INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Real-Time Independent Vector Analysis Using Semi-Supervised Nonnegative Matrix Factorization as a Source Model

Taihui Wang[1,2], *Feiran Yang*[1,2], *Rui Zhu*[3], *Jun Yang*[1,2]

[1] Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Tencent Ethereal Audio Lab, Tencent Corporation, Beijing, China
{wangtaihui, feiran, jyang}@mail.ioa.ac.cn raymondrzhu@tencent.com

Abstract

Online independent vector analysis (IVA) based on auxiliary technology is effective to separate audio source in real time. However, the separated signal may contain residual interference noise because the source model of IVA lacks flexibility and cannot treat the specific harmonic structures of sources. This paper presents a real-time IVA method where the amplitude spectrum of separated signal is modeled by semi-supervised nonnegative matrix factorization (SSNMF). Using the pre-trained basis matrix which contains source structures, we can extract the target source from the separated signal in real time. The advantage of the proposed method is that the extracted source can provide a more accurate variance than the separated signal and hence the proposed method can obtain a better separation performance than the oracle IVA. Experimental results in speech denoising task show the effectiveness and the robustness of the proposed method with different types of noise.

Index Terms: audio source separation, speech denoising, real-time independent vector analysis, semi-supervised nonnegative matrix factorization

1. Introduction

Audio source separation (ASS) is a crucial technology to estimate sources using information about their mixtures [1]. ASS has many underlying applications, such as speech recognition interfaces, hearing-aid devices and mobile telephones.

Most of ASS algorithms [2, 3, 4] are performed in a block-wise manner. For (over) determined case, independent vector analysis (IVA) [3] was developed. The cost function of IVA exploits the statistical independence between sources by assuming that the source frequency bins obey spherical multivariate Laplace distribution. Natural gradient (NG) algorithm and auxiliary function technology were used to optimize the cost function of IVA. In the literature, there are only few works that focus on the online implementation of ASS. Kim proposed a real-time IVA method based on NG (NGIVA) [5]. The convergence and stability of real-time NGIVA generally depend on the environment-sensitive learning rate [6]. Determining the learning rate suffers from the trade-off between convergence time and smoothness of steady-state solution [7]. The normalized Frobenius norm was used in [8] to introduce an adaptive learning rate. However, the calculation of adaptive learning rate is time-consuming [8]. Taniguchi et al. proposed an real-time IVA method based on auxiliary technology (AuxIVA), which shows better separation performance and faster convergence than real-time NGIVA [6]. Though other source prior distributions such as the Student's t distribution [9] and mixed source distribution [10] have been exploited, the separation performance of IVA-based methods is limited because the source model lacks flexibility and cannot treat the specific harmonic structures of sources [11]. It is well-known that nonnegative matrix factorization (NMF) can capture the structures of the audio spectrogram effectively [12]. Standard NMF is an unsupervised technology to factorize the observed matrix into the product of two matrices [13, 14]: one is the basis matrix containing the structures and the other is the activation matrix. NMF has been successfully used in ASS [11, 15, 16, 17], since audio signals consist of units such as phonemes, syllables and words.

In this paper, we propose a real-time AuxIVA method where semi-supervised NMF (SSNMF) [18] is used to model the amplitude spectrum of separated signal obtained by AuxIVA. Using the pre-trained basis matrix which contains source structures, we can extract the target source from the separated signal in real time. The strength of the proposed method is that the extracted source can provide a more accurate variance than the separated signal. Due to the accurate source variance estimation, a better separation performance can be achieved by the online IVA. Experiments are presented to validate the effectiveness of the proposed method.

