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

Alpha-Stable Autoregressive Fast Multichannel Nonnegative Matrix Factorization for Joint Speech Enhancement and Dereverberation

Mathieu Fontaine[1], *Kouhei Sekiguchi*[1], *Aditya Arie Nugraha*[1], *Yoshiaki Bando*[2,1], *Kazuyoshi Yoshii*[3,1]

[1]AIP, RIKEN, Tokyo, Japan
[2]National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
[3]Graduate School of Informatics, Kyoto University, Kyoto, Japan

`{mathieu.fontaine,kouhei.sekiguchi,adityaarie.nugraha,yoshiaki.bando,kazuyoshi.yoshii}@riken.jp`

Abstract

This paper proposes α-stable autoregressive fast multichannel nonnegative matrix factorization (α-AR-FastMNMF), a robust joint blind speech enhancement and dereverberation method for improved automatic speech recognition in a realistic adverse environment. The state-of-the-art versatile blind source separation method called FastMNMF that assumes the short-time Fourier transform (STFT) coefficients of a direct sound to follow a circular complex Gaussian distribution with jointly-diagonalizable full-rank spatial covariance matrices was extended to AR-FastMNMF with an autoregressive reverberation model. Instead of the light-tailed Gaussian distribution, we use the heavy-tailed α-stable distribution, which also has the reproductive property useful for the additive source modeling, to better deal with the large dynamic range of the direct sound. The experimental results demonstrate that the proposed α-AR-FastMNMF works well as a front-end of an automatic speech recognition system. It outperforms α-AR-ILRMA, which is a special case of α-AR-FastMNMF, and their Gaussian counterparts, i.e., AR-FastMNMF and AR-ILRMA, in terms of the speech signal quality metrics and word error rate.

Index Terms: speech enhancement, dereverberation, automatic speech recognition, α-stable model, joint diagonalization

1. Introduction

The multichannel speech enhancement and dereverberation have become crucial with the new technologies of automatic speech recognition that aims to exploit the spatial correlation between microphones [1]. Joint blind source separation (BSS) and dereverberation have been investigated actively because the good adaptability to various acoustic environments is offered by its unsupervised nature.

A popular approach to multichannel BSS is the combination of a full rank spatial covariance matrix (SCM) model and a nonnegative matrix factorization (NMF) model on the power spectrogram of each source signal, resulting in multichannel nonnegative matrix factorization (MNMF) [2]. Its computationally-efficient constrained version called FastMNMF [3, 4] assumes the source SCMs to be full-rank and jointly diagonalizable [5, 6]. The well-known independent low-rank matrix analysis (ILRMA) [7] is moreover a special case of FastMNMF with a degenerate rank-1 source SCMs. FastMNMF was recently combined with an autoregressive (AR) model by making use of the weighted prediction error (WPE) [8] for joint BSS and dereverberation. This combination called AR-FastMNMF [9] achieves better results than its degenerate version called AR-ILRMA [10]. The local Gaussian model (LGM) [11] on the short-time Fourier transform (STFT) coefficients of each source signal, however, would be insufficient to extract a direct sound in a noisy reverberant environment and naturally calls for models that deal with more impulsive sounds.

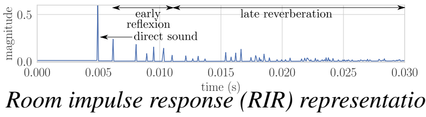

Figure 1: *Room impulse response (RIR) representation. The direct sound can be seen as an outlier realization of all reflections.*

Heavy-tailed extensions of MNMF [12–14], ILRMA [15], and FastMNMF [16, 17] have been proposed in the context of reverberant source image separation. Unlike the late reverberation, the direct sound and early reflection tend to be outliers. In [18], the room impulse response (RIR) and sources are statistically represented as a heavy-tailed Student's t distribution (Fig. 1). However, the Student's t law is usually not preserved under linear combination, whereas the sum of α-stable random vector gives an α-stable random vector [19]; the α-stable distribution satisfies the reproductive property (RP) [20]. Besides the RP, the α-stable model has capability of adjusting the tail heaviness with a characteristic exponent $\alpha \in (0, 2]$, including well-known distributions such as Levy ($\alpha = 0.5$), Cauchy ($\alpha = 1$), and Gaussian ($\alpha = 2$) distributions as its special cases. The multichannel α-stable model is usually restricted to the so-called elliptically-contoured α-stable distribution [13] that admits a positive definite shape matrix akin to the covariance matrix of a Gaussian distribution, yielding a robust extension to FastMNMF called α-FastMNMF [17]. Even if the RP is satisfied, the probability density function (PDF) of an α-stable random vector is usually not computable. Instead, the α-stable distribution can be seen as a Gaussian one, where the covariance is perturbed by a positive random variable called an impulse variable [21].

