SUPERVISED MONAURAL SOURCE SEPARATION BASED ON AUTOENCODERS

Keiichi Osako[1], Yuki Mitsufuji[1], Rita Singh[2] and Bhiksha Raj[2]

1. Sony Corporation, Minato-ku, Tokyo, Japan
2. Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

In this paper, we propose a new supervised monaural source separation based on autoencoders. We employ the autoencoder for the dictionary training such that the nonlinear network can encode the target source with high expressiveness. The dictionary is trained by each target source without the mixture signal, which makes the system independent from the context where the dictionaries will be used. In separation process, the decoder portions of the trained autoencoders are used as dictionaries to find the activations in a iterative manner such that a summation of the decoder outputs approximates the original mixture. The results of the instruments source separation experiments revealed that the separation performance of the proposed method was superior to that of the NMF.

Index Terms— source separation, autoencoder, neural networks, non-negative matrix factorization

1. INTRODUCTION

The problem of source separation has continued to receive the attention of researchers for several decades. In case of monaural source separation, given a monaural recording that includes a mixture of sounds from multiple sources, the goal is to separate out the individual sources.

Non-negative matrix factorization (NMF) [1] is currently one of the most popular techniques for source separation. NMF employs compositional models that are interpreted as low-rank representations, or *bases*, of the magnitude or power spectrum components of reference sources. Over the years, various NMF methods have been proposed [2, 3]. Sparse NMF [4] includes sparsity constraints explicitly into the NMF cost function. Convolutive model of NMF [5] employs spectro-temporal patterns as bases to extract meaningful components out of target spectrograms. Phoneme-dependent NMF [6] learns separate bases for each phoneme of the language. Although these methods have provided successful results, the performance depends on the expressiveness and selectivity of the dictionary.

The basic NMF approach is *generative* – the dictionaries for each source are trained independently. This has the advantage that no knowledge is required about the competing sources when training the model for any source; as a consequence, the approach is scalable.

As an alternative, several works to train *discriminative* bases for NMF have been proposed [7, 8, 9, 10, 11]. In contrast to conventional NMF where the basis vectors are trained independently on each source, these approaches optimize target and non-target basis vectors jointly using mixture signals. In other words, the discriminative methods learn the dictionaries as mixture-specific models. Therefore, while these models provide significantly improved performance over the basic *generative* NMF model, their performance is specific to the mixture contexts for which the dictionaries have been trained.

An alternate, widely-studied successful trend employs deep neural network (DNN) approaches for source separation [12, 13, 14, 15, 16]. DNN for monaural separation [17] works as a spectrum domain classifier which can classify its input spectrum into each source type. In the recurrent DNN approach [18], given a mixture spectrogram, the DNN directly reconstructs the target spectrograms. Although these DNN approaches are proved to outpeform the NMF approach, the frameworks cannot be easily extended to different contexts since DNNs are fundamentally discriminatively trained, and the mixture information is essential in the training stage, which limits the frameworks to the specific mixture context to which it is trained.

In our work we attempt to retain the advantages offered by both, the deep neural network architecture, and the natural scalability of the *generative* approach. Thus, our focus is twofold:

- Our dictionary employs compositional models represented as a constructive nonlinear combination as neural networks to improve the expressiveness of sources.
- In the training stage, the dictionary is trained independently for each target source which makes it independent of the mixture context in which it will be employed.

In this paper, we employ the autoencoder mechanism [19] for the dictionary training such that the nonlinear network can encode the target source with higher expressiveness than a generative NMF model. In effect, we are employing deep autoencoders as dictionaries to represent sources. Note that our intention is not to improve over the *discriminative* model, but rather the *generative* one. In separation process, the decoder portions of the trained autoencoders are used as dictionaries to find the activations in a iterative manner such that a summation of the decoder outputs approximates the original mixture.

This paper is organized as follows: Section 2 reviews supervised source separation based on NMF algorithm. Section 3 describes the proposed source separation method based on autoencoders. Section 4 shows our experimental setup and results. Finally, we present our conclusions in section 5.

