ALPHA-STABLE MULTICHANNEL AUDIO SOURCE SEPARATION

Simon Leglaive[1], *Umut Şimşekli*[1], *Antoine Liutkus*[2], *Roland Badeau*[1], *Gaël Richard*[1]

1: LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
2: Inria, Speech Processing Team, Villers-lès-Nancy, France

ABSTRACT

In this paper, we focus on modeling multichannel audio signals in the short-time Fourier transform domain for the purpose of source separation. We propose a probabilistic model based on a class of heavy-tailed distributions, in which the observed mixtures and the latent sources are jointly modeled by using a certain class of multivariate alpha-stable distributions. As opposed to the conventional Gaussian models, where the observations are constrained to lie just within a few standard deviations from the mean, the proposed heavy-tailed model allows us to account for spurious data or important uncertainties in the model. We develop a Monte Carlo Expectation-Maximization algorithm for inferring the sources from the proposed model. We show that our approach leads to significant performance improvements in audio source separation under corrupted mixtures and in spatial audio object coding.

***Index Terms*—** Alpha-stable distributions, Multichannel source separation, Informed source separation, Monte Carlo Expectation-Maximization.

1. INTRODUCTION

Multichannel audio source separation is the task that aims to recover a set of source audio signals from an observed mixture signal that has multiple channels (e.g. stereo audio). The problem is called 'underdetermined' if the number of channels in the observed mixture is less than the number of sources. In this paper we focus on modeling mixtures of punctual sources in reverberant conditions for underdetermined multichannel audio source separation.

In source separation applications, audio signals are often represented in the time-frequency domain since this representation provides a natural interpretation due to its sparseness. Moreover, reverberant mixtures are easily modeled in this domain under a short reverberation assumption [1]. One of the most common approaches for tackling audio source separation is based on *variance modeling* frameworks [2]. In these probabilistic approaches, the Short-Time Fourier Transform (STFT) coefficients of each source are modeled as latent random variables following a complex circularly symmetric distribution with a time and frequency-dependent scale parameter. As a particular special case of complex circularly symmetric distributions, the complex Gaussian distributions have been widely used in audio source separation [3, 4, 5, 6, 7]. Within this framework, Non-negative Matrix Factorization (NMF) techniques are popular to model the spectro-temporal characteristics of the sources [8, 9, 10, 11]. The generative Gaussian source model based on NMF was first introduced in [8] and then applied to multichannel audio source separation in [4], where the reverberant mixing model relies on a frequency-dependent mixing matrix. Recently, extensions based on neural network variance models have also been proposed [12].

Even though the Gaussian models have proven successful in several scenarios, they might fall short when the observations exhibit strong variations or if they contain outliers. For addressing this issue in the single channel case, several NMF models with heavy-tailed observation models have been proposed. These heavy-tailed models include the Student's t-distribution [13], the Cauchy [10], and the Lévy distributions [14]. A more general NMF framework was presented in [11], where the observations are assumed to be α-stable distributed; generalizing several NMF models such as the Itakura-Saito (IS), Cauchy and Lévy NMF.

Research on heavy-tailed models for multichannel source separation has drawn some interest recently [15, 16, 17]. In this study, we propose a novel probabilistic generative model for multichannel audio source separation that tackles the problem of jointly modeling all channels by using a family of multivariate heavy-tailed distributions. Our approach consists in extending the Gaussian model proposed in [4] to a certain class of multivariate α-stable distributions. We develop a Monte Carlo Expectation Maximization (MCEM) algorithm for inferring the sources from the proposed model. This algorithm can be implemented by applying minor modifications to the Expectation-Maximization (EM) algorithm for the Gaussian model [4, 5].

We evaluate the proposed approach under two challenging applications. First we consider a multichannel musical source separation scenario, where the observed mixture signals are heavily corrupted. Then, we consider a coding-based informed source separation application similar to the one described in [18]. In both applications, we report superior performance compared to the Gaussian models.

2. TECHNICAL BACKGROUND

Stable distributions [19] are heavy-tailed distributions and they appear as the limit distributions in the generalized central limit theorem, i.e. when the variance of the random variables or vectors considered is no longer constrained to be finite. They are defined by the property that a sum of stable random variables is stable. As such, those distributions comprise the Gaussian, Cauchy and Lévy particular cases, among many others.

In this paper, we will be concerned with two particular cases of stable distributions. The first one is the positive α-stable distribution, abbreviated as $\mathcal{P}\frac{\alpha}{2}\mathcal{S}$, referring to non-negative scalar random variables. The second one concerns complex random vectors and is called the complex multivariate elliptically contoured stable distribution, denoted by $\mathcal{E}\alpha\mathcal{S}_c$.