In this paper, we propose a robust and adaptive extension of AR-FastMNMF called α-AR-FastMNMF based on the α-stable model on the direct sound for joint BSS and dereverberation. Our method enforces the sparsity of the direct sound represented by the full-rank SCM model and satisfies the RP for the additive source model assumption. The sparsity is controlled by a time-dependent impulse variable [13] intended to more precisely identify the direct sound in a noisy reverberant mixture. Inspired by [22], adaptive estimation of the characteristic exponent α is furthermore proposed. In this context, the value of α also aims to evaluate the impulsiveness of the direct sound in the noisy reverberant observed signal.

The remainder of the paper is organized as follows: Section 2 reviews the conventional AR-FastMNMF. Section 3 describes the proposed α-AR-FastMNMF and derives its parameter estimation algorithm. Section 4 evaluates α-AR-FastMNMF in the context of noisy and reverberant speech recognition. The conclusion is finally drawn in Section 5.

http://dx.doi.org/10.21437/Interspeech.2021-742

2. Autoregressive FastMNMF

This section reviews AR-FastMNMF [9] that integrates an autoregressive model into FastMNMF2 [4].

2.1. Mixture, Source, and Spatial Models

We suppose that N sources are recorded by M microphones. Let $\mathbf{X} \triangleq \{\mathbf{x}_{ft}\}_{f,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be the tensor of a multichannel reverberant observed signal in the short-time Fourier transform (STFT) domain for all time-frequency (TF) bins, where F is the number of frequency bins and T is that of time frames. We consider an autoregressive (AR) model that decomposes $\mathbf{x}_{ft}$ into the direct sound $\mathbf{d}_{ft}$ and the reverberation $\mathbf{r}_{ft}$ as follows:

$$\mathbf{x}_{ft} = \mathbf{d}_{ft} + \mathbf{r}_{ft} \triangleq \underbrace{\sum_{n=1}^{N} \mathbf{d}_{nft}}_{\text{direct sound}} + \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{B}_{fl}\mathbf{x}_{f,t-l}}_{\text{reverberation}}, \tag{1}$$

where $\mathbf{B}_{fl} \triangleq [\mathbf{b}_{fl1}, \dots, \mathbf{b}_{flM}]^{\mathsf{T}} \in \mathbb{C}^{M\times M}$ are the coefficients of the AR process of order $L \geq 0$ called the *tap length*, $\Delta \geq 0$ is the *delay* and ${}^{\mathsf{T}}$ denotes the transposition. From the local Gaussian model, the direct sounds $\{\mathbf{d}_{nft}\}_{f,t=1}^{F,T}$ are assumed to be independent circular complex Gaussian distributions [11]:

$$\mathbf{d}_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(\lambda_{nft}\mathbf{G}_{nf} \triangleq \mathbf{Y}_{nft}\right), \tag{2}$$

where $\lambda_{nft} \geq 0$ represents the power spectral density (PSD) of the source s_{nft} and $\mathbf{G}_{nf}$ is the spatial covariance matrix. The source PSDs $\{\lambda_{nft}\}_{n,f,t=1}^{N,F,T}$ are represented by nonnegative matrix factorization (NMF) parameterized by $\mathbf{W} \triangleq \{w_{nkf}\}_{n,k,f=1}^{N,K,F}$, $\mathbf{H} \triangleq \{h_{nkt}\}_{n,k,t=1}^{N,K,T}$ as follows:

$$\lambda_{nft} = \sum_{k=1}^{K} w_{nkf}h_{nkt}, \tag{3}$$

where K is the total number of NMF bases, $w_{nfk} \geq 0$ and $h_{nkt} \geq 0$ are the magnitude of basis k of source n at frequency f and the activation of basis k of source n at time t, respectively. Moreover, we consider that $\mathbf{G}_{nf}$'s are jointly-diagonalizable full-rank spatial matrices with the decomposition:

$$\forall n, f,\ \mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_n)\,\mathbf{Q}_f^{-\mathsf{H}}, \tag{4}$$

where $\mathrm{Diag}(\tilde{\mathbf{g}}_n)$ is a diagonal matrix whose diagonal elements are $\tilde{\mathbf{g}}_n \triangleq [\tilde{g}_{n1}, \dots, \tilde{g}_{nM}] \in \mathbb{R}_+^M$, $\mathbf{Q}_f \triangleq [\mathbf{q}_{f1}, \dots, \mathbf{q}_{fM}]^{\mathsf{H}} \in \mathbb{C}^{M\times M}$ is the *diagonalizer* assumed to be non-singular, and ${}^{\mathsf{H}}$ denotes the conjugate transpose. Using the reproductive property (RP) of Gaussian distributions and combining Eqs. (2) and (4) provide the following mixture model:

$$\mathbf{d}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{Q}_f^{-1}\left[\sum_{n=1}^{N}\lambda_{nft}\mathrm{Diag}(\tilde{\mathbf{g}}_n)\right]\mathbf{Q}_f^{-\mathsf{H}} \triangleq \mathbf{Y}_{ft}\right). \tag{5}$$