2. Problem Formulation

Let the number of sources and channels (microphones) be N and M, respectively. The time-frequency domain representations of sources, observed and separated signals are described as

$$\mathbf{s}_{ft} = (s_{1,ft} \cdots s_{N,ft})^{\mathsf{T}}, \tag{1}$$

$$\mathbf{x}_{ft} = (x_{1,ft} \cdots x_{M,ft})^{\mathsf{T}}, \tag{2}$$

$$\mathbf{y}_{ft} = (y_{1,ft} \cdots y_{N,ft})^{\mathsf{T}}, \tag{3}$$

where $f = 1, \cdots, F$, $t = 1, \cdots, T$, $n = 1, \cdots, N$ and $m = 1, \cdots, M$ are the indices of frequency bins, time frames, sources, and channels, respectively. $(\cdot)^{\mathsf{T}}$ denotes vector transposition. When the window length of the short-time Fourier transform (STFT) is sufficiently long compared to the length of the room impulse responses between sources and microphones, the observed signal can be approximated by an instantaneous mixture

$$\mathbf{x}_{ft} \approx \mathbf{A}_f \mathbf{s}_{ft}, \tag{4}$$

where $\mathbf{A}_f = (\mathbf{a}_{f,1} \cdots \mathbf{a}_{f,N}) \in \mathbb{C}^{M \times N}$ is the mixing matrix and $\mathbf{a}_{f,n} \in \mathbb{C}^M$ is the steering vector between the n-th source and all M microphones. In the determined case $M = N$, the separated sources can be estimated as a linear demixing process

$$\mathbf{y}_{ft} = \mathbf{W}_{ft} \mathbf{x}_{ft}, \tag{5}$$

where $\mathbf{W}_{ft} = (\mathbf{w}_{ft,1} \cdots \mathbf{w}_{ft,N})^{\mathsf{H}}$ is the demixing matrix and $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose. Our goal is to recover the

http://dx.doi.org/10.21437/Interspeech.2021-146

separated signal $\mathbf{y}_{ft}$ by estimating the demixing matrix $\mathbf{W}_{ft}$ given the observed signal $\mathbf{x}_{ft}$.

3. Related Work

In this section, we briefly review how the online AuxIVA [6] estimates the demixing matrix in real time and introduce the standard NMF.

3.1. Real-time AuxIVA

In IVA, spherical Laplace distribution is assumed as the source prior

$$p(\mathbf{s}_{n,t}) = \rho \exp\left(-\frac{||\mathbf{s}_{n,t}||_2}{\sigma_{n,t}^2}\right), \tag{6}$$

where $\mathbf{s}_{n,t} = [s_{n,1t} \cdots s_{n,Ft}]^T$ and $\sigma_{n,t}^2$ is the uniform variance over the frequency bins. The higher-order correlation between frequency bins can be exploited by this multivariate spherical Laplace distribution. If we assume the independence between the observed channels, the demixing matrices can be estimated by minimizing the following negative log-likelihood function [3]:

$$Q(\mathbf{W}) = \sum_n \frac{1}{T} \sum_t G(\mathbf{y}_{n,t}) - \sum_f \text{log}|\text{det}\mathbf{W}_{ft}|, \tag{7}$$

where $\mathbf{W}$ denotes a set of $\mathbf{W}_f$, $\mathbf{y}_{n,t} = [y_{n,1t} \cdots y_{n,Ft}]^T$ and $G(\mathbf{y}_{n,t}) = -\text{log}p(\mathbf{y}_{n,t})$ is called a contrast function. To minimize (7), the auxiliary function technology is used in [19] to derive the offline updates rules of $\mathbf{W}$ as

$$\mathbf{V}_{n,f} = \frac{1}{T} \sum_t [\frac{1}{\sigma_{n,t}^2} \mathbf{x}_{ft} \mathbf{x}_{ft}^{\mathsf{H}}], \tag{8}$$

$$\mathbf{w}_{n,f} \leftarrow (\mathbf{W}_f \mathbf{V}_{n,f})^{-1} \mathbf{e}_n, \tag{9}$$

$$\mathbf{w}_{n,f} \leftarrow \frac{\mathbf{w}_{n,f}}{\sqrt{\mathbf{w}_{n,f}^{\mathsf{H}} \mathbf{V}_{n,f} \mathbf{w}_{n,f}}}, \tag{10}$$

where $\mathbf{e}_n$ is a column vector whose n-th element is one and all the others are zero. For online implementation of AuxIVA, $\mathbf{V}_{n,f}$ should be calculated at each time frame t in a frame-by-frame way. An autoregressive calculation of $\mathbf{V}_{n,f,t}$ is derived in [6] as follows:

$$\mathbf{V}_{n,f,t} = \alpha \mathbf{V}_{n,f,t-1} + (1-\alpha)[\frac{1}{\sigma_{n,t}^2} \mathbf{x}_{ft} \mathbf{x}_{ft}^{\mathsf{H}}], \tag{11}$$

where α is a forgetting factor with $0 \leq \alpha < 1$. The demixing matrix at the t-th time frame can be derived as

$$\mathbf{w}_{n,ft} \leftarrow (\mathbf{W}_{ft} \mathbf{V}_{n,f,t})^{-1} \mathbf{e}_n, \tag{12}$$

$$\mathbf{w}_{n,ft} \leftarrow \frac{\mathbf{w}_{n,ft}}{\sqrt{\mathbf{w}_{n,ft}^{\mathsf{H}} \mathbf{V}_{n,f,t} \mathbf{w}_{n,ft}}}. \tag{13}$$

Note that the uniform source variance $\sigma_{n,t}^2$ is calculated using the l_2 norm of the separated source

$$\sigma_{n,t}^2 = ||\mathbf{y}_{n,t}||_2^2 = \sum_f |y_{n,ft}|^2. \tag{14}$$

3.2. Standard NMF

NMF has been used in ASS successfully, because the spectral structure of audio source, such as musical notes and speech harmonics, can be extracted into the basis matrix [1], and pictures describing speech harmonics can be found in [20]. Standard NMF is an unsupervised technology used to decompose the nonnegative amplitude spectrum $|\mathbf{X}|$ into the product of two nonnegative matrices as

$$|\mathbf{X}| = \sum_k \mathbf{b}_k \mathbf{h}_k, \tag{15}$$

where $|\mathbf{X}| \in \mathbb{R}_+^{F \times T}$, $\mathbf{b}_k$ is the k-th column vector of $\mathbf{B} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{B}$ is the frequency-dependent basis matrix containing the features. $\mathbf{h}_k$ is the k-th row vector of $\mathbf{H} \in \mathbb{R}_+^{K \times T}$ and $\mathbf{H}$ is the time-dependent activation matrix. K is the basis number.

4. The Proposed Method

In this section, we present a real-time AuxIVA method where the amplitude spectrum of separated signal $\mathbf{y}_{n,t}$ is modeled by SSNMF. The proposed method has two cascaded stages. The demixing matrix is estimated by AuxIVA in the first stage and the separated signals can be obtained. In the second stage, SSNMF is used to extract the source from the separated signal by the pre-trained basis matrix which contains source structures. More accurate source variance can be estimated in the second stage and be used to perform the online AuxIVA in the first stage.

We derive the proposed algorithm in the determined case where $N = M = 2$ for a speech denoising task. One of the sources is speech and the other one is noise. At the t-th time frame, back-projection technique [21] should be used to recover the scales of the separated speech $\mathbf{y}_{1,t}$ and noise $\mathbf{y}_{2,t}$. In the oracle IVA, $|\mathbf{y}_{n,t}|$ is directly used to estimate $\sigma_{n,t}^2$. However, the separated speech $\mathbf{y}_{1,t}$ may contain the residual interference noise, and hence it may be difficult to obtain an exact estimation of $\sigma_{n,t}^2$. We here model the amplitude spectrum of separated speech as the sum of speech signal $|\hat{\mathbf{y}}_{1,t}|$ and residual noise $|\mathbf{v}_t|$ as

$$|\mathbf{y}_{1,t}| \approx |\hat{\mathbf{y}}_{1,t}| + |\mathbf{v}_t|. \tag{16}$$

To estimate the amplitude spectrum of speech signal $|\hat{\mathbf{y}}_{1,t}|$, the standard NMF can be used to decompose $|\mathbf{y}_{1,t}|$ as

$$|\mathbf{y}_{1,t}| = \sum_k \mathbf{b}_k h_k, \tag{17}$$

where h_k is the activation scalar. Then, the speech signal $|\hat{\mathbf{y}}_{1,t}|$ can be obtained by

$$|\hat{\mathbf{y}}_{1,t}| = \sum_{k \in speech} \mathbf{b}_k h_k. \tag{18}$$

Using the standard NMF to extract the speech signal has two problems. The first one is how to cluster these basis vectors into speech and residual noise, which is a challenging problem [1]. The second one is that NMF has the drawback of learning the structures in real-time, because the structures of audio signal are long-time dependent. The second one is also the difficulty for the real-time implementation of ILRMA [11]. To overcome these problems, the supervised NMF (SNMF) [22] can be used. In SNMF, all the basis vectors $\mathbf{b}_k$ are trained previously on the training dataset. When performing real-time NMF on test