Positive α-stable distributions $\mathcal{P}\frac{\alpha}{2}\mathcal{S}\,(x;\sigma)$ have the following characteristic function $\mathbb{E}[\exp(itx)] = \exp(-|\sigma t|^{\alpha/2} \left[1 - i\,\mathrm{sgn}(t)\tan\frac{\pi\alpha}{4}\right])$, where $\mathrm{sgn}(t)$ denotes the sign of t. They

This work is partly supported by the French National Research Agency (ANR) as a part of the EDISON 3D (ANR-13-CORD-0008-02) and KAMoulox (ANR-15-CE38-0003-01) projects.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

are parameterized by two scalars: $\alpha \in (0, 2)$ is called the characteristic exponent and determines the tail thickness of the distribution[1]. The smaller α, the thicker the tail of the distribution. $\sigma \in \mathbb{R}_+$ is a scale parameter measuring the spread of the random variable.

Stable random vectors may be defined in full generality through the analytical expression of their characteristic function, involving integration against a so-called *spectral measure* [19]. For practical purposes, in this study we will focus on a subclass of the complex multivariate stable distributions, called the *complex multivariate elliptically contoured stable distribution* and denoted by $\mathcal{E}\alpha\mathcal{S}_c$ [20]. Its characteristic function takes the following form: if $\mathbf{z} \in \mathbb{C}^K \sim \mathcal{E}\alpha\mathcal{S}_c\left(\boldsymbol{\Sigma}_\mathbf{z}\right)$, then

$$\mathbb{E}\left[\exp\left(i\Re\{\mathbf{t}^\star \mathbf{z}\}\right)\right] = \exp\left(-\left|(1/2)\mathbf{t}^\star \boldsymbol{\Sigma}_\mathbf{z} \mathbf{t}\right|^{\alpha/2}\right), \qquad (1)$$

where $\mathbf{t}$ is a $K \times 1$ complex vector and $\cdot^\star$ denotes Hermitian conjugation. As can be seen, $\mathcal{E}\alpha\mathcal{S}_c$ depends on two parameters: just like $\mathcal{P}\frac{\alpha}{2}\mathcal{S}$, the characteristic exponent $\alpha \in (0, 2]$ controls the thickness of the tails: choosing $\alpha < 2$ corresponds to a heavy-tailed model. Then, $\boldsymbol{\Sigma}_\mathbf{z}$ is a $K \times K$ positive definite matrix called the *shape matrix*. As a special case of (1) for $\alpha = 2$, we obtain the complex isotropic Gaussian distribution $\mathbf{z} \sim \mathcal{N}_c(0, 2\boldsymbol{\Sigma}_\mathbf{z})$. For $\alpha = 1$, we obtain the complex isotropic Cauchy distribution [10].

A very useful fact concerning $\mathcal{E}\alpha\mathcal{S}_c$ distributions is their *conditional Gaussianity*, justifying the fact they are also called "sub-Gaussian α-stable distributions", e.g. in [19]. Let $\mathbf{z} \sim \mathcal{E}\alpha\mathcal{S}_c\left(\boldsymbol{\Sigma}_\mathbf{z}\right)$. It can be shown that we may introduce the so called *impulse variable* $\phi \sim \mathcal{P}\frac{\alpha}{2}\mathcal{S}$ and get [19, 11]:

$$\mathbf{z} \sim \mathcal{E}\alpha\mathcal{S}_c\left(\boldsymbol{\Sigma}_\mathbf{z}\right) \Leftrightarrow \begin{cases} \phi \sim \mathcal{P}\frac{\alpha}{2}\mathcal{S}\left(2(\cos\frac{\pi\alpha}{4})^{2/\alpha}\right), \\ \mathbf{z} \mid \phi \sim \mathcal{N}_c\left(0, \phi\boldsymbol{\Sigma}_\mathbf{z}\right). \end{cases} \qquad (2)$$

As can be seen, the distribution of the impulse variable ϕ does not depend on the shape matrix $\boldsymbol{\Sigma}_\mathbf{z}$. This equivalent model (2) for $\mathbf{z} \sim \mathcal{E}\alpha\mathcal{S}_c$ can be interpreted in the following way. Realizations of $\mathbf{z}$ are expected overall to be distributed similarly to a Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}_\mathbf{z}$. However, its overall covariance is perturbed by the variable ϕ that is most of the time small, but may sometimes get significantly large, accounting for unexpected observations. The rate at which these spurious realizations occur is controlled by the characteristic exponent α of the $\mathcal{E}\alpha\mathcal{S}_c$ distribution. In the limiting case $\alpha = 2$, ϕ becomes deterministic and $\mathbf{z}$ is hence Gaussian. In all other $\alpha < 2$ cases, the $\mathcal{E}\alpha\mathcal{S}_c$ model permits us to model multivariate data in a way similar to the Gaussian case but accounting for data with high variability, or equivalently high uncertainty regarding the Gaussian assumption.