2. SUPERVISED SOURCE SEPARATION BASED ON NMF

This section reviews the formulation of NMF [1] for supervised source separation. Let $\mathbf{Y} \in \mathbb{R}_+^{I \times J}$ be the magnitude spectrogram of a monaural observation. An element of matrix $[\mathbf{Y}]_{i,j} = y(i,j)$ denotes the magnitude spectrum coefficient. The frequency index and the frame index are denoted by $i = 1, \ldots, I$ and $j = 1, \ldots, J$, respectively. Although, in this paper, the discussion to limited the case of mixtures of two sound sources for simplicity, it can easily be extended to include the case of separation of three or more sources.

2.1. Dictionary training

In the NMF algorithm, a target (magnitude) spectrogram is represented as

$$\mathbf{Y} \simeq \mathbf{WH}, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}_+^{I \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times J}$ are regarded as a dictionary of the target source and a set of corresponding activations, respectively. K is the number of NMF basis vectors. The beta divergence is often applied to minimize the error between $\mathbf{Y}$ and $\mathbf{WH}$ as:

$$D_\beta(\mathbf{Y}, \mathbf{WH}) = \sum_{i=1}^{I} \sum_{j=1}^{J} d_\beta(y(i,j), \hat{y}(i,j)) \tag{2}$$

with

$$\hat{y}(i,j) = \sum_{k=1}^{K} w(i,k) h(k,j), \tag{3}$$

where $d_\beta()$ denotes the divergence function: $\beta = 0$ is the Itakura-Saito divergence [3], $\beta = 1$ is the generalized Kullback-Leibler (KL) divergence and $\beta = 2$ is the squared Euclidean distance.

In the training process, $\mathbf{W}$ and $\mathbf{H}$ are obtained by multiplicative update rules and $\mathbf{W}$ is stored for the separation process whereas $\mathbf{H}$ is discarded. In the case of separation of two sources, $\mathbf{W}_{\text{dic1}}$ and $\mathbf{W}_{\text{dic2}}$ are trained individually for the corresponding target sources.

2.2. Separation process

For the separation, the trained matrices $\mathbf{W}_{\text{dic1}}$ and $\mathbf{W}_{\text{dic2}}$ are fixed while the activations $\mathbf{H}_1$ and $\mathbf{H}_2$ are updated to best represent the mixture $\mathbf{X}$ as below:

$$\mathbf{X} \simeq \begin{bmatrix} \mathbf{W}_{\text{dic1}} \mathbf{W}_{\text{dic2}} \end{bmatrix} \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix}. \tag{4}$$

We recover an estimate of the target source spectrogram as a matrix product of the target dictionary and its corresponding activation matrix, e.g., $\mathbf{W}_{\text{dic1}}\mathbf{H}_1$. The final separated spectrograms are obtained by multiplying the spectrogram of the mixed source by a Wiener filter composed from the estimated target source spectrograms [20], and inverting the result to a time-domain signal using the phase information of the observed signal.

3. SUPERVISED SOURCE SEPARATION BASED ON AUTOENCODERS

This section describes the supervised approach based on autoencoders. Similar to the NMF approach, the proposed algorithm consists of two stages: the first stage is to train sound dictionaries by autoencoders, the next stage is to separate a mixture signal using decoder portions of the trained dictionaries from the first stage. We will refer to the proposed algorithm as AESS (AutoEncoder based Source Separation).

3.1. Dictionary training

In the training process, autoencoders are individually trained to make dictionaries for target signals. The bottom part of figure 1 shows the training process of two target sources. The networks for two sources can be expressed by encoder $f_{\text{ENC1,2}}$ and decoder $f_{\text{DEC1,2}}$ functions described as below:

$$\begin{aligned} \hat{\mathbf{Y}}_1 &= f_{\text{DEC1}}(f_{\text{ENC1}}(\mathbf{Y}_1)) \\ &= f_{\text{DEC1}}(\mathbf{H}_1) \\ &= g\Big(\mathbf{W}_{\text{DEC1}}^{(L)} \cdots g\big(\mathbf{W}_{\text{DEC1}}^{(l)} g(\mathbf{W}_{\text{DEC1}}^{(1)} \mathbf{H}_1)\big)\Big), \end{aligned} \tag{5}$$