Algorithm 1 real-time SSNMF-AuxIVA

```
Input: observed signal, pre-trained $\bar{\mathbf{B}}_s$
output: separated speech and separated noise
Initialize $\mathbf{W}, \mathbf{B}_n, \mathbf{h}_s, \mathbf{h}_n$
for $t \leftarrow 1 \cdots T$ do
    Iterations for online AuxIVA optimization
        for $f \leftarrow 1 \cdots F$ do
            Update $\sigma_{1,t}^2, \sigma_{2,t}^2$ by (14)
            Update $\mathbf{V}_{n,f,t}$ by (11)
            Update $\mathbf{W}_{ft}$ by (12) and (13)
        end for
    Update $\mathbf{y}_{n,t}$ by (5)
    Fix scales of $\mathbf{y}_{n,t}$ with back projection
    Iteration for SSNMF optimization
        Update $\mathbf{B}_n$ by (21)
        Update $\mathbf{h}_s, \mathbf{h}_n$ by (22)
    Reconstruct $\mathbf{y}_{1,t}$ by (23)
    Update $\sigma_{1,t}^2$ and $\sigma_{2,t}^2$ by (24) and (25)
end for
```

dataset, the basis vectors are fixed while the time-dependent activation matrix is updated over time.

However, SNMF is not flexible because training datasets for all sources should be prepared, which is not suitable here because the training dataset for residual interference noise $\mathbf{v}_t$ is unavailable. Another disadvantage of SNMF is that all sources are assumed to have constant spectral structures, which is not appropriate for most noise sources. Therefore, the semi-supervised NMF (SSNMF) [18] is used in the proposed method, where only the speech basis vectors $\mathbf{b}_k$ ($k \in speech$) are pre-trained to extract the speech feature in real time when performing the speech denoising task.

The pre-trained basis matrix of speech can be learned by standard NMF or sparse NMF [23]. In the sparse NMF, the sparsity constraint of the activation matrix is considered. To make the basis matrix discriminative from each other, the discriminative NMF [24] was proposed by training NMF reconstruction basis that provides a satisfactory performance. Though these methods work well in specific sources separation, they need large training dataset and deal with a certain kind of source [1]. In our experiments, for small dataset containing different speakers, we use the exemplar-based NMF to learn the speech basis. The so-called exemplars are time-frequency slots of training data [25]. Exemplar-based NMF randomly select exemplars of the training data and is effective in small dataset.

With the pre-trained speech basis matrix $\bar{\mathbf{B}}_s \in \mathbb{R}_+^{F\times K_s}$, the speech signal separated by the online AuxIVA can be modeled as

$$|\mathbf{y}_{1,t}| = [\bar{\mathbf{B}}_s \; \mathbf{B}_n] \begin{bmatrix} \mathbf{h}_s \\ \mathbf{h}_n \end{bmatrix}, \qquad (19)$$

where $\mathbf{h}_s \in \mathbb{R}_+^{K_s\times 1}$ and $\mathbf{h}_n \in \mathbb{R}_+^{K_n\times 1}$ are the activation vectors for speech source and noise source, respectively, and $\mathbf{B}_n \in \mathbb{R}_+^{F\times K_n}$ is the basis matrix of noise source. K_s and K_n are user-guided basis numbers for speech and noise. Various criteria can be used to estimate the SSNMF model parameters. For the audio amplitude spectrum, the IS-divergence [26] is generally used as

$$f(\mathbf{B}_n, \mathbf{h}_s, \mathbf{h}_n) = \sum_f \frac{|y_{1,ft}|}{[\bar{\mathbf{B}}_s\mathbf{h}_s + \mathbf{B}_n\mathbf{h}_n]_f} + \log[\bar{\mathbf{B}}_s\mathbf{h}_s + \mathbf{B}_n\mathbf{h}_n]_f + cst, \qquad (20)$$