3. THE MODEL

Before describing the proposed model, we first present the multichannel audio source separation framework that was introduced in [4]. We consider an audio mixture of J source signals on I channels expressed in the STFT domain. We denote the observations and sources as $\mathbf{x}_{fn} = [x_{1,fn}, ..., x_{I,fn}]^T$ and $\mathbf{s}_{fn} = [s_{1,fn}, ..., s_{J,fn}]^T$, respectively, at the Time-Frequency (TF) point $(f, n) \in \{0, ..., F-1\} \times \{0, ..., N-1\}$. In [4], the observations and the sources are jointly modeled as follows:

$$\begin{pmatrix}\mathbf{x}_{fn} \\ \mathbf{s}_{fn}\end{pmatrix} \sim \mathcal{N}_c(0, \boldsymbol{\Sigma}_{fn}), \qquad (3)$$

where $\boldsymbol{\Sigma}_{fn} \in \mathbb{C}^{(I+J)\times(I+J)}$ is the positive definite covariance matrix. It is structured as follows:

$$\boldsymbol{\Sigma}_{fn} = \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{x},fn} & \mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},fn} \\ \boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^\star & \boldsymbol{\Sigma}_{\mathbf{s},fn}\end{pmatrix}, \qquad (4)$$

where

$$\boldsymbol{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^\star + \boldsymbol{\Sigma}_{\mathbf{b},f}. \qquad (5)$$

$\mathbf{A}_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I\times J}$ is called the mixing matrix, $\boldsymbol{\Sigma}_{\mathbf{b},f} = \sigma_{b,f}^2\mathbf{I}_I$ with $\sigma_{b,f}^2 > 0$, and $\mathbf{I}_I$ is the identity matrix of size $I \times I$. The source covariance matrices $\boldsymbol{\Sigma}_{\mathbf{s},fn}$ are further parametrized by using an NMF model:

$$\boldsymbol{\Sigma}_{\mathbf{s},fn} = \operatorname{diag}([v_{j,fn}]_j), \text{ with } v_{j,fn} = [\mathbf{W}_j\mathbf{H}_j]_{fn}, \qquad (6)$$

where $\operatorname{diag}([v_{j,fn}]_j)$ is the diagonal matrix constructed from the coefficients $\{v_{j,fn}\}_{j=1}^J$ and $\mathbf{W}_j \in \mathbb{R}_+^{F\times K_j}$, $\mathbf{H}_j \in \mathbb{R}_+^{K_j\times N}$ are called the dictionary and activation matrices of source j, respectively.

This model has been shown to be useful in various scenarios. However, as we have discussed in the earlier sections, the Gaussian assumption might be limiting in certain applications. Therefore, in this study we consider a heavy-tailed distribution for jointly modeling the sources and the mixture observations:

$$\begin{pmatrix}\mathbf{x}_{fn} \\ \mathbf{s}_{fn}\end{pmatrix} \sim \mathcal{E}\alpha\mathcal{S}_c(\boldsymbol{\Sigma}_{fn}), \qquad (7)$$

where $\boldsymbol{\Sigma}_{fn}$ is the shape matrix of the elliptically contoured α-stable distribution and it has the same structure as in (4)-(6). From (2) we can express this model as conditionally Gaussian as follows:

$$\phi_{fn} \sim \mathcal{P}\frac{\alpha}{2}\mathcal{S}\left(2\Big(\cos\frac{\pi\alpha}{4}\Big)^{2/\alpha}\right),$$

$$\begin{pmatrix}\mathbf{x}_{fn} \\ \mathbf{s}_{fn}\end{pmatrix} |\phi_{fn} \sim \mathcal{N}_c(0, \phi_{fn}\boldsymbol{\Sigma}_{fn}). \qquad (8)$$

Thanks to the heavy-tailed structure, we expect that this model will allow larger variations in the observed data and be more robust to outliers. Note that we obtain the Gaussian model if we set $\alpha = 2$.