$$\begin{aligned} \hat{\mathbf{Y}}_2 &= f_{\text{DEC2}}(f_{\text{ENC1}}(\mathbf{Y}_2)) \\ &= f_{\text{DEC2}}(\mathbf{H}_2) \\ &= g\Big(\mathbf{W}_{\text{DEC2}}^{(L)} \cdots g\big(\mathbf{W}_{\text{DEC2}}^{(l)} g(\mathbf{W}_{\text{DEC2}}^{(1)} \mathbf{H}_2)\big)\Big), \end{aligned} \tag{6}$$

where $\hat{\mathbf{Y}}_{1,2}$ and $\mathbf{Y}_{1,2}$ denote the output and input of the autoencoders, respectively. $\mathbf{H}_{1,2}$ denote the bottle-neck features from the hidden layer, superscript $\cdot^{(l)}$ is the index of the hidden layer, L denotes the number of layers, and $g()$ denotes the nonlinear activation functions such as sigmoid, hyperbolic tangent and ReLU (Rectified Linear Unit) [21].

The autoencoders are trained to minimize the squared error between the output and the input:

$$C(\hat{\mathbf{Y}}_1, \mathbf{Y}_1) = \frac{1}{2} \|\hat{\mathbf{Y}}_1 - \mathbf{Y}_1\|_F^2 + \lambda \|\mathbf{H}_1\|_1, \tag{7}$$

$$C(\hat{\mathbf{Y}}_2, \mathbf{Y}_2) = \frac{1}{2} \|\hat{\mathbf{Y}}_2 - \mathbf{Y}_2\|_F^2 + \lambda \|\mathbf{H}_2\|_1, \tag{8}$$

where $\|\cdot\|_F^2$ denotes Frobenious norm, $\|\cdot\|_1$ denotes L_1 norm and λ is a sparsity constraint parameter of L_1-norm regularization for $\mathbf{H}_1$ and $\mathbf{H}_2$.

3.2. Separation Process

The AESS conducts the separation using the decoder portions of the autoencoder, $f_{\text{DEC1}}()$ and $f_{\text{DEC2}}()$. The main idea of the separation process is to explore $\mathbf{H}_1$ and $\mathbf{H}_2$ such that a summation of decoded features approximates the observed mixture. The weights of the Wiener filter α_1 and α_2 are also updated to finally reconstruct the target spectrograms together with decoded features used as soft masks. The top part of figure 1 shows the separation process.

Prior to the optimization step, $\mathbf{H}_1$ and $\mathbf{H}_2$ are initialized by encoder portions of trained networks such as $\mathbf{H}_1 = f_{\text{ENC1}}(\mathbf{X})$ and $\mathbf{H}_2 = f_{\text{ENC2}}(\mathbf{X})$, respectively.

Additionally, we calculate $\hat{\mathbf{Y}}_1$ and $\hat{\mathbf{Y}}_2$ from initialized $\mathbf{H}_1$ and $\mathbf{H}_2$ using decoders described in Eq. (5) and (6).

The estimate of the mixture is expressed as a weighted sum of $\hat{\mathbf{Y}}_1$ and $\hat{\mathbf{Y}}_2$ as:

$$\hat{\mathbf{X}} = \alpha_1 \hat{\mathbf{Y}}_1 + \alpha_2 \hat{\mathbf{Y}}_2, \tag{9}$$