Such a model with the AR filter is called AR-FastMNMF [9] and includes AR-ILRMA [10] as a special case with $\tilde{\mathbf{G}} \triangleq [\tilde{\mathbf{g}}_1, \cdots, \tilde{\mathbf{g}}_N]^{\mathsf{T}}$ equal to the identity, $M = N$ and the column vectors of $\mathbf{Q}_f^{-1}$ acting like the column of the mixing matrix. The column vectors of $\mathbf{Q}_f^{-1}$ then serve as steering vectors pointing in a set of M directions with a weight-shared scale given by the entries of the nonnegative matrix $\tilde{\mathbf{G}}$.

2.2. Parameter Estimation and Filtering Methods

The whole parameters $\mathbf{\Theta} \triangleq \{\mathbf{W}, \mathbf{H}, \tilde{\mathbf{G}}, \mathbf{Q}, \mathbf{B}\}$ with $\mathbf{Q} \triangleq \{\mathbf{Q}_f\}_{f=1}^{F}$ and $\mathbf{B} \triangleq \{\mathbf{B}_{fl}\}_{f,l=1}^{F,L}$ are estimated jointly based on the maximization of the log-likelihood (LL) $\log p(\mathbf{X}|\mathbf{\Theta})$, where p is the probability density function (PDF) of a distribution. The observations $\mathbf{X}$ and the estimated parameters $\mathbf{\Theta}$ are then used to apply a Wiener filtering that provides an estimator of the direct sounds $\{\mathbf{d}_{nft}\}_{f,t=1}^{F,T}$. The readers is referred to [4, 9] for further implementation details.

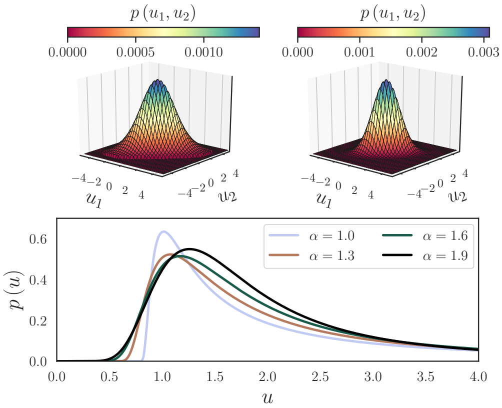

Figure 2: *PDFs of bivariate elliptically stable distributions for* $\alpha = 1.6$ *(top-left) and* $\alpha = 2$ *(i.e., the Gaussian distribution; top-right), and positive* α*-stable distributions for various* α *(bottom).*

3. Alpha-Stable Autoregressive FastMNMF

This section introduces the α-stable theory and the proposed α-stable autoregressive FastMNMF (α-AR-FastMNMF).

3.1. α-stable Theory

A multivariate random vector $\mathbf{u}$ that can be decomposed as the sum of two independent identically distributed (i.i.d) vectors with the same law as $\mathbf{u}$ is called α-stable [19]. The characteristic exponent $\alpha \in (0, 2]$ controls the heaviness of the tails with special instances such as Levy ($\alpha = 0.5$), Cauchy ($\alpha = 1$), and Gaussian ($\alpha = 2$) distributions. Assuming the complex-valued *isotropic symmetric* multivariate α-stable model [19], we have shown that the underlying vectors can be recovered [23]. However, due to incompatible representation, incorporating the computationally-efficient joint diagonalization (as in FastMNMF) is not possible.

Instead, we use in this paper the complex-valued *elliptically-contoured symmetric* multivariate α-stable distribution [12, 13], shorted in *elliptically stable distribution* hereafter (Fig. 2), to integrate the joint diagonalization. A zero-location elliptically stable distribution $\mathbf{u}$ can be represented using a positive-definite shape matrix $\mathbf{R} \succ 0$ as in the Gaussian case:

$$\mathbf{u} \sim \mathcal{S}_{\mathbb{C}}^{\alpha}(\mathbf{0}, \mathbf{R}). \tag{6}$$

The elliptically stable family holds the RP for a fixed α. However, the shape matrix of the sum is not the sum of shape matrices [24]. This non-linearity of shape matrices can be resolved by representing an elliptically stable distribution as a Gaussian scale mixture [25], where $\mathbf{u}$ is described as a conditional Gaussian distribution $\mathbf{u}|\phi \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \phi\mathbf{R})$ whose covariance is perturbed by a so-called impulse variable [21, 26] following a positive α-stable distribution (Fig. 2) [19], $\phi \sim \mathcal{PS}^{\alpha}\left(2\cos(\frac{\pi\alpha}{4})^{2/\alpha}\right)$.