In source separation applications, the purpose is to estimate the source signals given the observations and the model parameters. If we assume that the model parameters $\boldsymbol{\Theta} = \{\{\mathbf{W}_j, \mathbf{H}_j\}_j, \{\boldsymbol{\Sigma}_{\mathbf{b},f}, \mathbf{A}_f\}_f\}$ are known, we are naturally interested in a Minimum Mean Squared Error (MMSE) estimate of the sources $\mathbf{s}_{fn}$. From (4) and (8) we can first identify the following conditional distribution:

$$\mathbf{s}_{fn}|\mathbf{x}_{fn}, \phi_{fn}; \boldsymbol{\Theta} \sim \mathcal{N}_c\Big(\hat{\mathbf{s}}_{fn}, \phi_{fn}\boldsymbol{\Sigma}_{\mathbf{s},fn}^{cond}\Big), \qquad (9)$$

where $\hat{\mathbf{s}}_{fn}$ and $\boldsymbol{\Sigma}_{\mathbf{s},fn}^{cond}$ are defined at lines 4 and 5 of Algorithm 1 respectively. The MMSE estimate of $\mathbf{s}_{fn}$ is then given as follows:

$$\mathbb{E}_{\mathbf{s}_{fn}|\mathbf{x}_{fn};\boldsymbol{\Theta}}[\mathbf{s}_{fn}] = \mathbb{E}_{\phi_{fn}|\mathbf{x}_{fn};\boldsymbol{\Theta}}\left[\mathbb{E}_{\mathbf{s}_{fn}|\mathbf{x}_{fn},\phi_{fn};\boldsymbol{\Theta}}[\mathbf{s}_{fn}]\right] = \hat{\mathbf{s}}_{fn}.$$

Note that this estimate does not depend on ϕ_{fn}. Moreover, it is exactly the same estimate as in the Gaussian model [4]; if the parameters are identical, both models lead to the same source estimate. In our applications, we further need to compute the posterior covariance of the sources, that is given as follows:

$$\mathbb{E}_{\mathbf{s}_{fn}|\mathbf{x}_{fn};\boldsymbol{\Theta}}\Big[\big(\mathbf{s}_{fn} - \hat{\mathbf{s}}_{fn}\big)\big(\mathbf{s}_{fn} - \hat{\mathbf{s}}_{fn}\big)^\star\Big] = \mathbb{E}_{\phi_{fn}|\mathbf{x}_{fn};\boldsymbol{\Theta}}[\phi_{fn}]\boldsymbol{\Sigma}_{\mathbf{s},fn}^{cond}, \qquad (10)$$

where $\mathbb{E}_{\phi_{fn}|\mathbf{x}_{fn};\boldsymbol{\Theta}}[\phi_{fn}]$ is the posterior mean of ϕ_{fn}, which does not admit an analytical form. In the next section, we will develop a method for approximately computing this expectation.

[1] The choice of defining $\mathcal{P}\frac{\alpha}{2}\mathcal{S}$ through $\frac{\alpha}{2}$ with $\alpha < 2$ instead of simply $\alpha < 1$ comes out of notational convenience.

4. INFERENCE

In this section we derive an MCEM algorithm [21] for estimating the parameters of the proposed model. Let $\mathbf{X} = \{\mathbf{x}_{fn}\}_{f,n}$ be the set of observed data while $\mathbf{S} = \{\mathbf{s}_{fn}\}_{f,n}$ and $\mathbf{\Phi} = \{\phi_{fn}\}_{f,n}$ denote the set of hidden variables.

Our aim is to estimate the parameters in a Maximum Likelihood (ML) sense, i.e. by maximizing the log-likelihood $\ln p(\mathbf{X}|\mathbf{\Theta})$. This estimation can be done by maximizing a lower bound of the log-likelihood, typically with an EM algorithm [22]. At the E-step of the algorithm we compute the following lower-bound from the current estimation of the parameters $\mathbf{\Theta}'$:

$$Q(\mathbf{\Theta}|\mathbf{\Theta}') = \mathbb{E}_{\mathbf{S},\mathbf{\Phi}|\mathbf{X},\mathbf{\Theta}'}[\ln p(\mathbf{X}, \mathbf{S}, \mathbf{\Phi}|\mathbf{\Theta})]. \tag{11}$$

It is defined as the conditional expectation of the complete data log-likelihood. The M-step then aims to maximize this lower bound in order to obtain a new estimate of the parameters. These two steps are iterated until convergence. One iteration of the proposed MCEM algorithm is summarized in Algorithm 1 and we detail its derivation in the following sub-sections. Further derivation details can be found at [23].