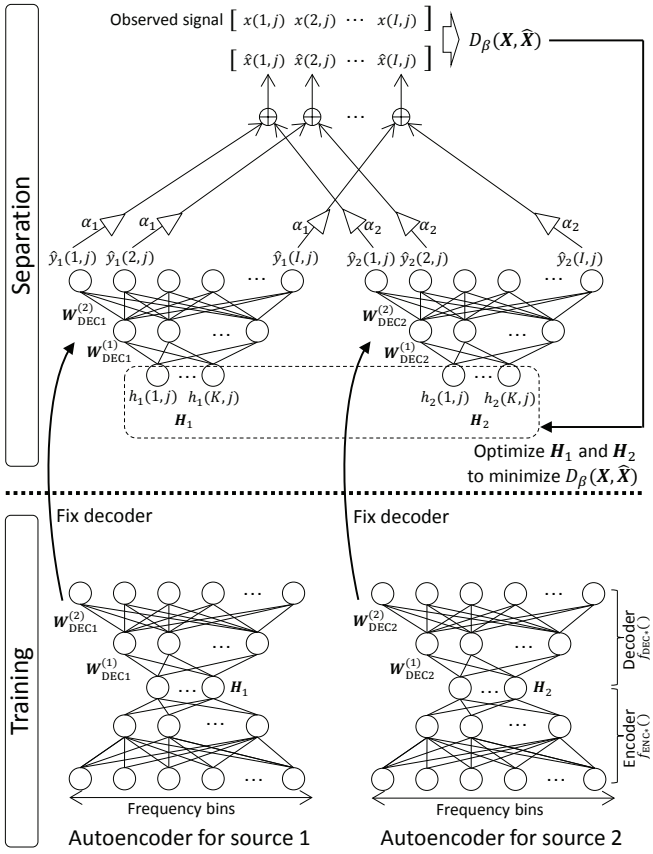

Fig. 1. Diagram of the autoencoder based source separation (AESS) algorithm. AESS consists of a dictionary training process (bottom) and a separation process (top).

where α_1 and α_2 denote scalar coefficients to control the mixture ratio of decoded features. The cost function for the AESS approach is also based on beta divergence described in Eq. (2):

$$D_\beta(\mathbf{X}, \hat{\mathbf{X}}) = \sum_{i=1}^{I} \sum_{j=1}^{J} d_\beta(x(i,j), \hat{x}(i,j)). \tag{10}$$

To achieve the separation, we try to minimize the divergence by updating $\mathbf{H}_1$, $\mathbf{H}_2$, α_1 and α_2:

$$\{\mathbf{H}_1, \mathbf{H}_2, \alpha_1, \alpha_2\} = \underset{\mathbf{H}_1, \mathbf{H}_2, \alpha_1, \alpha_2}{\arg\min} \; D_\beta(\mathbf{X}, \hat{\mathbf{X}}). \tag{11}$$

For the updates, an arbitrary optimization method such as gradient descent can be used. In our work we use gradient descent through backpropagation. When the optimization is completed by minimizing the divergence, the decoded features are expected to represent the estimates of the magnitude spectrograms for target sources.

Finally, the separated spectrograms are obtained by Wiener filtering with the mixture ratio α_1 and α_2:

$$\hat{\mathbf{S}}_1 = \mathbf{X} \circ \frac{\alpha_1 \hat{\mathbf{Y}}_1}{\alpha_1 \hat{\mathbf{Y}}_1 + \alpha_2 \hat{\mathbf{Y}}_2}, \tag{12}$$

$$\hat{\mathbf{S}}_2 = \mathbf{X} \circ \frac{\alpha_2 \hat{\mathbf{Y}}_2}{\alpha_1 \hat{\mathbf{Y}}_1 + \alpha_2 \hat{\mathbf{Y}}_2}, \tag{13}$$

where the symbol $\circ$ denotes the Hadamard product (element-wise multiplication).

Although the basic structure of representing spectrograms based on a multiplication of matrices in Eq. (5) is analogous to the supervised NMF formulation in Eq. (1), there are two differences. First, AESS uses nonlinear functions and multistage matrix operations enabling us to increase the expressiveness of dictionaries. Second, if ReLU is used as the nonlinear function, the decoder coefficients can both be negative and non-negative values while keeping the non-negativity of the reconstructed spectrograms. On top of the above features, the trained dictionary is not dependent on the mixture context of mixture signals, as described in section 1. Thus, the process is scalable, in that a model that is trained once can be applied to any arbitrary mixture that includes the signal.

4. EVALUATION

4.1. Experimental setup

To evaluate the proposed algorithm, we compared our proposed method with sparse NMF [4]. We used the Bach10 dataset [22] consisting of ten monaural recordings of four instruments: violin, clarinet, saxophone and bassoon. All recordings were downsampled to 16 kHz. Magnitude spectrums were computed from the dataset using 1024-point STFT with half overlap.