3.2. α-AR-FastMNMF Model and Filtering Method

The first impulse in a room impulse response (RIR, Fig. 1), *i.e.*, the direct sound, can be regarded as an outlier compare to the rest consisting of early reflections and late reverberations. Combined with the additive source model, it naturally calls for an elliptically stable model, whose impulse variable can enforce the sparseness of the direct sound compared to the reverberant observation in the time domain and thus, it can be assumed to be source- and time-dependent. For each source n and TF-bin (f, t), we consider the following conditional Gaussian model to represent the elliptically stable model in Eq. (6):

$$\begin{cases} \mathbf{d}_{nft} \mid \phi_{nt} & \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \phi_{nt} \mathbf{Y}_{nft} \right), \\ \phi_{nt} & \sim \mathcal{PS}^{\alpha} \left(2 \cos \left(\frac{\pi\alpha}{4} \right)^{2/\alpha} \right), \end{cases} \tag{7}$$

where $\mathbf{Y}_{nft}$ is defined in Eq. (2). Because $\mathbf{x}_{ft}|\mathbf{\Phi}$ with $\mathbf{\Phi} \triangleq \{\phi_{nt}\}_{n,t=1}^{N,T}$ is also a conditional Gaussian model, we estimate the direct sounds $\{\mathbf{d}_{nft}\}_{f,t=1}^{F,T}$ using [9, 17]:

$$\begin{aligned} &\mathbb{E}_{\mathbf{\Phi}|\mathbf{X}} \left[\mathbb{E} \left[\mathbf{d}_{nft} \mid \mathbf{\Theta}, \mathbf{\Phi}, \mathbf{x}_{ft}, \alpha \right] \right] \\ &= \mathbb{E}_{\mathbf{\Phi}|\mathbf{X}} \left[\phi_{nt} \mathbf{Y}_{nft} \left(\sum_{n'=1}^{N} \phi_{n't} \mathbf{Y}_{n'ft} \right)^{-1} \right] \mathbf{d}_{ft}. \end{aligned} \tag{8}$$

3.3. Parameter Estimation

We first describe the estimation of the characteristic exponent $\alpha \in (0, 2]$ in Section 3.3.1. Sections 3.3.2 and 3.3.3 then formulate the estimations of $\mathbf{\Theta}$ and $\mathbf{\Phi}$, respectively, following the expectation-maximization (EM) framework [27].

3.3.1. Estimation of α

We propose the estimation of α for the complex-valued multivariate isotropic α-stable distribution by generalizing the method for the real-valued counterpart proposed in [22] since all assumptions still hold when all entries of $\mathbf{X}$ are assumed to be i.i.d. with the same α. Given the mixture $\mathbf{X}$, we create non-overlapping segments consisting of T' frames ($T' \leq T$) and consider each segment as a set $B \triangleq \{\mathbf{x}_{ft}\}_{f,t=1}^{F,T'} \triangleq \{\mathbf{x}_b\}_{b=1}^{FT'}$, in which the shape matrices $\mathbf{R}_{ft}$ are assumed to be i.i.d. We randomly split the set into B_2 minibatches of B_1 samples with $|B| = B_1 B_2 = FT'$ being the total number of samples. The values B_1 and B_2 are arbitrary as long as $1 < B_1 < B_2$ [22]. The estimated α for set B, denoted by $\widehat{\alpha}_B$, can be computed as

$$\frac{1}{\widehat{\alpha}_B} \triangleq \frac{1}{\log B_1} \left(\frac{1}{B_2} \sum_{b'=1}^{B_2} \log \|\boldsymbol{\xi}_{b'}\| - \frac{1}{|B|} \sum_{b=1}^{|B|} \log \|\mathbf{x}_b\| \right), \tag{9}$$

where $\boldsymbol{\xi}_{b'} \triangleq \sum_{b''=1}^{B_1} \mathbf{x}_{b''+(b'-1)B_1}$ and $\|.\|$ the Frobenius norm. Assuming that α is the same for the whole mixture, we then obtain $\widehat{\alpha}$ by averaging all $\widehat{\alpha}_B$ from the different segments.

3.3.2. Estimation of $\mathbf{\Theta}$

We maximize the discretization of the LL function $\log\, p \left(\mathbf{X} | \Theta, \alpha \right) = \log \int p(\mathbf{X}|\mathbf{\Theta}, \boldsymbol{\phi}, \alpha) p(\boldsymbol{\phi}) d\boldsymbol{\phi}$ [9, 17]:

$$\begin{aligned} \log\, p(\mathbf{X}|\mathbf{\Theta}) \geq &-\frac{1}{P} \sum_{f,t,m=1}^{F,T,M} \sum_{p=1}^{P} \left(\frac{\tilde{d}_{ftm}}{\tilde{y}_{ftmp}} + \log \tilde{y}_{ftmp} \right) \\ &+ T \sum_{f=1}^{F} \log \left| \mathbf{Q}_f \mathbf{Q}_f^{\mathsf{H}} \right| - \mathrm{KL}[q(\phi_{nt}) \| p(\phi_{nt})], \end{aligned} \tag{10}$$