4.1. E-step

At the E-step we compute the lower bound (11):

$$\begin{aligned} Q(\mathbf{\Theta}|\mathbf{\Theta}') \overset{c}{=} -N \sum_{f=0}^{F-1} \Big[\ln\det(\mathbf{\Sigma}_{\mathbf{b},f}) + \mathrm{tr}\Big(\mathbf{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\hat{\mathbf{R}}_{\phi\mathbf{ss},f}\mathbf{A}_f^\star \\ - \mathbf{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\hat{\mathbf{R}}_{\phi\mathbf{xs},f}^\star - \mathbf{\Sigma}_{\mathbf{b},f}^{-1}\hat{\mathbf{R}}_{\phi\mathbf{xs},f}\mathbf{A}_f^\star + \mathbf{\Sigma}_{\mathbf{b},f}^{-1}\hat{\mathbf{R}}_{\phi\mathbf{xx},f}\Big)\Big] \\ - \sum_{j=1}^{J}\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[\ln(v_{j,fn}) + \frac{\hat{p}_{j,fn}}{v_{j,fn}}\right], \end{aligned} \tag{12}$$

where $\overset{c}{=}$ denotes equality up to an additive constant, $\hat{\mathbf{R}}_{\phi\cdot\cdot,f} = \frac{1}{N}\sum_{n=0}^{N-1}\hat{\mathbf{R}}_{\phi\cdot\cdot,fn}$, $\hat{\mathbf{R}}_{\phi\mathbf{ss},fn} = \mathbb{E}_{\mathbf{S},\mathbf{\Phi}|\mathbf{X},\mathbf{\Theta}'}[\phi_{fn}^{-1}\mathbf{s}_{fn}\mathbf{s}_{fn}^\star]$, $\hat{\mathbf{R}}_{\phi\mathbf{xs},fn} = \mathbb{E}_{\mathbf{S},\mathbf{\Phi}|\mathbf{X},\mathbf{\Theta}'}[\phi_{fn}^{-1}\mathbf{x}_{fn}\mathbf{s}_{fn}^\star]$, $\hat{\mathbf{R}}_{\phi\mathbf{xx},fn} = \mathbb{E}_{\mathbf{S},\mathbf{\Phi}|\mathbf{X},\mathbf{\Theta}'}[\phi_{fn}^{-1}\mathbf{x}_{fn}\mathbf{x}_{fn}^\star]$ and $\hat{p}_{j,fn} = \mathbb{E}_{\mathbf{S},\mathbf{\Phi}|\mathbf{X},\mathbf{\Theta}'}[\phi_{fn}^{-1}|s_{j,fn}|^2]$. After straightforward calculations, these statistics can be further written as in Algorithm 1. As can be seen, they involve the computation of $\mathbb{E}_{\mathbf{\Phi}|\mathbf{X},\mathbf{\Theta}'}[\phi_{fn}^{-1}]$. Unfortunately, similarly as in (10), this expectation cannot be written in a closed-form analytical expression. In the sequel, we will derive a method for approximately computing these expectations.

Algorithm 1: One iteration of the MCEM algorithm.

```
E-step:
 1: $\mathbf{\Sigma}_{\mathbf{s},fn} = \mathrm{diag}([v_{j,fn}]_j)$ with $v_{j,fn} = [\mathbf{W}_j\mathbf{H}_j]_{fn}$
 2: $\mathbf{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^\star + \mathbf{\Sigma}_{\mathbf{b},f}$
 3: $\mathbf{G}_{\mathbf{s},fn} = \mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^\star\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}$
 4: $\hat{\mathbf{s}}_{fn} = \mathbf{G}_{\mathbf{s},fn}\mathbf{x}_{fn}$
 5: $\mathbf{\Sigma}_{\mathbf{s},fn}^{cond} = (\mathbf{I}_J - \mathbf{G}_{\mathbf{s},fn}\mathbf{A}_f)\mathbf{\Sigma}_{\mathbf{s},fn}$
 6: Compute $\hat{\phi}_{fn}^{-1} = \mathbb{E}_{\mathbf{\Phi}|\mathbf{X},\mathbf{\Theta}'}[\phi_{fn}^{-1}]$ with the MH algorithm
 7: $\hat{\mathbf{R}}_{\phi\mathbf{ss},fn} = \hat{\phi}_{fn}^{-1}\hat{\mathbf{s}}_{fn}\hat{\mathbf{s}}_{fn}^\star + \mathbf{\Sigma}_{\mathbf{s},fn}^{cond}$
 8: $\hat{p}_{j,fn} = [\hat{\mathbf{R}}_{\phi\mathbf{ss},fn}]_{j,j}$
 9: $\hat{\mathbf{R}}_{\phi\mathbf{ss},f} = \frac{1}{N}\sum_n \hat{\mathbf{R}}_{\phi\mathbf{ss},fn}$
10: $\hat{\mathbf{R}}_{\phi\mathbf{xx},f} = \frac{1}{N}\sum_n \hat{\phi}_{fn}^{-1}\mathbf{x}_{fn}\mathbf{x}_{fn}^\star$
11: $\hat{\mathbf{R}}_{\phi\mathbf{xs},f} = \frac{1}{N}\sum_n \hat{\phi}_{fn}^{-1}\mathbf{x}_{fn}\hat{\mathbf{s}}_{fn}^\star$
M-step:
12: $\mathbf{A}_f = \hat{\mathbf{R}}_{\phi\mathbf{xs},f}\hat{\mathbf{R}}_{\phi\mathbf{ss},f}^{-1}$
13: $\mathbf{\Sigma}_{\mathbf{b},f} = \mathrm{tr}(\hat{\mathbf{R}}_{\phi\mathbf{xx},f} - \mathbf{A}_f\hat{\mathbf{R}}_{\phi\mathbf{xs},f}^\star - \hat{\mathbf{R}}_{\phi\mathbf{xs},f}\mathbf{A}_f^\star$
        $+\mathbf{A}_f\hat{\mathbf{R}}_{\phi\mathbf{ss},f}\mathbf{A}_f^\star)\mathbf{I}_I/I$
14: $\mathbf{W}_j, \mathbf{H}_j = \text{IS-NMF}(\hat{\mathbf{P}}_j)$ with $\hat{\mathbf{P}}_j = [\hat{p}_{j,fn}]_{fn}$
```