We trained four autoencoders individually from the instruments recordings to extract the decoder portions. Eight musical pieces (set No.1-8) were used as training data, the set No.9 as validation data, and set No.10 as test data. The key parameters of the autoencoders in our experiments were (a) the layer of the network: three, five or seven, (b) learning rate: 0.01, (c) nonlinear function: ReLU, (d) sparsity coefficient: $\lambda = 10^{-4}$, (e) dimensions of input and output layer: 513, and (f) L2 regularization: 10^{-4}. The parameters of separation process were (i) step size of gradient descent: 10^{-3}, (ii) iterations: 3000, and (iii) optimization frames: all frames per input data as batch process.

The separation performance was evaluated by signal-to-distortion ratio (SDR) improvement obtained from the BSS Eval Toolbox [23]. For training NMF, the number of bases was set to 320 as this was found to provide the best results on the dataset.

4.2. Results

4.2.1. Comparison of instruments pairs

Table 1 compares the separation performance of different algorithms applied to various pairs of instruments. Euclid (EU) distance and generalized Kullback-Leibler (KL) divergence were used for cost functions. Compared with sparse NMF on average, both EU and KL-based AESS show better SDR improvements. In particular, KL-based AESS shows significantly better results for the most of the instrument pairs except S-B.

4.2.2. Comparison of decoder configuration

Table 2 shows averaged SDR improvements of all instrument pairs in case of different decoder configurations. For example, a decoder configuration 20-600-513 denotes that the autoencoder used for trainings comprised fully connected five layers of orders 513-600-20-600-513. The result shows that using a decoder with deeper layers leads to a better performance. In addition, regardless of the

Pair of Instruments	SDR Improvement [dB]		
	sNMF [4]	AESS (EU)	AESS (KL)
V-C	10.86	10.53	12.59
V-S	9.43	10.91	11.28
V-B	8.61	8.63	9.10
C-S	11.18	12.17	12.85
C-B	10.93	10.80	11.57
S-B	8.79	7.78	8.00
Average	9.97	10.14	10.90

Table 1. Source separation performance of various pair of instruments. Alphabets denote the instrument names, V: Violin, C: Clarinet, S: Saxophone and B: Bassoon. The cost function was used EU: Euclid and KL: generalized Kullback-Leibler divergence.

decoder configurations, the decoders with 20 hidden-layer units show the best results.

Decoder Configuration	Average SDR Improvement [dB]
20-513	7.85
20-600-513	10.69
10-200-600-513	9.09
20-200-600-513	10.58
50-200-600-513	9.39
10-200-800-513	9.50
20-200-800-513	10.90
50-200-800-513	9.31

Table 2. Source separation performance of different decoders configuration. The cost function of AESS was used generalized Kullback-Leibler divergence.

