*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Unsupervised Robust Speech Enhancement Based on Alpha-Stable Fast Multichannel Nonnegative Matrix Factorization

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

[1]AIP, RIKEN, Tokyo, Japan
[2]Graduate School of Informatics, Kyoto University, Kyoto, Japan
{mathieu.fontaine, kouhei.sekiguchi, adityaarie.nugraha, kazuyoshi.yoshii}@riken.jp

## Abstract

This paper describes multichannel speech enhancement based on a probabilistic model of complex source spectrograms for improving the intelligibility of speech corrupted by undesired noise. The univariate complex Gaussian model with the reproductive property supports the additivity of source complex spectrograms and forms the theoretical basis of nonnegative matrix factorization (NMF). Multichannel NMF (MNMF) is an extension of NMF based on the multivariate complex Gaussian model with spatial covariance matrices (SCMs), and its state-of-the-art variant called FastMNMF with jointly-diagonalizable SCMs achieves faster decomposition based on the univariate Gaussian model in the transformed domain where all time-frequency-channel elements are independent. Although a heavy-tailed extension of FastMNMF has been proposed to improve the robustness against impulsive noise, the source additivity has never been considered. The multivariate $\alpha$-stable distribution does not have the reproductive property for the shape matrix parameter. This paper, therefore, proposes a heavy-tailed extension called $\alpha$-stable FastMNMF which works in the transformed domain to use a univariate complex $\alpha$-stable model, satisfying the reproductive property for any tail lightness parameter $\alpha$ and allowing the $\alpha$-fractional Wiener filtering based on the element-wise source additivity. The experimental results show that $\alpha$-stable FastMNMF with $\alpha = 1.8$ significantly outperforms Gaussian FastMNMF ($\alpha = 2$).

**Index Terms**: speech enhancement, nonnegative matrix factorization, $\alpha$-stable distribution, joint diagonalization

## 1. Introduction

Multichannel speech enhancement aims to reduce the noise from corrupted speech signals captured by multiple sensors. It is an essential part of modern automatic speech recognition systems and hearing aids [1, 2]. Those applications require a flexible and tractable model to handle scenarios ranging from easy (anechoic, noiseless) to complex (strongly reverberant, highly noisy with impulsive noises).

A common approach is to consider in the short-time Fourier transform (STFT) domain a probabilistic model for all time-frequency (TF) bins of the sources. It is then convenient to assume that the observation is a linear combination of audio components and that all TF bins and sources are independent. Many works, for instance, assume a local Gaussian model that satisfies the linear stability condition (the reproductive property). The covariance matrix of each TF bin is then usually decomposed into a full-rank positive semidefinite matrix called a spatial covariance matrix (SCM) and a positive scalar representing the power spectral density (PSD) of the signal [3]. The direct estimation of those parameters, however, often results in a sub-optimal performance because it is hard to obtain accurate estimates.

The multichannel nonnegative matrix factorization (MNMF) is a well-known technique that decomposes the PSDs into separate frequency-dependent and time-dependent matrices [4]. This low-rank representation is effective to reduce the degree of freedom of the model and to improve the speech enhancement performance. Several extensions have been proposed to further improve both computation cost and enhancement performance. While independent low-rank matrix analysis (ILRMA) [5] proposes a fast and effective technique for a determined case, a recent approach called FastMNMF [6, 7, 8] uses a joint diagonalization technique to reduce the algorithm complexity. A heavy-tailed extension of FastMNMF based on a Student's $t$ model is described in [9]. The Student's $t$ model achieves good results for a more complex scenario of audio source separation. However, neither the Student's $t$ model [9, 10] nor the generalized Gaussian model [11] has the reproductive property.

The multivariate $\alpha$-stable distribution, where $\alpha \in (0, 2]$ is referred to as characteristic exponent, is the family of distributions satisfying the reproductive property [12]. A distribution with a smaller $\alpha$ has heavier tails and includes as a special case the Gaussian ($\alpha = 2$), Cauchy ($\alpha = 1$), and Levy ($\alpha = 0.5$) distributions. In speech enhancement, $\alpha$ has been used to characterize the impulsiveness of noise [13, 14]. The parameter estimation for a multivariate $\alpha$-stable model can be summarized in three major approaches. The first approach is to use a maximum likelihood technique as in [15], where a projected Cauchy multivariate distribution is used as a proxy. However, not all values of $\alpha$ have a closed-form probability density function (pdf.). The second approach circumvents this issue by using a unique spatial representation of parameter satisfying the reproductive property for non-Gaussian symmetric $\alpha$-stable vectors [16]. The third approach is to consider a sub-family called elliptically-contoured symmetric multivariate complex $\alpha$-stable distribution [17, 18], simply referred to as *elliptically stable distribution* hereafter, that can be seen as a Gaussian model given some positive $\alpha$-stable random variable, known as the impulse variable in speech enhancement literature [13, 14, 19].