4.2. Estimating Intractable Expectations via MCMC

In this section, we develop a Markov Chain Monte Carlo (MCMC) algorithm for approximating the intractable expectations by computing an average of samples that are drawn from $p(\mathbf{\Phi}|\mathbf{X}, \mathbf{\Theta})$. We approximate the intractable expectations in question as follows:

$$\mathbb{E}[f(\phi_{fn})] = \int f(\phi_{fn})\pi(\phi_{fn})d\phi_{fn} \approx \frac{1}{M}\sum_{m=1}^{M} f\left(\phi_{fn}^{(m)}\right) \tag{13}$$

where $\phi_{fn}^{(m)}$ are samples that are drawn from the target distribution $\pi(\phi_{fn}) = p(\phi_{fn}|\mathbf{X}, \mathbf{\Theta})$. More precisely, we develop a Metropolis-Hastings (MH) algorithm that generates samples from $\pi(\phi_{fn})$ in two steps. In the m-th iteration of this algorithm, we firstly draw a random sample ϕ'_{fn} from the prior distribution $\phi'_{fn} \sim \mathcal{P}\frac{\alpha}{2}\mathcal{S}\left(2(\cos\frac{\pi\alpha}{4})^{2/\alpha}\right)$ and we compute an acceptance probability, given as follows:

$$\mathrm{acc}(\phi_{fn} \to \phi'_{fn}) = \min\left\{1, \frac{\mathcal{N}_c(\mathbf{x}_{fn}; 0, \phi'_{fn}\mathbf{\Sigma}_{\mathbf{x},fn})}{\mathcal{N}_c(\mathbf{x}_{fn}; 0, \phi_{fn}\mathbf{\Sigma}_{\mathbf{x},fn})}\right\}. \tag{14}$$

Then, we draw a uniform random number $u \sim \mathcal{U}([0, 1])$. If $u < \mathrm{acc}(\phi_{fn}^{(m-1)} \to \phi'_{fn})$, we accept the sample and set $\phi_{fn}^{(m)} = \phi'_{fn}$; otherwise we reject the sample and set $\phi_{fn}^{(m)} = \phi_{fn}^{(m-1)}$.

4.3. M-step

Zeroing the gradient of $Q(\mathbf{\Theta}|\mathbf{\Theta}')$ with respect to the mixing matrix $\mathbf{A}_f$ and the shape matrix $\mathbf{\Sigma}_{\mathbf{b},f}$ leads to the updates given at the M-step of Algorithm 1. Moreover, up to an additive constant independent of $v_{j,fn}$, we can recognize in the last line of (12) the IS divergence [8] between $\hat{p}_{j,fn}$ and $v_{j,fn} = [\mathbf{W}_j\mathbf{H}_j]_{fn}$. Therefore, similarly as in [5], the update of the source parameters at line 14 of Algorithm 1 is done by computing an NMF on $\hat{\mathbf{P}}_j = [\hat{p}_{j,fn}]_{fn} \in \mathbb{R}_+^{F\times N}$ using the IS divergence. It can be done with the standard multiplicative update rules (see [8]).

5. EXPERIMENTS

Our experiments are conducted on audio tracks provided by the Musical Audio Signal Separation (MASS) dataset [24]. We created 8 stereo mixtures by simulating mixing filters with the Roomsimove toolbox [25]. The room was a $4.45 \times 3.55 \times 2.5$ m shoebox with a reverberation time of 128 ms[2]. One mixture lasts between 12 and 28 seconds and contains between 2 and 4 sources. The mixtures are instrumental and do not contain singing voice as voice signals are not accurately modeled by NMF. As the MASS dataset provides stereo sources, each one is first converted to mono, downsampled to 16 kHz and filtered with the associated RIRs to create a source image. We finally sum the source images to create a mixture.

