explore a discriminative training objective and a joint optimization framework for the proposed model. Systematic experiments show that the proposed model is competitive with the previous methods in monaural speech separation tasks.

Index Terms— Speech Separation, Deep Neural Network, Non-negative Matrix Factorization, Spectro-Temporal Structures

1. INTRODUCTION

Segregating the interested speech from the mixture has many important realistic applications. A good speech separation system can significantly improve the speech intelligibility and the performance of automatic speech recognition [1–4]. However, in real-world environments, speech separation is still a challenging task, especially when noise is non-stationary and only one microphone is available.

Speech separation can be formulated as a supervised learning problem [5–8]. A typical supervised speech separation system usually learns a mapping function from noisy features to certain ideal masks or magnitude spectrograms of interest through a supervised learning algorithm [5]. Recently, supervised speech separation has been extensively studied and shown to be substantially promising for the challenging acoustic conditions [9–11].

Due to speech production mechanisms, speech presents prominent harmonic structures and temporal continuities, which results that the targets of speech separation, whether ideal masks or magnitude spectrograms of interest, have strong spectro-temporal structures [12]. Explicitly exploiting these characteristics will probably

improve the separation performance. However, previous methods usually ignore these structure patterns and directly predict ideal masks or magnitude spectrograms of interest using a deep neural network (DNN) [5, 6].

Nonnegative matrix factorization (NMF) is a well-known technique that can discover parts-based representations underlying non-negative data [13]. When performed on the targets of speech separation, such as the magnitude spectrograms of interest, NMF can capture the spectra patterns of output targets [14–16]. In [12], Wang uses NMF to preserve the spectro-temporal structures of the square-root ideal ratio mask, and DNN is used to predict the activation coefficients of mask-level spectro-temporal bases. Although the output structures are captured by NMF, DNN optimizes the intermediate separation objective rather than the actual separation objective, and the learned basis spectra are completely independent of training for DNN, which probably lead to the separation that is more sensitive to the estimation errors of DNN.

In this paper, we propose to use NMF to exploit the spectro-temporal structures of speech and noise and incorporate the basis spectra learned by NMF into DNN-based speech separation. In advance, we perform NMF on the magnitude spectrograms of the clean speech and noise to obtain their basis spectra. Then the learned basis spectra of speech and noise are integrated into the original output layer of DNN. And DNN is trained to simultaneously reconstruct the magnitude spectrograms of speech and noise with the nonnegative linear combination of the basis spectra. Its original outputs are used as the activation coefficients of the bases rather than to compute the error metric. To enforce that the sum of the reconstruction results is equal to the original mixture, we use a Wiener-type filtering to obtain the final estimates for each source. Using the final estimation results, we further explore a discriminative training objective and a joint optimization framework for the proposed model. In fact, the basis spectra and the Wiener-type filtering can be viewed as extra layers that are added to the original output of DNN, but they are deterministic and has no connective weights to be optimized.

2. PROBLEM FORMULATION

The task of speech separation is to obtain an estimate $\hat{s}(k)$ of target speech $s(k)$ from a mixture signal $x(k)$ containing additive noises $n(k)$. For this problem, the short time Fourier transform (STFT) is the commonest technique. We define $X(t, f)$, $Y_s(t, f)$ and $Y_n(t, f)$ as the STFT coefficients of $x(k)$, $s(k)$ and $n(k)$, respectively, where

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

t represents the frame index and f is the frequency-index. Due to the sparse nature of speech, the magnitude spectrum of the mixture speech can be approximated as follows [17, 18]:

$$|X(t, f)| \approx |Y_s(t, f)| + |Y_n(t, f)| \quad (1)$$

where $|\cdot|$ denotes the absolute value operator in the complex domain. For simplicity, we rewrite (1) in vector form as follows:

$$\mathbf{x} \approx \mathbf{y}_s + \mathbf{y}_n \quad (2)$$

where $\mathbf{x} \in \mathbb{R}_+^{F \times 1}$, $\mathbf{y}_s \in \mathbb{R}_+^{F \times 1}$ and $\mathbf{y}_n \in \mathbb{R}_+^{F \times 1}$ denote the magnitude spectrums of the mixture, the speech and the noise at the time frame t , respectively. For simplicity, unless mentioned explicitly, the time frame index t is omitted. F is the number of frequency bins.