This paper proposes a possibly ($\alpha < 2$) heavy-tailed elliptically stable source model and its convergence-guaranteed parameter estimation, which exploits its equivalent conditional Gaussian model in [18] and the joint-diagonalization technique of FastMNMF in [6, 7] applied to the shape matrix parameters to satisfy the reproductive property in the transformed domain. We first review the state-of-the-art Gaussian FastMNMF [6, 7] in Section 2. We then describe the proposed $\alpha$-stable FastMNMF in Section 3 and the estimation of its parameters in Section 4. We present speech enhancement experiments on subsets of the CHiME-4 dataset [20] in Section 5 to compare the performance of $\alpha$-stable FastMNMF to that of Gaussian FastMNMF [6], and ILRMA [5]. We finally draw a conclusion in Section 6.

http://dx.doi.org/10.21437/Interspeech.2020-3202

# 2. Gaussian FastMNMF

This section introduces the state-of-the-art blind source separation (BSS) method called FastMNMF, which is based on the complex Gaussian likelihood [6].

## 2.1. Model Formulation

Suppose that a mixture of $N$ sources are recorded by $M$ microphones. Let $\mathbf{X} = \{\mathbf{x}_{ft}\}_{f,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be the observed multichannel mixture complex spectrogram, where $F$ and $T$ denote the number of frequency bins and time frames, respectively. Let $\mathbf{S}_n = \{s_{nft}\}_{f,t=1}^{F,T} \in \mathbb{C}^{F\times T}$ be the single-channel source complex spectrogram and $\mathbf{X}_n = \{\mathbf{x}_{nft}\}_{f,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be its *image*. Assuming the additivity of spectra, $\mathbf{x}_{ft} \in \mathbb{C}^M$ is given by

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{x}_{nft} \tag{1}$$

and given $\mathbf{X}$ as observed data, the goal of BSS is to estimate the latent source images $\{\mathbf{X}_n\}_{n=1}^N$.

### 2.1.1. Source Model

The source model represents a probabilistic generative process of the source spectrogram $\mathbf{S}_n$, where $s_{nft}$ is assumed to follow a univariate complex Gaussian distribution as follows:

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{nft}\right), \tag{2}$$

where $\lambda_{nft}$ represents the PSD of source $n$ at frequency $f$ and time $t$. In the low-rank source model based on NMF, the source PSDs are assumed to have low-rank structure as follows:

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

where $K$ is the number of bases, $w_{nkf} \geq 0$ is the magnitude of basis $k$ of source $n$ at frequency $f$, and $h_{nkt} \geq 0$ is the activation of basis $k$ of source $n$ at time $t$.

### 2.1.2. Spatial Model

The spatial model represents a probabilistic generative model of the source image $\mathbf{X}_n$. If the sound propagation process (room acoustics) is time-invariant, we have

$$\mathbf{x}_{nft} = \mathbf{a}_{nf} s_{nft}, \tag{4}$$

where $\mathbf{a}_{nf} \in \mathbb{C}^M$ is the steering vector of source $n$ at frequency $f$. Using Eqs. (2) and (4), we have

$$\mathbf{x}_{nft} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \lambda_{nft}\mathbf{G}_{nf}) \triangleq \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{Y}_{nft}), \tag{5}$$

where $\triangleq$ means ”equals by definition”, $\mathbf{G}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}} \in \mathbb{S}_+^M$ is the rank-1 SCM of source $n$ at frequency $f$ with $.^{\mathsf{H}}$ the Hermitian transposition and $\mathbb{S}_+^M$ indicates the set of positive semidefinite matrices of size $M$. This is called the rank-1 spatial model used in ILRMA [5]. In the full-rank spatial model of MNMF, in contrast, the rank-1 constraint is removed for dealing with more realistic echoic conditions, *i.e.*, $\mathbf{G}_{nf}$ is regarded as a full-rank matrix. Using Eqs. (1) and (5) and the reproductive property of the Gaussian distribution, we have