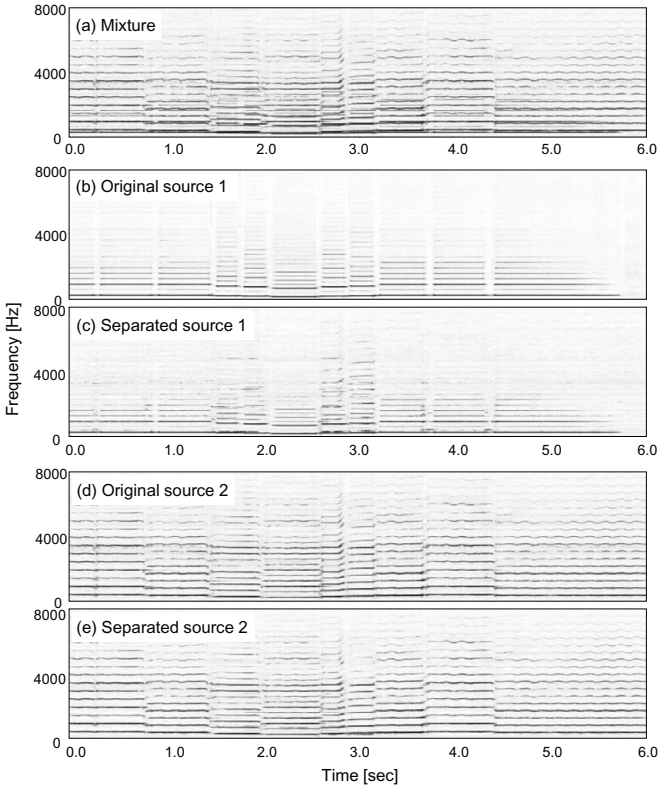

Fig. 2. Source separation example using the Bach10 dataset. (a) The mixture of clarinet and violin; (b) original source 1 of clarinet; (c) separated source 1; (d) original source 2 of violin; and (e) separated source 2

4.2.3. Spectrograms of separated sources

Figure 2 shows the spectrograms in the case of two instrument separation of V-C. The separated source 1 and 2 were obtained by KL-based AESS which layers are 20-200-800-513. Comparing the separated sources with original sources, it is clear that a structure of harmonics and vibrato components were kept through the separation process.

The results of experiments revealed that source separation performance can be improved by the autoencoder trainings. Additional examples can be found at http://mlsp.cs.cmu.edu/projects/AESS

5. CONCLUSION

In this paper, we have proposed a new supervised monaural source separation based on autoencoders. The algorithm employs deep autoencoders as the dictionary for signals, such that the nonlinear network can encode the target source with high expressiveness. The approach is generative – the dictionaries are not dependent on the mixture context of signals. The results of the instruments source separation experiments reveal that the separation performance of the proposed AESS was superior to that of sparse NMF.

However, the formalism leads us to many potential extensions. Although our dictionaries have many parameters, they effectively represent only a small number of degrees of freedom, i.e. the size of the bottleneck. These are effectively *low-rank* non-linear representations. We are currently investigating the other end of the domain, namely *high-rank* domains where, instead of a bottleneck, the network has a large number of intermediate nodes with appropriate sparsity constraints.

The nature of the model also naturally extends to *recurrent* models. The auto-encoder framework is easily extended to RNNs and LSTMs, to obtain more expressive dictionaries that also capture temporal structure. We are also investigating this avenue.

Finally, we also plan to extend our evaluations to more challenging problems, such as with more than two sources, possibly in the presence of noise that must also be estimated. The generative framework easily extends to these scenarios.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems (NIPS)*. MIT Press, 2001, vol. 13.

[2] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *the International Computer Music Conference (ICMC)*, 2003.

[3] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[4] J. Eggert and E. Korner, "Sparse coding and NMF," in *Neural Networks*, 2004, vol. 4, pp. 2529–2533.

[5] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. on audio, speech and language processing*, vol. 15, no. 1, pp. 1–12, 2007.

[6] B. Raj, R. Singh, and T. Virtanen, "Phoneme-dependent NMF for speech enhancement in monaural mixtures," in *Interspeech 2011*, 2011, pp. 1217–1220.

[7] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *Interspeech 2014*, 2014, pp. 865–869.

[8] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3777–3781.

[9] J. L. Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 66–70.

[10] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation," in *Interspeech 2013*, 2013, pp. 808–812.

[11] A. Gang and P. Biyani, "On discriminative framework for single channel audio source separation," in *Interspeech 2016*, 2016, pp. 565–569.

[12] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2135–2139.

[13] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[14] H. Li, S. Nie, X. Zhang, and H. Zhang, "Jointly optimizing activation coefficients of convolutive NMF using DNN for speech separation," in *Interspeech 2016*, 2016, pp. 550–554.

[15] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, 2015.

[16] M. Sun, X. Zhang, H. V. Hamme, and T. F. Zheng, "Unseen noise estimation using separable deep auto encoder for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 93–104, 2016.

[17] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3734–3738.

[18] P. S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[19] G. E. Hinton and R. R. Aalakhutdinov, "Reducing the dimensionality of data with neural networks," in *Science*, July 2006, vol. 313, pp. 504–507.

[20] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[21] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on artificial intelligence and statistics (AISTATS)*, 2011, pp. 315–323.

[22] Z. Duan and B. Pardo, "Soundprism: an online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Process*, vol. 5, no. 6, pp. 1205–1215, 2011.

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