[2] The reverberation time is defined as the time it takes for the sound energy to decrease by 60 dB after extinction of the source.

Initialization	Oracle		Blind	
Model	Gaussian	α-stable	Gaussian	α-stable
SDR (dB)	-3.7	**4.5**	-6.7	**0.6**

Table 1. Average SDR: Gaussian model with standard Wiener filtering versus α-stable model with the modified estimation procedure.

5.1. Musical Source Separation on Corrupted Audio Mixtures

This section aims to illustrate the behavior and the capabilities of our model given by (8). We would expect that the impulse variable ϕ_{fn} represents the uncertainty about the model at TF point (f, n). Indeed, we observe in (10) that the posterior covariance of the sources depends on the posterior mean of ϕ_{fn}. If this quantity is high we can conclude that the estimate of the sources at this TF point is uncertain. We can thus expect to obtain a high posterior mean of ϕ_{fn} for the TF points that are not accurately represented by the source or mixing models. For demonstrating this claim we conduct a first experiment that is meant to be a proof of concept. For each mixture in the dataset, we corrupt a few number of TF points by setting them to very high values, resulting in a highly audible noise. There is no structure in this noise, so that it does not fit an NMF model. As can be seen in line 4 of Algorithm 1, the sources are estimated by Wiener filtering of the noisy mixture. The corrupted TF points thus propagate to the source estimates, which results in noisy estimated sources. For overcoming this issue and showing that the impulse variables behave as expected, we propose a modified estimator given by:

$$\hat{\mathbf{s}}_{fn}^{modif} = \mathbb{E}_{\mathbf{s}_{fn},\phi_{fn}|\mathbf{x}_{fn};\Theta}[\mathbf{s}_{fn}\phi_{fn}^{-1}] = \hat{\mathbf{s}}_{fn}\hat{\phi}_{fn}^{-1}, \tag{15}$$

where $\hat{\mathbf{s}}_{fn}$ and $\hat{\phi}_{fn}^{-1}$ are defined at lines 4 and 6 of Algorithm 1 respectively. We are thus scaling the standard Wiener source estimate by the posterior mean of the inverse impulse variable.

We perform the source separation on all the corrupted mixtures in the dataset. We compare the Gaussian model [4, 5] using standard Wiener filtering with the proposed α-stable model (with α empirically chosen as equal to 1.5) using the modified estimation procedure (15). The separation is performed from an oracle initialization (the parameters are initialized from the true sources and mixing filters) and from a blind initialization. We evaluate the source separation performance using the Signal-to-Distortion Ratio (SDR) [26] expressed in decibels (dB). We used the BSS Eval Toolbox to compute this measure [27]. The average SDR over all the sources in the dataset is shown in Table 1. We can observe a great improvement in the SDR with the α-stable model. This improvement is due to the fact that scaling $\hat{\mathbf{s}}_{fn}$ with $\hat{\phi}_{fn}^{-1}$ suppresses the noise, by setting to zero the corrupted TF points. This experiment thus confirms that the impulse variable is able to capture outliers that do not fit for example the NMF model.

5.2. Coding-Based Informed Source Separation

Informed source separation (ISS) is a framework where the audio sources are known at an encoding stage. Side information can thus be computed at the encoder and transmitted along with the mixtures to the decoder, where source separation is performed. ISS was introduced in [28] and relies on transmitting parameters that permit to recover the sources using the mixtures at the decoder. The main drawback of ISS is that the separation quality is limited by the separation method, which usually cannot recover the sources perfectly whatever the allocated bitrate. To overcome this issue, Coding-based ISS (CISS) was introduced, with the idea of also encoding the source signals themselves [29, 18]. CISS makes use of the source posterior distribution, given the mixture. The sources are represented by their posterior means, which are characterized by a set of transmitted parameters. The main idea of CISS is to further encode the error between the true sources and their posterior means, through the use of the posterior covariance matrix.

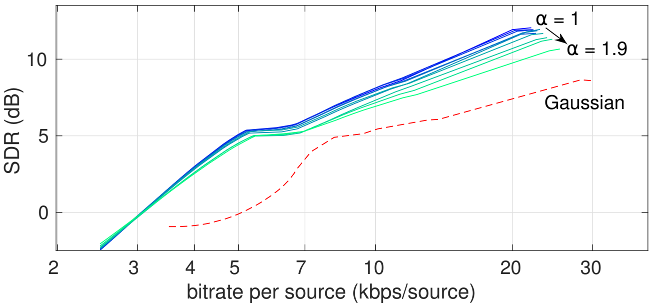

Fig. 1. Rate-SDR curves for multiple values of α (blue to green curves) compared with the Gaussian model (red dashed curve).