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft}\mathbf{G}_{nf}\right) \triangleq \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{Y}_{ft}). \tag{6}$$

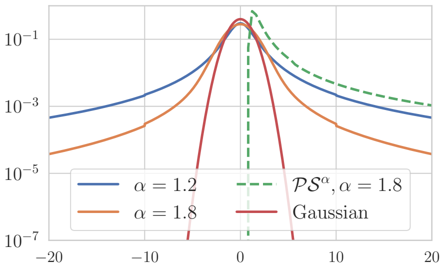


Figure 1: *Pdf. of symmetric and positive $\alpha$-stable distributions in the real scalar case*

In FastMNMF, a constrained version of MNMF, the SCMs of $N$ sources $\{\mathbf{G}_{nf}\}_{n=1}^N$ are assumed to be jointly diagonalizable as follows:

$$\forall n,\ \mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_n)\mathbf{Q}_f^{-\mathsf{H}}, \tag{7}$$

where $\tilde{\mathbf{g}}_n = [\tilde{g}_{n1}, \ldots, \tilde{g}_{nM}]^\top \in \mathbb{R}_+^M$ is a nonnegative vector with $.^\top$ denoting the transposition, $\mathrm{Diag}(\cdot)$ returns a diagonal matrix, and $\mathbf{Q}_f = [\mathbf{q}_{f1}, \ldots, \mathbf{q}_{fM}]^{\mathsf{H}} \in \mathbb{C}^{M\times M}$ is a nonsingular matrix called a *diagonalizer*, which is not limited to a unitary matrix. Using Eq. (7) into Eq. (6), we have

$$\mathbf{Q}_f\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft}\mathrm{Diag}(\tilde{\mathbf{g}}_n)\right). \tag{8}$$

This means that the elements of $\mathbf{Q}_f\mathbf{x}_{ft}$ are all independent because they follow a multivariate complex Gaussian distribution with a diagonal covariance matrix. This makes FastMNMF much more computationally efficient than MNMF because the factorization of $\mathbf{Q}_f\mathbf{x}_{ft}$ can be performed faster than that of $\mathbf{x}_{ft}$ with inter-element correlation.

## 2.2. Parameter Estimation

The parameters $\mathbf{W} = \{w_{nkf}\}_{n,k,f=1}^{N,K,F}$, $\mathbf{H} = \{h_{nkt}\}_{n,k,t=1}^{N,K,T}$, $\tilde{\mathbf{G}} = \{\tilde{\mathbf{g}}_n\}_{n=1}^N$, and $\mathbf{Q} = \{\mathbf{Q}_f\}_{f=1}^F$ are estimated jointly such that the log-likelihood function $\log p(\mathbf{X} \mid \mathbf{\Theta})$ is maximized where $\mathbf{\Theta} = \{\mathbf{Q}, \tilde{\mathbf{G}}, \mathbf{W}, \mathbf{H}\}$. Given parameters $\mathbf{\Theta}$, the source image $\mathbf{x}_{nft}$ are then estimated with a Wiener filtering. We refer to [6] for further details.

# 3. $\alpha$-Stable FastMNMF

This section briefly introduces the elliptically stable distribution and presents the proposed model called $\alpha$-stable FastMNMF.

## 3.1. Univariate and Multivariate $\alpha$-Stable Distributions

Suppose that a random variable $x$ follows an univariate symmetric isotropic zero-location $\alpha$-stable distribution $x \sim \mathcal{S}_{\mathbb{C}}^{\alpha}(0, v)$, where $v \geq 0$ is the scale parameter akin to the variance of a Gaussian distribution ($\alpha = 2$). The characteristic exponent $\alpha \in (0, 2]$ controls the heaviness of the distribution tails: the smaller $\alpha$ is, the heavier the tails are (*cf.* Figure 1). It has the reproductive property [12] and can also be expressed as a Gaussian scale mixture [21], *e.g.*, $x \sim \mathcal{S}_{\mathbb{C}}^{\alpha}(0, v)$ equals in distribution to $x \mid \phi \sim \mathcal{N}_{\mathbb{C}}(0, \phi v)$, where $\phi \sim \mathcal{PS}^{\alpha}\left(2\cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}\right)$ has a positive $\alpha$-stable distribution [12].

The multivariate $\alpha$-stable distribution includes a sub-family called elliptically-contoured symmetric multivariate complex $\alpha$-stable distribution [17, 18], referred to as *elliptically stable distribution* hereafter, which is a subset of Gaussian scale mixture.