As we perform NMF on \mathbf{y}_s and \mathbf{y}_n , we can obtain the approximate factorization of \mathbf{y}_s and \mathbf{y}_n via sets of basis vectors and their activation coefficients as follows:

$$\mathbf{y}_s \approx \mathbf{B}_s \mathbf{a}_s; \quad \mathbf{y}_n \approx \mathbf{B}_n \mathbf{a}_n \quad (3)$$

where $\mathbf{B}_s \in \mathbb{R}_+^{F \times L_s}$ and $\mathbf{B}_n \in \mathbb{R}_+^{F \times L_n}$ are the basis spectra of speech and noise, respectively. L_s and L_n are the numbers of basis vectors of speech and noise, respectively. $\mathbf{a}_s \in \mathbb{R}_+^{L_s \times 1}$ and $\mathbf{a}_n \in \mathbb{R}_+^{L_n \times 1}$ are the time-varying activation levels of the corresponding basis vectors. At this study, the basis matrices \mathbf{B}_s and \mathbf{B}_n are learned using the appropriate training data in advance. With \mathbf{B}_s and \mathbf{B}_n held fixed, the magnitude spectrum \mathbf{x} of the mixture signal can be approximated as follows:

$$\mathbf{x} \approx \hat{\mathbf{y}}_s + \hat{\mathbf{y}}_n \approx \mathbf{B}_s \hat{\mathbf{a}}_s + \mathbf{B}_n \hat{\mathbf{a}}_n \quad (4)$$

where $\hat{\mathbf{a}}_s$ and $\hat{\mathbf{a}}_n$ are the unknown activation coefficients and at run time need to be estimated using the observed mixture signal. Then a Wiener-type filtering can be used to reconstruct the magnitude spectrums of the speech and the noise while ensuring that the estimates of speech and noise sum to the mixture [8]:

$$\tilde{\mathbf{y}}_s = \frac{\mathbf{B}_s \hat{\mathbf{a}}_s}{\mathbf{B}_s \hat{\mathbf{a}}_s + \mathbf{B}_n \hat{\mathbf{a}}_n} \otimes \mathbf{x}; \quad \tilde{\mathbf{y}}_n = \frac{\mathbf{B}_n \hat{\mathbf{a}}_n}{\mathbf{B}_s \hat{\mathbf{a}}_s + \mathbf{B}_n \hat{\mathbf{a}}_n} \otimes \mathbf{x} \quad (5)$$

where the division is performed element-wise, and \otimes denotes an element-wise multiplication. Finally, the speech $\hat{s}(k)$ and the noise $\hat{n}(k)$ are obtained using the noisy phase and the inverse STFT.

3. PROPOSED METHOD

From the above problem formulation, we can see that the fundamental issues of speech separation are to obtain the basis spectra \mathbf{B}_s and \mathbf{B}_n of speech and noise, and estimate the corresponding activation coefficients $\hat{\mathbf{a}}_s$ and $\hat{\mathbf{a}}_n$. \mathbf{B}_s and \mathbf{B}_n can be obtained by performing NMF on the magnitude spectrograms of speech and noise in training phase. With the bases held fixed, we use a DNN to learn the corresponding activation coefficients from the mixture input. Finally, the resulting output from Wiener-type filtering is used to compute the error metric for optimizing the network weights.