where P is the number of samples, KL is the Kullback-Leibler divergence, $q(\phi)$ a variational distribution that satisfies the equality in (10) for all n, t if $q(\phi_{nt}) = p \left(\phi_{nt} | \mathbf{X} \right)$, $\tilde{d}_{ftm} \triangleq |\mathbf{q}_{fm}^{\mathsf{H}} \mathbf{d}_{ft}|^2$ and $\tilde{y}_{ftmp} = \sum_{n,k=1}^{N,K} \tilde{\phi}_{ntp} w_{nkf} h_{nkt} \tilde{g}_{nm}$, where $\tilde{\phi}_{ntp} \sim p \left(\phi_{nt} | \mathbf{X} \right)$. The multiplicative updates rules for $\mathbf{W}$, $\mathbf{H}$, $\tilde{\mathbf{G}}$ are given by [9, 17]:

$$w_{nkf} \leftarrow w_{nkf} \sqrt{\frac{\sum_{t,m,p=1}^{T,M,P} \tilde{\phi}_{ntp} h_{nkt} \tilde{g}_{nm} \tilde{d}_{ftm} \tilde{y}_{ftmp}^{-2}}{\sum_{t,m,p=1}^{T,M,P} \tilde{\phi}_{ntp} h_{nkt} \tilde{g}_{nm} \tilde{y}_{ftmp}^{-1}}}, \tag{11}$$

$$h_{nkt} \leftarrow h_{nkt} \sqrt{\frac{\sum_{f,m,p=1}^{F,M,P} \tilde{\phi}_{ntp} w_{nkf} \tilde{g}_{nm} \tilde{d}_{ftm} \tilde{y}_{ftmp}^{-2}}{\sum_{f,m,p=1}^{F,M,P} \tilde{\phi}_{ntp} w_{nkf} \tilde{g}_{nm} \tilde{y}_{ftmp}^{-1}}}, \tag{12}$$

$$\tilde{g}_{nm} \leftarrow \tilde{g}_{nm} \sqrt{\frac{\sum_{f,t,k,p=1}^{F,T,K,P} \tilde{\phi}_{ntp} w_{nkf} h_{nkt} \tilde{d}_{ftm} \tilde{y}_{ftmp}^{-2}}{\sum_{f,t,k,p=1}^{F,T,K,P} \tilde{\phi}_{ntp} w_{nkf} h_{nkt} \tilde{y}_{ftmp}^{-1}}}. \tag{13}$$

The vectors $\mathbf{q}_{fm}$ of $\mathbf{Q}$ are updated as in [28] given the matrix $\mathbf{V}_{fm} \triangleq \frac{1}{TP} \sum_{t,p=1}^{T,P} \mathbf{d}_{ft} \mathbf{d}_{ft}^{\mathsf{H}} \tilde{y}_{ftmp}^{-1}$ as follows:

$$\mathbf{q}_{fm} \leftarrow (\mathbf{Q}_f \mathbf{V}_{fm})^{-1} \mathbf{e}_m, \tag{14}$$

$$\mathbf{q}_{fm} \leftarrow (\mathbf{q}_{fm}^{\mathsf{H}} \mathbf{V}_{fm} \mathbf{q}_{fm})^{-\frac{1}{2}} \mathbf{q}_{fm}. \tag{15}$$

The updating rules for $\mathbf{B}$ are similar to the ones in [9, 29] :

$$\mathbf{\Psi}_f \triangleq \sum_{m=1}^{M} \mathbf{q}_{fm} \otimes \left(\frac{1}{P} \sum_{t,p=1}^{T,P} \frac{\mathbf{x}_{ft}^{\mathsf{H}} \mathbf{q}_{fm}}{\tilde{y}_{ftmp}} \bar{\mathbf{x}}_{ft} \right)^{*}, \tag{16}$$

$$\mathbf{\Omega}_f \triangleq \sum_{m=1}^{M} (\mathbf{q}_{fm} \mathbf{q}_{fm}^{\mathsf{H}}) \otimes \left(\frac{1}{P} \sum_{t,p=1}^{T,P} \frac{\bar{\mathbf{x}}_{ft} \bar{\mathbf{x}}_{ft}^{\mathsf{H}}}{\tilde{y}_{ftmp}} \right)^{\mathsf{T}}, \tag{17}$$

$$\hat{\mathbf{b}}_f = \mathbf{\Omega}_f^{-1} \mathbf{\Psi}_f, \tag{18}$$

where * stands for the complex conjugate, $\otimes$ stands for the Kronecker product and

$$\hat{\mathbf{b}}_f \triangleq [\mathbf{b}_{f:1}^{\mathsf{T}}, \ldots, \mathbf{b}_{f:M}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{M^2 L}, \tag{19}$$

$$\mathbf{b}_{f:m} \triangleq [\mathbf{b}_{f,\Delta,m}^{\mathsf{T}}, \ldots, \mathbf{b}_{f,\Delta+L-1,m}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{ML}, \tag{20}$$

$$\bar{\mathbf{x}}_{ft} \triangleq [\mathbf{x}_{f,t-\Delta}^{\mathsf{T}}, \ldots, \mathbf{x}_{f,t-(\Delta+L-1)}^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{C}^{ML}. \tag{21}$$