A zero-location elliptically stable distribution parameterized by a shape matrix $\mathbf{V} \in \mathbb{S}_{+}^{M}$ [17]:

$$\mathbf{x} \sim \mathcal{S}_{\mathbb{C}}^{\alpha}(\mathbf{0}, \mathbf{V}) \tag{9}$$

can be expressed as $\mathbf{x} \,|\, \phi \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \phi\mathbf{V})$. The elliptically stable distribution can then be seen as a Gaussian distribution whose covariance $\mathbf{V}$ is randomly perturbed by a positive scalar $\phi$, whose value may be very large (*cf.* Figure 1). The so-called *impulse variable* $\phi$ [13] can then be used to characterize the impulsiveness of speech and noise. The elliptically stable distribution has the reproductive property according to the spatial representation theorem [12, 16] but the linearity of shape matrix representation does not coincide. We then exploit the conditional Gaussian distribution $\mathbf{x} \mid \phi$ in developing the $\alpha$-stable FastMNMF.

### 3.2. Model Formulation

We assume that a source image follows a zero-location elliptical stable distribution $\mathbf{x}_{nft} \sim \mathcal{S}_{\mathbb{C}}^{\alpha}(\mathbf{0}, \lambda_{nft}\mathbf{Q}_f^{-1}\mathrm{Diag}(\tilde{\mathbf{g}}_n)\mathbf{Q}_f^{-\mathrm{H}})$, where the characteristic exponent $\alpha \in (0, 2)$, $\lambda_{nft}$ is the source PSD and $\tilde{\mathbf{g}}_n$ is the diagonal elements of the diagonalized spatial shape matrix. As a surrogate of the elliptical stable distribution, we consider the conditional Gaussian model:

$$\begin{cases} \mathbf{x}_{nft} \mid \phi_{nft} & \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{Q}_f^{-1}\mathrm{Diag}\left(\tilde{\mathbf{y}}_{nft}\right)\mathbf{Q}_f^{-\mathrm{H}}\right) \\ \phi_{nft} & \sim \mathcal{PS}^{\alpha}\left(2\cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}\right) \end{cases} \tag{10}$$

where $\mathrm{Diag}\left(\tilde{\mathbf{y}}_{nft}\right) \triangleq \phi_{nft}\lambda_{nft}\mathrm{Diag}\left(\tilde{\mathbf{g}}_n\right)$. Since Eq. (10) implies that $\sum_n \mathbf{x}_{nft} \mid \phi_{nft}$ is also Gaussian, we can then formulate a marginalized Wiener filter given $\boldsymbol{\phi} = \{\phi_{nft}\}_{nft}$, $\mathrm{Diag}\left(\tilde{\mathbf{y}}_{ft}\right) \triangleq \sum_n \mathrm{Diag}\left(\tilde{\mathbf{y}}_{nft}\right)$ and $\mathbb{E}_{\boldsymbol{\phi}|\mathbf{X}}$ simplified as $\mathbb{E}_{\boldsymbol{\phi}}$:

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\phi}}\left[\mathbb{E}\left[\mathbf{x}_{nft}|\boldsymbol{\Theta}, \boldsymbol{\phi}, \mathbf{x}_{ft}\right]\right] \\ &\quad = \mathbb{E}_{\boldsymbol{\phi}}\left[\mathbf{Q}_f^{-1}\mathrm{Diag}\left(\tilde{\mathbf{y}}_{nft}\right)\mathrm{Diag}\left(\tilde{\mathbf{y}}_{ft}\right)^{-1}\mathbf{Q}_f^{-\mathrm{H}}\right]\mathbf{x}_{ft}. \end{aligned} \tag{11}$$

If we assume that all components of $\mathbf{Q}_f\mathbf{x}_{nft}$ are independent for all $n, f, t$, then the reproductive property holds and the filtering method in Eq. (11) can be replaced by an $\alpha$-fractional Wiener filter [22] applied element-wise on vector $\mathbf{Q}_f\mathbf{x}_{nft}$. To sum up, in addition to $\boldsymbol{\Theta} = \{\mathbf{Q}, \tilde{\mathbf{G}}, \mathbf{W}, \mathbf{H}\}$ as in Gaussian FastMNMF, the proposed $\alpha$-stable FastMNMF is also parameterized by $\boldsymbol{\phi}$.