3.1. Model Architecture

With the bases \mathbf{B}_s and \mathbf{B}_n held fixed, we construct a DNN to estimate the magnitude spectrograms $\hat{\mathbf{y}}_s$ and $\hat{\mathbf{y}}_n$ of speech and noise in the spaces spanned by \mathbf{B}_s and \mathbf{B}_n , respectively, and then the outputs

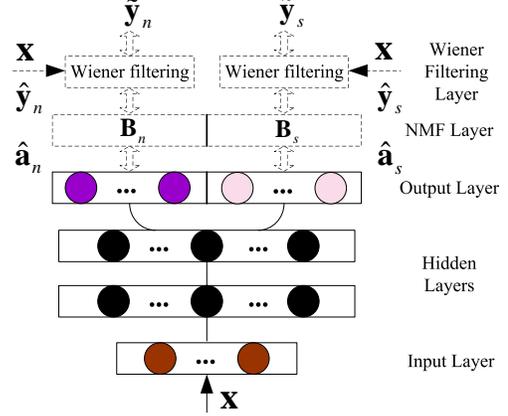


Fig. 1. The architecture of the proposed model.

follow a Wiener-type filtering operator to obtain the final estimates $\tilde{\mathbf{y}}_s$ and $\tilde{\mathbf{y}}_n$, as shown in Fig. 1. Different from the standard DNN, the original output layer of the proposed DNN is followed by a NMF layer and a Wiener-type filtering layer. As extra layers, they are deterministic and has no connective weights to be optimized, but their resulting outputs are used to compute the error metric for optimizing the network weights.

3.2. Training Objectives

Commonly, modeling all sources in a model can improve separation performance due to the complementarity between different sources in the mixture [7]. In this paper, we adapt the architecture of DNN to simultaneously predict the magnitude spectrograms of speech and noise, as shown in Fig. 1. In order to suppress more noise and preserve more speech components, we further explore a discriminative objective for the proposed network. Given the output predictions $\tilde{\mathbf{y}}_s$ and $\tilde{\mathbf{y}}_n$, the discriminative objective not only increases the similarity between the prediction and its target, but also decreases the similarity between the prediction and the targets of other sources, as shown in Eq. (6).

$$J = \frac{1}{2} (\|\mathbf{y}_s - \tilde{\mathbf{y}}_s\|_2^2 + \|\mathbf{y}_n - \tilde{\mathbf{y}}_n\|_2^2) - \frac{\lambda}{2} (\|\mathbf{y}_s - \tilde{\mathbf{y}}_n\|_2^2 + \|\mathbf{y}_n - \tilde{\mathbf{y}}_s\|_2^2) \quad (6)$$

where λ specifies the relative importance of the term and can be experimentally chosen. Introducing Eq. (5) into Eq. (6), we obtain the final objective function as follows:

$$J = \frac{1}{2} \left(\left\| \mathbf{y}_s - \frac{\mathbf{B}_s \hat{\mathbf{a}}_s}{\mathbf{B}_s \hat{\mathbf{a}}_s + \mathbf{B}_n \hat{\mathbf{a}}_n} \otimes \mathbf{x} \right\|_2^2 + \left\| \mathbf{y}_n - \frac{\mathbf{B}_n \hat{\mathbf{a}}_n}{\mathbf{B}_s \hat{\mathbf{a}}_s + \mathbf{B}_n \hat{\mathbf{a}}_n} \otimes \mathbf{x} \right\|_2^2 \right) - \frac{\lambda}{2} \left(\left\| \mathbf{y}_s - \frac{\mathbf{B}_n \hat{\mathbf{a}}_n}{\mathbf{B}_s \hat{\mathbf{a}}_s + \mathbf{B}_n \hat{\mathbf{a}}_n} \otimes \mathbf{x} \right\|_2^2 + \left\| \mathbf{y}_n - \frac{\mathbf{B}_s \hat{\mathbf{a}}_s}{\mathbf{B}_s \hat{\mathbf{a}}_s + \mathbf{B}_n \hat{\mathbf{a}}_n} \otimes \mathbf{x} \right\|_2^2 \right) \quad (7)$$