3.3.3. Estimation of $\mathbf{\Phi}$

Since $\mathbf{\Phi}$ is not tractable because $p \left(\phi_{nt} | \mathbf{X} \right)$ does not admit a closed-form, we approximate the PDF by iteratively drawing samples using the Metropolis-Hastings (MH) algorithm [13]:

1. Draw samples from $\phi_{nt,f}^{\text{new}} \sim \mathcal{PS}^{\alpha} \left(2 \cos \left(\frac{\pi\alpha}{4} \right)^{2/\alpha} \right)$.
2. Sample $\nu \sim \mathcal{U} \left([0, 1] \right)$ from the uniform distribution.
3. Compute the acceptance probability for all f:

   $$\mathrm{acc} \left(\phi_{nt,f}^{\text{old}} \rightarrow \phi_{nt,f}^{\text{new}} \right) = \min \left(1, \frac{u_{nft} \left(\phi_{nt,f}^{\text{new}} \right)}{u_{nft} \left(\phi_{nt,f}^{\text{old}} \right)} \right), \tag{22}$$

   where $u_{nft} \left(\phi_{nt,f} \right)$ is the PDF of a zero-mean Gaussian distribution whose covariance matrix is given by $\phi_{nt,f} \lambda_{nft} \mathrm{Diag} \left(\tilde{\mathbf{g}}_n \right) + \sum_{n' \neq n} \phi_{n't,f} \lambda_{n'ft} \mathrm{Diag} \left(\tilde{\mathbf{g}}_{n'} \right)$.
4. Acceptance test:
   - if $\nu < \mathrm{acc} \left(\phi_{nt,f}^{\text{old}} \rightarrow \phi_{nt,f}^{\text{new}} \right)$, $\phi_{nt,f} = \phi_{nt,f}^{\text{new}}$ (accept);
   - otherwise, $\phi_{nt,f} = \phi_{nt,f}^{\text{old}}$ (reject).
5. Average by computing $\tilde{\phi}_{ntp} \triangleq \frac{1}{F} \sum_{f=1}^{F} \phi_{nt,f}$.

4. Evaluation

We assess the performance of the proposed α-AR-FastMNMF and α-AR-ILRMA in the context of speech enhancement and dereverberation for automatic speech recognition (ASR). The enhanced speech signal is evaluated in terms of the signal-to-distortion ratio (SDR) [30] and the perceptual evaluation speech

Table 1: *Average SDR/PESQ/WER scores for experiments described in Section 4. Boldface numbers show the best performance.*

Dist.	SNR	M	L	α-AR-FastMNMF variants		α-AR-ILRMA variants		Unprocessed
				$\alpha\in(0,2]$	$\alpha=2$	$\alpha\in(0,2]$	$\alpha=2$	
Near	0 dB	2	0	**5.4/1.6/49.1**	4.8/**1.6**/50.4	2.5/**1.6**/51.1	2.6/ 1.5/ 49.2	−0.3/1.4/68.0
			4	**7.0/1.6/34.4**	6.2/**1.6**/35.5	3.9/**1.6**/36.0	4.2/**1.6**/ 36.4	
		8	0	**12.1/2.3**/ 12.2	11.3/ 2.1/13.3	8.9/ 1.9/11.3	8.8/ 2.0/**10.2**	
			4	**13.7/2.4**/ **8.6**	12.8/ 2.2/ 9.7	10.9/ 2.0/ 8.9	10.8/ 2.0/ 9.6	
	5 dB	2	0	**9.3/1.8/11.1**	8.9/**1.8**/12.9	7.1/**1.8**/13.4	7.2/**1.8**/ 14.8	4.1/1.5/65.9
			4	**10.9/1.9/10.2**	10.2/**1.9**/11.7	8.6/ 1.8/11.4	8.6/ 1.8/ 11.1	
		8	0	**13.9/2.6**/ 7.2	13.2/**2.6**/ 8.1	11.9/ 2.3/ **4.9**	12.2/ 2.3/ 5.3	
			4	**16.2/2.7**/ **4.0**	15.1/ 2.6/ 5.8	14.2/ 2.4/ 4.4	14.0/ 2.4/ 4.7	
Far	0 dB	2	0	**2.7/1.4/44.3**	1.8/**1.4**/45.7	0.5/**1.4**/45.9	0.7/**1.4**/ 47.9	−1.3/1.3/68.4
			4	**5.1/1.5/40.4**	4.2/**1.5**/42.4	2.3/**1.5**/41.3	2.4/**1.5**/ 42.3	
		8	0	**7.7/1.7/19.6**	7.1/**1.7**/21.0	5.1/**1.7**/22.7	5.5/**1.7**/ 24.1	
			4	**10.8/1.9/12.2**	10.0/**1.9**/16.0	8.4/ 1.8/15.3	8.6/ 1.8/ 14.2	
	5 dB	2	0	**5.4/1.6**/ 27.4	4.8/**1.6**/27.9	4.1/**1.6**/28.3	4.1/**1.6/26.3**	2.1/1.5/66.5
			4	**8.1/1.7/19.4**	7.6/ 1.6/21.2	6.2/ 1.6/20.1	6.3/ 1.6/ 19.7	
		8	0	**9.7/2.0**/ 8.6	9.0/**2.0**/ 9.3	7.6/ 1.9/ **8.0**	8.0/ 1.9/ 8.3	
			4	**13.4/2.2**/ **4.9**	12.5/**2.2**/ 7.6	11.9/ 2.0/ 6.5	11.7/ 2.0/ 6.8	