## 4. Parameter Estimation

This section presents the parameter estimation of $\alpha$-stable FastMNMF. We first explain the estimation of $\boldsymbol{\Theta}$ by using an Expectation-Maximization (EM) technique and iterative projection method similar to the one explained in [6]. We then describe the estimation of $\boldsymbol{\phi}$ by a Metropolis-Hastings (MH) algorithm.

### 4.1. Estimation of $\boldsymbol{\Theta}$

In estimating $\boldsymbol{\Theta}$, we assume that $\mathbf{X}$ and $\phi$ are known. We consider the probabilistic model described in Section 3.2 with NMF in Eq. (3) for modeling the source PSD $\{\lambda_{nft}\}_{n,f,t=1}^{N,F,T}$. We minimize the log-likelihood $\log p(\mathbf{X}|\boldsymbol{\Theta}) = \log\int p(\mathbf{X}|\boldsymbol{\Theta}, \boldsymbol{\phi})p(\boldsymbol{\phi})d\boldsymbol{\phi}$ as follow:

$$\begin{aligned} \log p(\mathbf{X}|\boldsymbol{\Theta}) \geq & -\frac{1}{L}\sum_{f,t,m=1}^{F,T,M}\sum_{l=1}^{L}\left(\frac{\tilde{x}_{ftm}}{\tilde{y}_{ftml}} + \log\tilde{y}_{ftml}\right) \\ & + T\sum_{f=1}^{F}\log\left|\mathbf{Q}_f\mathbf{Q}_f^{\mathrm{H}}\right| - \mathrm{KL}[q(\boldsymbol{\phi})\|p(\boldsymbol{\phi})] \end{aligned} \tag{12}$$

where $L$ is the number of samples $\boldsymbol{\phi}$ for averaging the integral, $\tilde{x}_{ftm} = |\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{x}_{ft}|^2$, $\tilde{y}_{ftml} = \sum_{n,k}\phi_{nftl}w_{nkf}h_{nkt}\tilde{g}_{nm}$, $\phi_{nftl} \sim q(\phi_{nft})$, and KL is the Kullback-Leibler divergence. The multiplicative update rules are given by:

$$w_{nkf} \leftarrow w_{nkf}\sqrt{\frac{\sum_{t,m,l=1}^{T,M,L}\phi_{nftl}h_{nkt}\tilde{g}_{nm}\tilde{x}_{ftm}\tilde{y}_{ftml}^{-2}}{\sum_{t,m,l=1}^{T,M,L}\phi_{nftl}h_{nkt}\tilde{g}_{nm}\tilde{y}_{ftml}^{-1}}}, \tag{13}$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{\sum_{f,m,l=1}^{F,M,L}\phi_{nftl}w_{nkf}\tilde{g}_{nm}\tilde{x}_{ftm}\tilde{y}_{ftml}^{-2}}{\sum_{f,m,l=1}^{F,M,L}\phi_{nftl}w_{nkf}\tilde{g}_{nm}\tilde{y}_{ftml}^{-1}}}, \tag{14}$$

$$\tilde{g}_{nm} \leftarrow \tilde{g}_{nm}\sqrt{\frac{\sum_{f,t,k,l=1}^{F,T,K,L}\phi_{nftl}w_{nkf}h_{nkt}\tilde{x}_{ftm}\tilde{y}_{ftml}^{-2}}{\sum_{f,t,k,l=1}^{F,T,K,L}\phi_{nftl}w_{nkf}h_{nkt}\tilde{y}_{ftml}^{-1}}}. \tag{15}$$

The vectors $\mathbf{q}_{fm}$ are updated as in [23] by applying the following update with $\mathbf{V}_{fm} \triangleq \frac{1}{TL}\sum_{t,l=1}^{T,L}\mathbf{X}_{ft}\tilde{y}_{ftml}^{-1}$:

$$\mathbf{q}_{fm} \leftarrow \left(\mathbf{Q}_f\mathbf{V}_{fm}\right)^{-1}\mathbf{e}_m; \tag{16}$$

$$\mathbf{q}_{fm} \leftarrow \left(\mathbf{q}_{fm}^{\mathrm{H}}\mathbf{V}_{fm}\mathbf{q}_{fm}\right)^{-\frac{1}{2}}\mathbf{q}_{fm}, \tag{17}$$

where $\mathbf{e}_m$ is a one-hot vector where $m^{\mathrm{th}}$ component is equal to 1 and $\mathbf{q}_{fm}$ is the $m^{\mathrm{th}}$ column of the matrix $\mathbf{Q}_f$.