3.3. Optimization

Taking \mathbf{x} as the input of DNN, except the NMF layer and the Wiener-type filtering layer, we recursively compute the activations of all layers in the DNN as follows:

$$\mathbf{z}_{l+1} = \mathbf{W}_l \mathbf{a}_l; \quad \mathbf{a}_{l+1} = f(\mathbf{z}_{l+1}) \quad (8)$$

where \mathbf{a}_l is the activations of the layer (l), \mathbf{z}_l is the total weighted sums of inputs to the layer (l), including the bias term, and \mathbf{W}_l is the connection weights between the layer l and the layer ($l + 1$). $f(\cdot)$ is the element-wise activation function. After obtaining the activations \mathbf{a}_{n_l} of the output layer (n_l), we let $\hat{\mathbf{a}}_n = \mathbf{a}_{n_l}(1, \dots, L_n)$ and $\hat{\mathbf{a}}_s = \mathbf{a}_{n_l}(L_n + 1, \dots, L_n + L_s)$, and then the final outputs $\tilde{\mathbf{y}}_s$ and $\tilde{\mathbf{y}}_n$ can be computed by Eq. (5).

According to the objective function in Eq. (7), we can compute the error metric, and then through the backward propagation of target errors, the network weights can be iteratively updated.

Using the chain rule, the gradients of the objective function with respects to the network weights can be computed as follows:

$$\nabla \mathbf{W}_l = \frac{\partial J}{\partial \mathbf{W}_l} = \frac{\partial J}{\partial \mathbf{z}_{l+1}} \frac{\partial \mathbf{z}_{l+1}}{\partial \mathbf{W}_l} = \frac{\partial J}{\partial \mathbf{z}_{l+1}} (\mathbf{a}_l)^T \quad (9)$$

To simplify notations, we introduce a variable δ and let $\delta_l = \frac{\partial J}{\partial \mathbf{z}_l}$. For the output layer ($l = n_l$), we have:

$$\delta_{n_l} = \frac{\partial J}{\partial \mathbf{a}_{n_l}} \otimes \frac{\partial \mathbf{a}_{n_l}}{\partial \mathbf{z}_{n_l}} = \left[\frac{\partial J}{\partial \hat{\mathbf{a}}_n}; \frac{\partial J}{\partial \hat{\mathbf{a}}_s} \right] \otimes f'(\mathbf{z}_{n_l}) \quad (10)$$

where,

$$\frac{\partial J}{\partial \hat{\mathbf{a}}_n} = (\mathbf{B}_n)^T [(y_s - \tilde{y}_s) - \lambda(y_n - \tilde{y}_s) - (y_n - \tilde{y}_n) + \lambda(y_s - \tilde{y}_n)] \otimes \frac{\tilde{y}_s}{\hat{y}_s + \hat{y}_n} \quad (11)$$

$$\frac{\partial J}{\partial \hat{\mathbf{a}}_s} = (\mathbf{B}_s)^T [-(y_s - \tilde{y}_s) + \lambda(y_n - \tilde{y}_s) + (y_n - \tilde{y}_n) - \lambda(y_s - \tilde{y}_n)] \otimes \frac{\tilde{y}_n}{\hat{y}_s + \hat{y}_n} \quad (12)$$

For the l -th layer ($l = n_l - 1, n_l - 2, \dots, 2$), we have:

$$\delta_l = ((\mathbf{W}_l)^T \delta_{l+1}) \otimes f'(\mathbf{z}_l) \quad (13)$$

After obtaining all δ terms, the partial derivatives of the objective function with respects to the network weights can be computed as follows:

$$\nabla \mathbf{W}_l = (\delta_{l+1}) (\mathbf{a}_l)^T \quad (14)$$

Then, we use the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [19] to update the weights \mathbf{W}_l .

4. EXPERIMENTS

4.1. Dataset and evaluation metrics

In this section, we systematically evaluate the performance of the proposed model with the TIMIT corpus [20] and the NOISEX corpus [21]. The TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The NOISEX contains 15 common types of noises in real-world environments, with length of about 4 min for each. These noises mainly cover a variety of daily noises and most of them are highly non-stationary.