As mentioned previously, for a fixed set of parameters, the posterior mean $\hat{\mathbf{s}}_{fn}$ is the same in the Gaussian and α-stable models. However the posterior covariance of the sources differs. Indeed, it can be seen in (10) that the α-stable model involves the posterior mean of ϕ_{fn} in the definition of the posterior covariance matrix. In this experiment we adapt the Gaussian CISS method presented in [18] to our α-stable model. It consists in applying the exact same source encoding procedure but using the posterior covariance matrix given in (10). Due to space limitations we do not detail the encoding algorithm but it can be found in [18].

For each mixture of the dataset, we run the CISS algorithms for the Gaussian model [18] and for the proposed α-stable one. We studied various values of α going from $\alpha = 1$ to 1.9 with a step of 0.1. Both algorithms are run with different levels of quality, corresponding to different source quantization step-sizes. For all the mixtures in the dataset we then obtain a scatter-plot of the SDR in dB according to the bitrate in kbit per second per source (kbps/source). The higher the SDR, the lower the distortion. We then smooth this scatter-plot using the local regression (LOESS) method [30]. The results are presented in figure 1. As can be seen, the α-stable model leads to higher SDR (or lower distortion) than the Gaussian model. Interestingly, we can observe that as α tends to 2, the CISS results with the α-stable model are getting smoothly closer to the ones with the Gaussian model. Indeed, if $\alpha = 2$ the elliptically contoured stable distribution becomes the complex isotropic Gaussian one. We can also mention that choosing $\alpha < 1$ did not improve the results.

6. CONCLUSION

In this paper we proposed an extension of the multichannel audio source separation framework [4] by using elliptically contoured stable distributions. We developed an MCEM algorithm that turned out to be a modified version of the EM algorithm for the Gaussian model [4, 5]. We showed that the use of heavy-tailed distributions permits to add some flexibility, as they are more robust to outliers that do not fit the structure of the model. The effectiveness of our approach has been shown for separating sources from corrupted mixtures and for a CISS application. Future work will include investigating a more justified way of using the impulse variable in the source estimate than the procedure given in (15).

7. REFERENCES

[1] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[2] E. Vincent, M. G. Jafari, S. A. Abdallah, M. Plumbley, and M. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*, pp. 162–185. 2010.

[3] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency gaussian source models," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2005, pp. 78–81.

[4] A. Ozerov and C. Févotte, "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] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 257–260.

[6] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830 –1840, 2010.

[7] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

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

[9] T. Virtanen, T. Cemgil, and S. Godsill, "Bayesian extensions to nonnegative matrix factorisation for audio signal modelling," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 1825–1828.

[10] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy Nonnegative Matrix Factorization," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.

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

[12] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, June 2016.

[13] K. Yoshii, K. Itoyama, and M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 51–55.

[14] P. Magron, R. Badeau, and A. Liutkus, "Lévy NMF for robust nonnegative source separation," *arXiv preprint arXiv:1608.01844*, 2016.

[15] D. Fitzgerald, A. Liutkus, and R. Badeau, "Projet – spatial audio separation using projections," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 36–40.

[16] D. Fitzgerald, A. Liutkus, and R. Badeau, "Projection-based demixing of spatial audio," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1560–1572, 2016.

[17] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, 2016, pp. 1–5.

[18] A. Liutkus, A. Ozerov, R. Badeau, and G. Richard, "Spatial coding-based informed source separation," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 2012, pp. 2407–2411.

[19] G. Samoradnitsky and M. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*, vol. 1, CRC Press, 1994.

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

[21] G. C. G. Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *Journal of the American statistical Association*, vol. 85, no. 411, pp. 699–704, 1990.

[22] 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)*, pp. 1–38, 1977.

[23] "Supporting document," `http://perso.telecom-paristech.fr/simsekli/multialpha/alphaStableMASS_icassp17.pdf`.

[24] M. Vinyes, "MTG MASS dataset," `http://mtg.upf.edu/download/datasets/mass`, 2008.

[25] E. Vincent and D.R. Campbell, "Roomsimove," `http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip`, 2008.

[26] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[27] E. Vincent, "BSS Eval Toolbox Version 3.0 for Matlab," `http://bass-db.gforge.inria.fr/bss_eval/`, 2007.

[28] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1721–1733, 2011.

[29] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed source separation: source coding meets source separation," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2011, pp. 257–260.

[30] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988.