### 4.2. Estimation of $\boldsymbol{\phi}$

A direct estimation of $\boldsymbol{\phi}$ is untractable because it generally does not have an analytical expression for the pdf. We can, however, do approximation by obtaining multiple samples of $\boldsymbol{\phi}$ via a Metropolis-Hastings algorithm with Gibbs sampling [18]:

1. Draw a random sample from the prior disribution $\phi_{nft}^{\text{new}} \sim \mathcal{PS}^{\alpha}\left(2\cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}\right)$.
2. Draw $\nu \sim \mathcal{U}\left([0,1]\right)$ where $\mathcal{U}$ is the uniform distribution.
3. Compute the following acceptance probability:

$$\mathrm{acc}\left(\phi_{nft}^{\text{old}} \to \phi_{nft}^{\text{new}}\right) = \min\left(1, \frac{u_{nft}\left(\phi_{nft}^{\text{new}}\right)}{u_{nft}\left(\phi_{nft}^{\text{old}}\right)}\right) \tag{18}$$

   where $u_{nft}\left(\phi_{nft}\right)$ is the pdf. of a zero-mean Gaussian distribution whose covariance matrix $\phi_{nft}\lambda_{nft}\mathrm{Diag}\left(\tilde{\mathbf{g}}_n\right) + \sum_{n' \neq n}\phi_{n'ft}\lambda_{n'ft}\mathrm{Diag}\left(\tilde{\mathbf{g}}_{n'}\right)$ evaluated on $\mathbf{x}_{ft}$.
4. Test the acceptance:
   - if $\nu < \mathrm{acc}\left(\phi_{nft}^{\text{old}} \to \phi_{nft}^{\text{new}}\right)$, then $\phi_{nft} = \phi_{nft}^{\text{new}}$ (acceptance)
   - otherwise, $\phi_{nft} = \phi_{nft}^{\text{old}}$ (rejection).

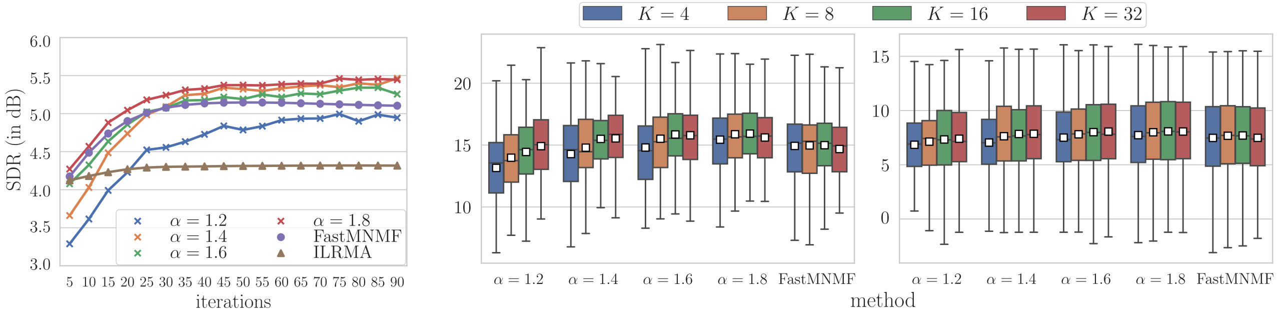
The Gaussian-FastMNMF complexity algorithm for one EM iteration is $\mathcal{O}\left(FM^2 + NM\right)$ and become $\mathcal{O}\left(L\left(FM^2 + NM\right)\right)$ for $\alpha-$FastMNMF.

## 5. Evaluation

This section evaluates the speech enhancement performance of the proposed $\alpha$-stable FastMNMF model and compares it to that of the state-of-the-art Gaussian FastMNMF and ILRMA.