quality (PESQ) [31], while the speech transcription is evaluated in terms of word error rate (WER). A higher score is better for SDR and PESQ; a lower score is better for WER. The methods discussed in this paper act as a frontend of a transformer-based ASR [32], as implemented in SpeechBrain [33].

4.1. Data and Settings

We consider 20 utterances from the REVERB challenge dataset [34] composed of 8-channel mixtures ($M=8$) sampled at 16kHz with a distance of 2.0 m (far) or 0.5 m (near) between the center of microphone array and the speaker. The reverberant time (RT$_{60}$) is either 0.25, 0.5, or 0.7 s, while the signal-to-noise ratio is either 0 or 5 dB. The STFT coefficients are computed using a 1024-point Hann window with 75% overlap ($F=513$ frequency bins).

The compared methods include α-AR-FastMNMF, α-AR-ILRMA, and their Gaussian variants ($\alpha = 2$) [4, 10]. Those AR variants use a tap length $L = 4$ and a delay $\Delta = 3$. We also consider the corresponding non-AR variants [4, 7, 17] by setting $L = 0$. For all methods, the number of NMF bases is $K=16$ and that of microphones is either $M=2$ or $M=8$. For a fair comparison, we only perform the determined separation ($N=M$), where the channel with the highest energy is selected as the enhanced speech signal afterwards. All methods initialize $\mathbf{\Theta}$ using AR-FastMNMF run for 50 iterations, and then perform parameter optimization for 150 iterations. The parameters α of the α-stable variants are estimated using segments with $T'=100$ (see Section 3.3.1) and MH sampling for 40 iterations, including a burning period of 30 iterations, so $P=10$.

4.2. Results and Discussion

Fig. 3 illustrates spectrograms obtained by α-AR-FastMNMF with different α, where α_{opt} is estimated using the proposed method in Section 3.3.1. For $\alpha = 1.5$, we can see that some harmonics of the phoneme are vanished while the Gaussian case ($\alpha = 2$) is noisier on the silent part than α_{opt}. Table 1 shows that α-AR-FastMNMF achieves the best SDR, PESQ, and WER scores for almost all settings. For non-AR methods ($L=0$), ILRMA achieves the best WER scores in some cases. If we compare the ILRMA variants for $L\in\{0,4\}$, the α-stable version tends to achieve similar results as Gaussian ILRMA. ILRMA and FastMNMF gap scores between α and Gaussian extensions shed the light on the necessity to consider a full-rank model for the α-stable version.

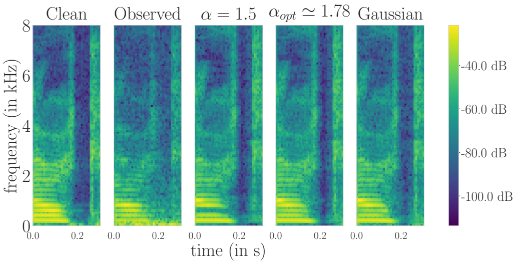

Figure 3: *Spectrograms of the clean speech reference* `c30c0201`*, the observed signal (SNR=*5 *dB, RT*$_{60}$=0.5 *s, far), and the enhanced speech obtained by* α*-AR-FastMNMF (*$M=8$*,* $K=16$*) with different* α*.*

In summary, we first pointed out that the estimated value α by the method of Section 3.3.1 is essential to have a good reconstruction of the speech spectrograms. We demonstrate that the combination of α-stable model with an autoregressive FastMNMF model significantly improves the performances of speech enhancement, dereverberation and ASR. The results of ILRMA methods also show that a full-rank SCM model is relevant for the α-stable variants to outperform their Gaussian counterpart.