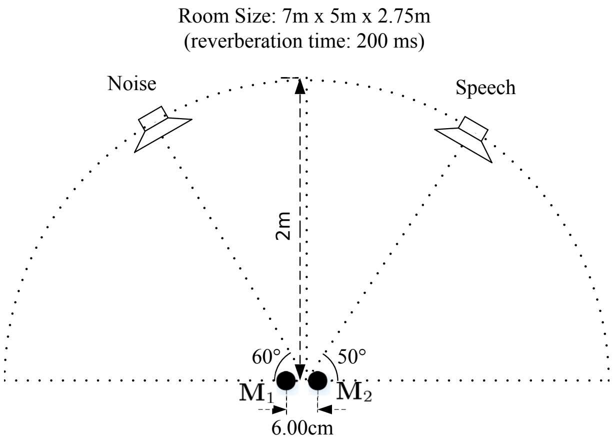


Figure 1: *Simulated room environment.*

where *cst* is a constant term w.r.t. parameters. To minimize the IS-divergence (20), majorization-minimization algorithm [27] can be used. The multiplicative update (MU) rules of SSNMF parameters read

$$\mathbf{B}_n \leftarrow \mathbf{B}_n \otimes \frac{\frac{|\mathbf{y}_{1,t}|}{\mathbf{Bh}}\mathbf{h}_n^\mathsf{T}}{\mathbf{1}^{F\times 1}\mathbf{h}_n^\mathsf{T}}, \qquad (21)$$

$$\begin{bmatrix} \mathbf{h}_s \\ \mathbf{h}_n \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{h}_s \\ \mathbf{h}_n \end{bmatrix} \otimes \frac{\mathbf{B}^\mathsf{T}\frac{|\mathbf{y}_{1,t}|}{\mathbf{Bh}}}{\mathbf{B}^\mathsf{T}\mathbf{1}^{I\times 1}}, \qquad (22)$$

where $\mathbf{B} = [\bar{\mathbf{B}}_s \; \mathbf{B}_n]$, $\mathbf{h} = [\mathbf{h}_s; \mathbf{h}_n]$, $\mathbf{1}^{F\times 1}$ is a all-one vector, $\otimes$ denotes entry-wise multiplication and the fraction bar denotes entry-wise division. To reconstruct the speech signal, we directly use the product of pre-trained basis matrix and activation vector

$$\hat{\mathbf{y}}_{1,t} = \bar{\mathbf{B}}_s\mathbf{h}_s. \qquad (23)$$

Compared with the AuxIVA in Section 2, (23) can model the speech variance with $F \times K_s$ parameters, which is larger than (14) with F parameters. Therefore, SSNMF can make the source model more flexible. Moreover, the prior information about the speech basis $\bar{\mathbf{B}}_s$, called structures of speech amplitude spectrum, can be used to estimated variance $\sigma_{1,t}^2$ more accurately. Therefore, the speech variance can be calculated as

$$\sigma_{1,t}^2 = \|\hat{\mathbf{y}}_{1,t}\|_2^2. \qquad (24)$$

Because it may be difficult to extract the feature of noise by NMF, we do not use SSNMF to estimate the variance of noise. Two methods can be used to obtain the variance of noise. Firstly, we can use (14) to estimate $\sigma_{2,t}^2$ as $\sigma_{2,t}^2 = \|\hat{\mathbf{y}}_{2,t}\|_2^2$. Secondly, the variance of noise can be calculated as

$$\sigma_{2,t}^2 = (\|\mathbf{y}_{1,t}\|_2^2 + \|\mathbf{y}_{2,t}\|_2^2) - r_{1,t}^2. \qquad (25)$$

We use (25) to calculate the variance of noise to provide a better estimation.

The proposed method is summarized as shown in Algorithm 1. Though the effective MU rules can be used to optimize the SSNMF parameters, the complexity of the proposed method is higher than that of the online AuxIVA.