### 5.1. Experimental Settings

We performed $M$-channel speech enhancement tasks ($N = 2$, *i.e.*, speech and noise) with $M \in \{2, 5\}$ on simulated and real data subsets of the CHiME-4 dataset [20]. Each subset consists of 20 randomly selected utterances for each of four different



(a) *Mean SDRs along the update iterations for $M = 2$ and $K = 16$ on the real data subset.*

(b) *Boxplots of SDRs on the simulated (left) and real (right) data subsets with $M = 5$ for various values of $\alpha$ and $K$. White squares show the mean SDRs.*

Figure 2: *SDR comparison for the different algorithms. 'FastMNMF' refers to the Gaussian FastMNMF.*

Table 1: *STOI and PESQ comparison for the different algorithms. STOI score ranges from* 0 *to* 1 *and PESQ score ranges from* $-0.5$ *to* 4.5*. Bold text indicates the best performance for each test set taking into account both mean and standard deviation.*

| **K = 16, STOI** | | $\alpha = 1.2$ | $\alpha = 1.4$ | $\alpha = 1.6$ | $\alpha = 1.8$ | FastMNMF | ILRMA |
|---|---|---|---|---|---|---|---|
| $M = 2$ | *Simu* | 0.87 ±0.06 | 0.88 ±0.06 | 0.89 ±0.05 | **0.89** ±0.05 | 0.88 ±0.06 | 0.87 ±0.06 |
| | *Real* | 0.74 ±0.11 | 0.75 ±0.10 | 0.75 ±0.10 | **0.76** ±0.10 | 0.75 ±0.10 | 0.75 ±0.10 |
| $M = 5$ | *Simu* | 0.92 ±0.05 | 0.93 ±0.04 | **0.94** ±0.04 | 0.94 ±0.07 | 0.92 ±0.03 | — |
| | *Real* | 0.75 ±0.09 | 0.76 ±0.09 | 0.77 ±0.09 | **0.78** ±0.09 | 0.77 ±0.09 | — |

| **K = 16, PESQ** | | $\alpha = 1.2$ | $\alpha = 1.4$ | $\alpha = 1.6$ | $\alpha = 1.8$ | FastMNMF | ILRMA |
|---|---|---|---|---|---|---|---|
| $M = 2$ | *Simu* | 2.11 ±0.44 | 2.14 ±0.46 | 2.15 ±0.43 | **2.15** ±0.40 | 2.12 ±0.47 | 2.03 ±0.47 |
| | *Real* | 2.12 ±0.43 | 2.15 ±0.40 | 2.16 ±0.47 | 2.18 ±0.41 | 2.14 ±0.44 | **2.19** ±0.43 |
| $M = 5$ | *Simu* | 2.42 ±0.58 | 2.49 ±0.51 | 2.51 ±0.47 | **2.54** ±0.48 | 2.48 ±0.50 | — |
| | *Real* | 2.27 ±0.52 | 2.3 ±0.56 | 2.32 ±0.50 | **2.33** ±0.43 | 2.29 ±0.51 | — |

environments (bus, cafe, pedestrian area, and street junction), amounting to 80 utterances in total. The performance is evaluated in terms of the signal to distortion ratio (SDR) [24], the perceptual evaluation speech quality (PESQ) [25], and the short-time objective measure (STOI) [26].

We compare the performance of the proposed $\alpha$-stable FastMNMF to that of the state-of-the-art Gaussian FastMNMF [7] and ILRMA [5]. The NMF coefficients $\mathbf{W}, \mathbf{H}$ whose number of bases varies $K \in \{4, 8, 16, 32\}$ are initialized with the absolute values of random samples from a Gaussian distribution. The demixing matrix for each frequency bin in ILRMA is initialized as an identity matrix. For both Gaussian and $\alpha$-stable FastMNMFs, the column vectors $\mathbf{q}_{fm}$ are initialized with the eigenvectors of the matrix $T^{-1} \sum_t \mathbb{E}[\mathbf{x}_{ft}\mathbf{x}_{ft}^H]$ and the diagonals of speech and noise SCMs are initialized as $[1, \epsilon, \ldots, \epsilon]^\top$ and $[M^{-1}, \ldots, M^{-1}]^\top$, respectively, where $\epsilon = 10^{-2}$. For the $\alpha$-stable FastMNMF, the same exponent $\alpha$ is used for both speech and noise with $\alpha \in \{1.2, 1.4, 1.6, 1.8\}$ and the number of MH iterations is set to $L = 5$. All algorithms run for 100 iterations.

### 5.2. Experimental Results & Discussions

Figure 2a illustrates the SDR evolution on the real data along the parameter update iterations for $M = 2$ and $K = 16$. With sufficient parameter updates, all FastMNMF variants outperform ILRMA and with appropriate setting of $\alpha$, the proposed $\alpha$-stable FastMNMFs outperform the Gaussian counterpart. We observe that $\alpha = 1.8$ is optimal among our settings. Conversely, we notice that the performance of $\alpha = 1.2$ is poor and relatively unstable, due to the fact that for a smaller $\alpha$, the impulse variable $\boldsymbol{\phi}$ has a wider range of values.