For training, 100 speech utterances from 50 different speakers, with 2 utterances for each speaker, are randomly chosen from the TIMIT training part. They are randomly mixed with 9 types of noises from NOISEX at a SNR of continuous variation from -5 to 5 dB,

which generates 2000 mixture utterances. To obtain the basis spectra \mathbf{B}_s and \mathbf{B}_n of speech and noise, we perform NMF on all utterances of the TIMIT training part and 9 types of noises from NOISEX, respectively. The chosen noises for NMF are the same as those chosen to generate the training set. For testing, we randomly choose 100 clean speech utterances from the TIMIT testing part. Each utterance is recorded by a different speaker. They are randomly mixed with 15 types of noises from NOISEX at -10, -7, -5, -2, 0, 2, 5, 7 and 10 dB, which generates 500 mixture utterances. 6 types of noises among them are unseen in training set for testing the generalization to the unmatched noise. In the same way, we construct the validation set of 500 mixture utterances. But the chosen speech utterances for the validating set are different from those for the training set and the testing set. In addition, we randomly cut each noise utterance of NOISEX into different parts to ensure that the different parts of each noise are used to construct the different datasets.

We take Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR), Source to Distortion Ratio (SDR) and Perceptual Evaluation of Speech Quality (PESQ) [22] as evaluation metrics. SIR, SAR and SDR measure the ratios of source to interference, artifacts and distortion, and can be computed by the BSS Eval toolbox [23]. The PESQ score quantifies the objective speech quality. For all evaluation metrics, higher values mean the better separation quality.

4.2. Related models and setting

To evaluate the proposed model (denoted as ‘P-DNN-NMF’), we systematically compare with the DNN-based (denoted as ‘DNN-1’) [6] and the previous DNN-NMF-based (denoted as ‘DNN-NMF’) speech separation models [12, 24]. In all experiments, a standard DNN, with two hidden layers of 1000 rectified linear units (ReLU) [25], is used as the learning model. All DNNs are trained from a random initialization, and the network weights are updated using the L-BFGS algorithm. The maximum epoch is set to 500. The magnitude spectrograms, extracted by applying a 512-point STFT with 50% overlap to the mixture signals, is used as the input features. To capture the context information, a window (5 frames) of features are concatenated together to form a long feature vector. Although the final objectives of all separation models in this paper are to estimate the magnitude spectrograms of speech, they are achieved by different means. DNN-1 directly predicts the magnitude spectrograms of speech and noise. DNN-NMF predicts the activation coefficients of speech and noise inferred by NMF, and the NMF is trained in advance using the clean speech and noise those are mixed to construct the training set. Then the predicted activation coefficients are used to generate the corresponding magnitude spectrograms with nonnegative linear combination of the basis spectra. And P-DNN-NMF integrates a NMF layer and a Wiener-type filtering layer into the DNN to reconstruct the magnitude spectrograms of speech and noise. The NMF layer is composed of the basis spectra of speech and noise, which are obtained by performing NMF on the TIMIT training part and NOISEX, respectively. In all experiments, we set the number of NMF bases to be 256. To capture temporal structures, NMF is trained on a window of 5 frames rather than single time slices. The original output layers of DNN-NMF and P-DNN-NMF both have 256×2 units and DNN-1 has 257×2 output units. All original output layers use ReLU as the activation function. In addition, To ensure that the estimates of speech and noise sum to the mixture, in testing, a Wiener-type filtering is applied to the outputs of DNN-1

and DNN-NMF to obtain the final estimates of magnitude spectrograms of speech and noise.

4.3. Results and discussions

Firstly, we explore the effects of different λ on the performance of P-DNN-NMF. Table 1 reports the average gains of SDR (GSDR) on the validating set using different λ , which can be computed as follows:

$$GSDR(\hat{s}, s, x) = SDR(\hat{s}, s) - SDR(x, s) \quad (15)$$

GSDR reflects the improvement of overall performance. We can observe that $\lambda = 0.05$ provides better results compared with other values. Hence, we fix λ to 0.05 in the following experiments.