5. Experimental evaluation

5.1. Setup

We conduct source separation experiments to confirm the performance of the proposed method, where the observed signals

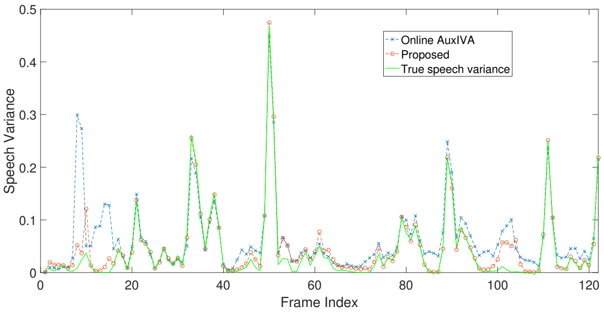


Figure 2: *The estimated speech variances.*

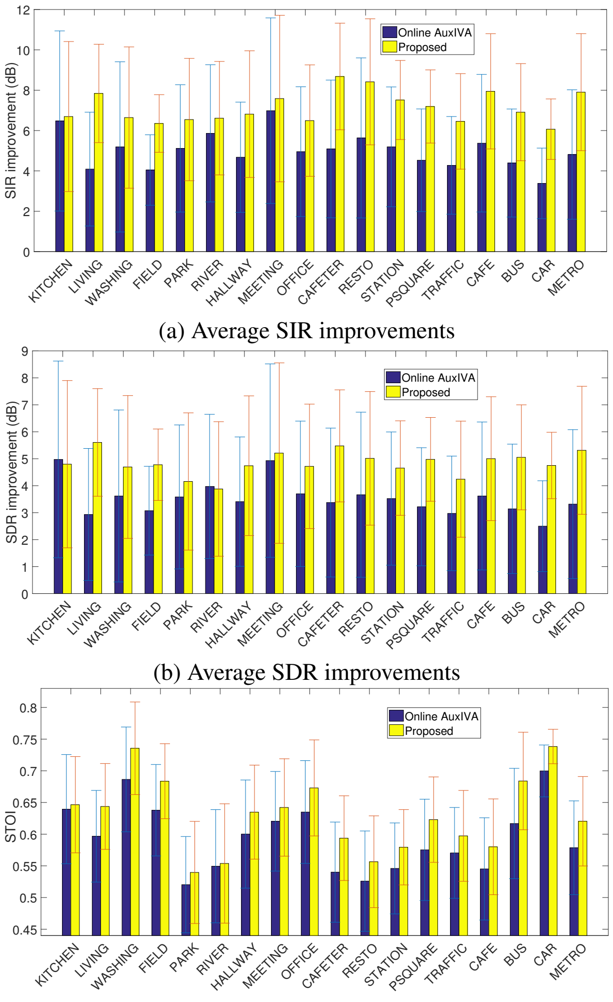
are mixtures of speech and different types of noise. The proposed method is compared with the online AuxIVA [6]. We use speech data selected from TIMIT [28] and 18 types of noise in the DEMAND database [29], and the sampling rate is 8 kHz. To simulate reverberant mixing system, the reverberant signals are produced by convoluting the room impulse response (RIR) with sources. The RIR is generated based on image method [30] and the recording condition is shown in Figure 1. The observed signals are synthesized by mixing the speech and noise at a signal-to-noise ratio of -5 dB. A 256-ms-long Hanning window is used for STFT with 75% overlap. Speech basis can be trained by different speakers in TIMIT dataset with $K_s = 80$. K_n is also set to 80 and α in (11) is set to 0.98. The iteration numbers in IVA and SSNMF are set to 2 and 15, respectively. Speech separation performance is evaluated using the PEASS toolkit [31], and speech intelligibility is quantified with the STOI [32] measure.

5.2. Comparison on estimation of speech variance

Figure 2 shows the estimated speech variances obtained by the online AuxIVA and the proposed algorithm. The results indicate that the proposed algorithm can work well in real time. Compared with the online AuxIVA, the proposed algorithm can obtain a more accurate estimation of speech variance. The speech signal $\mathbf{y}_{1,f}$ separated by the online AuxIVA may contain residual interference noise, leading to a larger variance estimation as shown in Figure 2. Therefore, the online AuxIVA cannot track the time-varying speech variance well. However, using pre-trained basis matrix containing speech features, the proposed algorithm can extract the speech signal and presents a good tracking performance of speech variance. Thanks to the more accurate speech variance, a higher separation performance can be achieved as shown later.