5. Conclusion

This paper described a probabilistic integration of an autoregressive model with the FastMNMF framework, where the original elliptically-contoured complex Gaussian model is replaced by an α-stable one to form the proposed α-AR-FastMNMF. By doing so, we enforce the sparsity between the direct sound and the reverberant component using the impulse variable induced by the elliptically-contoured α-stable model. The proposed method achieves promising results in terms of source separation and speech recognition scores. Future directions may include characteristic exponent α depending on time or replacing the NMF decomposition of the speech by a heavy-tailed deep speech prior model on the spectrograms as in [14].

6. Acknowledgements

Supported by JSPS KAKENHI Nos. 19H04137, 20K19833, and 20H01159; and NII CRIS-Line Collaborative Research.

7. References

[1] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[2] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[3] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 371–375.

[4] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE Trans. ASLP*, vol. 28, pp. 2610–2625, 2020.

[5] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[6] N. Ito, S. Araki, and T. Nakatani, "FastFCA: A joint diagonalization based fast algorithm for audio source separation using a full-rank spatial covariance model," in *Proc. Eur. Signal Process. Conf.*, 2018, pp. 1667–1671.

[7] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.

[9] K. Sekiguchi, Y. Bando, A. A. Nugraha, M. Fontaine, and K. Yoshii, "Autoregressive fast multichannel nonnegative matrix factorization for joint blind source separation and dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 511–515.

[10] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2018, pp. 31–35.

[11] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE/ACM Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[12] M. Fontaine, F.-R. Stöter, A. Liutkus, U. Şimşekli, R. Serizel, and R. Badeau, "Multichannel audio modeling with elliptically stable tensor decomposition," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2018, pp. 13–23.

[13] S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, and G. Richard, "Alpha-stable multichannel audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 576–580.

[14] M. Fontaine, A. A. Nugraha, R. Badeau, K. Yoshii, and A. Liutkus, "Cauchy multichannel speech enhancement with a deep speech prior," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[15] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 1, p. 28, 2018.

[16] K. Kamo, Y. Kubo, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Joint-diagonalizability-constrained multichannel nonnegative matrix factorization based on multivariate complex student's t-distribution," in *APSIPA*. IEEE, 2020, pp. 869–874.

[17] M. Fontaine, K. Sekiguchi, A. A. Nugraha, and K. Yoshii, "Unsupervised robust speech enhancement based on alpha-stable fast multichannel nonnegative matrix factorization," in *Proc. Interspeech*, 2020, pp. 4541–4545.

[18] S. Leglaive, R. Badeau, and G. Richard, "Student's t source and mixing models for multichannel audio source separation," *IEEE Trans. ASLP*, vol. 26, no. 6, pp. 1154–1168, 2018.

[19] G. Samoradnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall/CRC, 1994.

[20] W. Feller, *An introduction to probability theory and its applications, vol 2*. John Wiley & Sons, 2008.

[21] U. Şimşekli, A. Liutkus, and A. T. Cemgil, "Alpha-stable matrix factorization," *IEEE Signal Process. Letters.*, vol. 22, no. 12, pp. 2289–2293, 2015.

[22] M. Mohammadi, A. Mohammadpour, and H. Ogata, "On estimating the tail index and the spectral measure of multivariate α-stable distributions," *Metrika*, vol. 78, no. 5, pp. 549–561, 2015.

[23] M. Fontaine, R. Badeau, and A. Liutkus, "Separation of alpha-stable random vectors," *Signal Process.*, vol. 170, p. 107465, 2020.

[24] J. P. Nolan, "Multivariate elliptically contoured stable distributions: theory and estimation," *Comput. Stat.*, vol. 28, no. 5, pp. 2067–2089, 2013.

[25] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. Roy. Statist. Soc. Ser. B*, vol. 36, no. 1, pp. 99–102, 1974.

[26] U. Şimşekli, H. Erdoğan, S. Leglaive, A. Liutkus, R. Badeau, and G. Richard, "Alpha-stable low-rank plus residual decomposition for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 651–655.

[27] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[28] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 189–192.

[29] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE Trans. ASLP*, vol. 28, pp. 2267–2282, 2020.

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

[31] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Recommendation P.862, 2001.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

[33] M. Ravanelli, T. Parcollet, A. Rouhe, P. Plantinga, E. Rastorgueva, L. Lugosch, N. Dawalatabad, C. Ju-Chieh, A. Heba, F. Grondin, W. Aris, C.-F. Liao, S. Cornell, S.-L. Yeh, H. Na, Y. Gao, S.-W. Fu, C. Subakan, R. De Mori, and Y. Bengio, "Speechbrain," https://github.com/speechbrain/speechbrain, 2021.

[34] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2013, pp. 1–4.