Figure 2b shows the SDR comparison on both simulated and real data for $M = 5$ and various $K$ after 100 update iterations. We observe that $\alpha = 1.8$ provides the best performance. For $\alpha \in \{1.2, 1.4\}$, the performance is improving along the increase of $K$. For $\alpha \in \{1.6, 1.8\}$, conversely, $K = 16$ is the optimal one.

Table 1 presents the PESQ and STOI comparison for $M \in \{2, 5\}$. Note that ILRMA only works for the determined case ($M = 2$, $N = 2$). In most cases, $\alpha = 1.8$ outperforms the others. The number of microphones $M$ increases the STOI and PESQ scores slightly for all methods. To summarize, we found that the $\alpha$-stable FastMNMF with an appropriate setting of $\alpha$ outperforms both Gaussian FastMNMF and ILRMA in terms of both signal quality and objective perceptual quality. We also found from our experiments that $\alpha = 1.8$ is the optimal choice to model the non-stationary behavior of both speech and noise.

## 6. Conclusion & Future Works

This paper proposes the $\alpha$-stable FastMNMF based on the elliptically-contoured multivariate complex $\alpha$-stable distribution as an extension to the state-of-the-art FastMNMF, which is a variant of MNMF where the source SCMs are jointly-diagonalizable and based on the multivariate complex Gaussian distribution. Experimental results show that the $\alpha$-stable FastMNMF outperforms the Gaussian counterpart indicating that an $\alpha$-stable distribution facilitates better modeling of the source non-stationary behavior. Future works include the direct estimation of $\alpha$-stable parameters, including the covariation matrices and the characteristic exponent $\alpha$, and the development of a Cauchy FastMNMF ($\alpha = 1$), for which a closed-form pdf. is available. An extension with a deep speech prior [27] can also be proposed and compare with DNN state-of-the-art speech enhancement algorithms.

## 7. Acknowledgements

This work was supported by JSPS KAKENHI Nos. 19H04137, 20K19833, and 20H01159; JST ERATO No. JPMJER1401; and NII CRIS-Line Collaborative Research.

## 8. References

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

[2] S. Makino, Ed., *Audio Source Separation*. Springer, 2018.

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

[4] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2009.

[5] 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.

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

[7] K. Sekiguchi, Y. Bando, 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/ACM Trans. ASLP*, 2020, under review.

[8] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 371–375.

[9] K. Kamo, Y. Kubo, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "FastMNMF based on multivariant complex Student's t distribution for blind source separation," *IEICE Tech. Rep.*, vol. 119, no. 253, pp. 23–29, 2019.

[10] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *Proc. IWAENC*. IEEE, 2016, pp. 1–5.

[11] 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. on Adv. in Signal Process.*, vol. 2018, no. 1, p. 28, 2018.

[12] G. Samoradnitsky, *Stable non-Gaussian random processes: stochastic models with infinite variance*. Routledge, 2017.

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

[14] M. Fontaine, F.-R. Stöter, A. Liutkus, U. Şimşekli, R. Serizel, and R. Badeau, "Multichannel audio modeling with elliptically stable tensor decomposition," in *Proc. LVA/ICA*, 2018, pp. 13–23.

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

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

[17] J. Nolan, "Multivariate elliptically contoured stable distributions: theory and estimation," *Computational Statistics*, vol. 28, no. 5, pp. 2067–2089, 2013.

[18] S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, and G. Richard, "Alpha-stable multichannel audio source separation," in *Proc. IEEE ICASSP*, 2017, pp. 576–580.

[19] S. Leglaive, U. Şimşekli, A. Liutkus, L. Girin, and R. Horaud, "Speech enhancement with variational autoencoders and alpha-stable distributions," in *Proc. IEEE ICASSP*, 2019, pp. 541–545.

[20] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[21] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 1, pp. 99–102, 1974.

[22] A. Liutkus and R. Badeau, "Generalized wiener filtering with fractional power spectrograms," in *Proc. IEEE ICASSP*. IEEE, 2015, pp. 266–270.

[23] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[24] 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.

[25] *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.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[27] K. Sekiguchi, Y. Bando, A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM Trans. ASLP*, vol. 27, no. 12, pp. 2197–2212, 2019.