Table 1. The performances using different λ

λ	0.005	0.01	0.05	0.10	0.25	0.50
GSDR	10.00	9.99	10.08	10.06	9.81	9.40

Figure 2 presents the results of different speech separation models (DNN-1, DNN-NMF and P-DNN-NMF) in matched and unmatched noise conditions. We can observe that P-DNN-NMF achieves best results on all evaluation metrics (SDR, SIR, SAR and PESQ) and consistently outperforms DNN-1 and DNN-NMF in matched noise condition. It mainly owes to that the proposed network architecture integrating NMF can capture the invariable spectro-temporal structures in signals and the discriminative separation objective can suppress more noise. Although DNN-NMF also uses NMF to exploit the spectra patterns of speech and noise, NMF and DNN are independently treated. It will lead to a double error problem and that the separation is more sensitive to estimation errors. Therefore, DNN-NMF only achieves limited improvement compared to DNN, especially in unmatched noise conditions.

Finally, we show several examples of separation results in Fig. 3. We can observe that the proposed P-DNN-NMF can recover the speech with high quality. Compared to that from DNN-1, the separated spectrogram from P-DNN-NMF presents cleaner spectra structures and richer details, especially in high frequency bands. The results suggest that P-DNN-NMF can suppress more interferences with less artifacts and speech distortion. It mainly owes to that the nonnegative linear combination of the basis spectra learned by NMF can preserve the speech spectra structures and the discriminative training objective can suppress more interferences.

5. CONCLUSIONS AND RELATED WORKS

In this work, a novel DNN integrating NMF is proposed to exploit spectro-temporal structures of speech and noise for speech separation, and a discriminative training objective is further explored for the proposed model. In the proposed model, NMF is used to learn the spectra patterns of speech and noise, which are added to the original output of DNN as an extra layer. The discriminative training objective enhances the training of the discrimination between speech

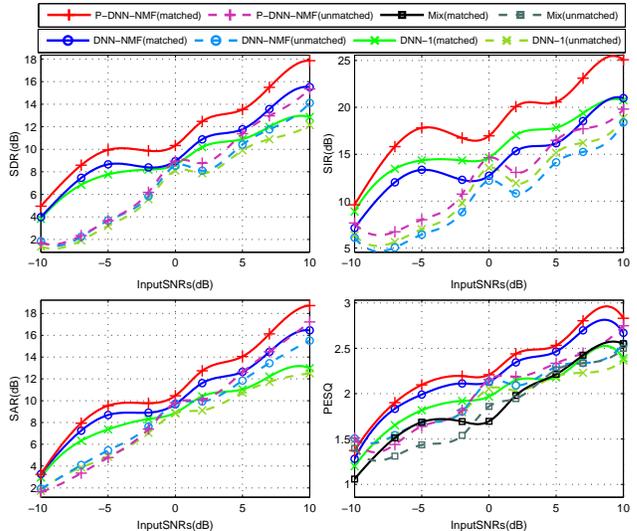


Fig. 2. The results of different speech separation models in matched noise and unmatched noise cases, ‘(matched)’ means the matched noise case and ‘(unmatched)’ means the unmatched noise case. ‘Mix’ means the results of the original mixture signals.

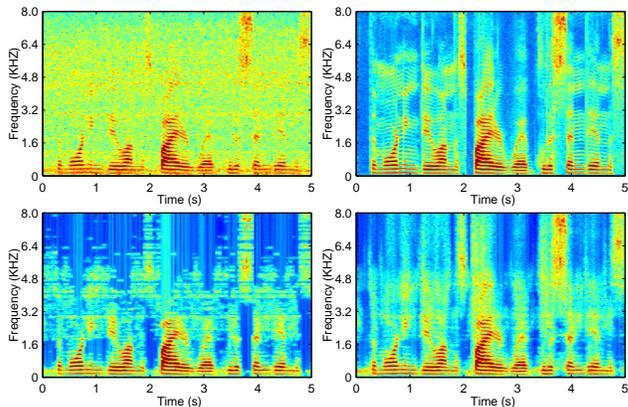


Fig. 3. Top left: the spectrogram of the mixture speech (in log scale); Top right: the groundtruth spectrogram for the speech; Bottom left: the separation result from DNN-1; Bottom right: the separation result from P-DNN-NMF.