5.3. Comparison on speech denoising task

Figure 3(a) shows the SIR improvements on noisy speech mixed by speech and different types of noise. The proposed algorithm achieves higher averaged SIR improvements than the online AuxIVA. The averaged standard deviation of the SIR improvements of the proposed algorithm over all types of noise is 2.70, which is smaller than that of the online AuxIVA with 3.15. Figure 3(b) shows the SDR improvements. The averaged SDR improvements of the proposed method are higher than that of the online AuxIVA except for Kitchen and River noise. The obtained STOIs by the proposed algorithm and the online AuxIVA are shown in Figure 3(c). Similarly to the results on averaged SIR improvements, the proposed method can separate speech with better STOI measure in all types of noise. The results in Figure 3 show that the proposed algorithm achieves better and more stable performance than the online AuxIVA.



(a) Average SIR improvements

(b) Average SDR improvements

(c) Averaged STOIs

Figure 3: *Separation performance on test noisy speech with different types of noise, (a) average SIR improvements, (b) averaged SDR improvements, and (c) averaged STOIs.*

6. Conclusions

We have proposed a real-time AuxIVA method where the separated signal is modeled by SSNMF. We model the amplitude spectrum of separated signal from AuxIVA as the sum of amplitudes of source and residual interference noise. The source signal can be extracted in real time with pre-trained basis matrix which contains source structures. Instead of directly using the separated signal from AuxIVA, the proposed method uses the source signal to estimate the source variance. A more flexible source model can be provided by SSNMF. Experimental results demonstrated that the proposed algorithm can achieve more accurate estimation of speech variance and outperforms the online AuxIVA in terms of SDR, SIR and STOI.

7. Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 11804368.

8. References

[1] S. Makino, *Audio source separation*. Springer, 2018.

[2] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. International Conference on ICA and BSS*, 2000, pp. 215–220.

[3] T. Kim, H. T. Attias, S. Lee, and T. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[4] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[5] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 7, pp. 1431–1438, 2010.

[6] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, 2014, pp. 107–111.

[7] S. Erateb and J. Chambers, "Enhanced online IVA with switched source prior for speech separation," in *Proc. IEEE 11th Sensor Array and Multichannel Signal Processing Workshop*, 2020, pp. 1–5.

[8] S. Erateb, M. Naqvi, and J. Chambers, "Online IVA with adaptive learning for speech separation using various source priors," in *Proc. Sensor Signal Processing for Defence Conference*, 2017, pp. 1–5.

[9] J. Harris, B. Rivet, S. M. Naqvi, J. A. Chambers, and C. Jutten, "Real-time independent vector analysis with student's t source prior for convolutive speech mixtures," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 1856–1860.

[10] W. Rafique, S. Erateb, S. M. Naqvi, S. S. Dlay, and J. A. Chambers, "Independent vector analysis for source separation using an energy driven mixed student's t and super gaussian source prior," in *Proc. 24th European Signal Processing Conference*, 2016, pp. 858–862.

[11] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

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

[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] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[15] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[16] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[17] E. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation," in *Proc. the Annual Conference of the International Speech Communication Association*, 2013, pp. 808–812.

[18] Y. Jia, S. Kwong, J. Hou, and W. Wu, "Semi-supervised non-negative matrix factorization with dissimilarity and similarity regularization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2510–2521, 2020.

[19] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 189–192.

[20] Z. Liu and A. Dong, "A speech denoising algorithm based on harmonic regeneration," in *Proc. IOP Conference Series: Earth and Environmental Science*, 2019.

[21] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[22] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[23] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. 9th International Conference on Spoken Language Processing*, 2006.

[24] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative nmf and its application to single-channel source separation," in *Proc. 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 865–869.

[25] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[26] N. Hashimoto, S. Nakano, K. Yamamoto, and S. Nakagawa, "Speech recognition based on Itakura-Saito divergence and dynamics/sparseness constraints from mixed sound of speech and music by non-negative matrix factorization," in *Proc. 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 2749–2753.

[27] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[28] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[29] J. Thiemann, N. Ito, and E. Vincent, "Demand: a collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust*, 2013, pp. 1–6.

[30] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

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

[32] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.