and noise, which can achieve higher SIR. The experimental results show that the proposed model significantly improves the performance of speech separation, and also suggest that exploiting spectro-temporal structures of output targets is substantially promising for speech separation. Based on this idea, several methods have been proposed. Wang [12] proposes to use NMF to exploit the spectro-temporal patterns of the square-root ideal ratio mask. Hershey [26] extends NMF to a deep architecture for speech separation. In our previous work [27], we use an autoencoder to exploit the spectro-temporal structures of the ideal ratio mask and the Gammatone frequency power spectrum. In fact, any kind of generative models can be used to exploit the spectro-temporal patterns of signals. These can be further studied in the future.

6. REFERENCES

- [1] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, pp. 1486–1494, 2009.
- [2] H. Dillon, *Hearing aids*, Thieme, 2001.
- [3] J. Allen, "Articulation and intelligibility," *Synthesis Lectures on Speech and Audio Processing*, vol. 1, no. 1, pp. 1–124, 2005.
- [4] M. Seltzer, B. Raj, and R. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.
- [5] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 22, no. 12, pp. 1849–1858, 2014.
- [6] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.
- [7] P.-L. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2014)*, 2014, pp. 1562–1566.
- [8] F. Weninger, J. Le Roux, J. R. Hershey, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE GlobalSIP 2014 Symposium on Machine Learning Applications in Speech Processing*, 2014.
- [9] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2015)*, 2015, pp. 4391–4394.
- [10] A. J. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," *arXiv preprint arXiv:1504.04658*, 2015.
- [11] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep nmf for speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2015)*, 2015, pp. 66–70.
- [12] Y. Wang and D. Wang, "A structure-preserving training target for supervised speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2014)*. IEEE, 2014, pp. 6107–6111.
- [13] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [14] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2008)*. Citeseer, 2008, pp. 4029–4032.
- [15] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 2015.
- [16] Pablo Sprechmann, Alexander Bronstein, Michael Bronstein, and Guillermo Sapiro, "Learnable low rank sparse models for speech denoising," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2013)*. IEEE, 2013, pp. 136–140.
- [17] S. Liang, W. Liu, W. Jiang, and W. Xue, "The optimal ratio time-frequency mask for speech separation in terms of the signal-to-noise ratio," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. EL452–EL458, 2013.
- [18] S. Liang, W. Liu, W. Jiang, and W. Xue, "The analysis of the simplification from the ideal ratio to binary mask in signal-to-noise ratio sense," *Speech Communication*, vol. 59, pp. 22–30, 2014.
- [19] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.
- [20] J. S. Garofolo, Linguistic Data Consortium, et al., *TIMIT: acoustic-phonetic continuous speech corpus*, Linguistic Data Consortium, 1993.
- [21] A. Varga and H. J. 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.
- [22] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2001)*, 2001, vol. 2, pp. 749–752.
- [23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [24] Tae Gyoong Kang, Kisoo Kwon, Jong Won Shin, and Nam Soo Kim, "Nmf-based target source separation using deep neural network," *Signal Processing Letters, IEEE*, vol. 22, no. 2, pp. 229–233, 2015.
- [25] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, 2011, vol. 15, pp. 315–323.
- [26] J. R. Hershey, J. L. Roux, and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," *arXiv preprint arXiv:1409.2574*, 2014.
- [27] S. Nie, S. Liang, W. Xue, X. Zhang, W. Liu, L. Dong, and H. Yang, "Two-stage multi-target joint learning for monaural speech separation," in *Proc. 16th Annual Conference of the International Speech Communication Association (INTER-SPEECH'2015)*, 2015.