Multichannel Audio Modeling with Elliptically Stable Tensor Decomposition

Mathieu Fontaine[1], Fabian-Robert Stöter[2], Antoine Liutkus[2], Umut Şimşekli[3], Romain Serizel[1], Roland Badeau[3]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France.
[2] Inria and LIRMM, Montpellier, France.
[3] LTCI, Télécom ParisTech, Université Paris-Saclay, Paris, France.

Abstract. This paper introduces a new method for multichannel speech enhancement based on a versatile modeling of the residual noise spectrogram. Such a model has already been presented before in the single channel case where the noise component is assumed to follow an alpha-stable distribution for each time-frequency bin, whereas the speech spectrogram, supposed to be more regular, is modeled as Gaussian. In this paper, we describe a multichannel extension of this model, as well as a Monte Carlo Expectation - Maximisation algorithm for parameter estimation. In particular, a multichannel extension of the Itakura-Saito nonnegative matrix factorization is exploited to estimate the spectral parameters for speech, and a Metropolis-Hastings algorithm is proposed to estimate the noise contribution. We evaluate the proposed method in a challenging multichannel denoising application and compare it to other state-of-the-art algorithms.

1 Introduction

In many contexts, speech denoising is studied and applied in order to obtain, among other things, a comfortable listening or broadcast of a talk [3], by exploiting the observed noisy signal, obtained by several microphones. From an audio source separation perspective, this denoising is achieved through a probabilistic model, where the observed signal is divided into two latent sources: a noise component and a target source.

Both speech and noise components are usually considered in the *time-frequency* (TF) domain where all TF-bins are supposed to be independent and follow a Gaussian law [8, 17]. A common approach to speech enhancement is the spectral subtraction method [9, 10]. Its principle is to estimate an a priori *signal to noise ratio* (SNR) in order to infer a *short-time spectral amplitude* (STSA) estimator of the noise which will be subtracted to the STSA of the observations. Another popular trend is to decompose the *power spectral densities* (PSD) of sources into a product of two matrices. The *non-negative matrix factorization* (NMF) model assumes that the PSDs admit low-rank structures and it performs well in denoising [25].

To the best of our knowledge, NMF models for multichannel speech enhancement have been proposed only in a Gaussian probabilistic context, whereas a non-Gaussian approach could offer a more flexible model for noise and speech. Moreover, a good initialization in a Gaussian NMF model is crucial to avoid staying stuck in a local minimum [4]. Many studies in the single-channel case have shown a better robustness to initialization when the signal is modeled in the TF domain with as heavy tail distribution [27, 23].

Among this type of distributions, α-stable distributions preserve interesting properties satisfied by Gaussian laws, and they can model distributions ranging from light tails as in the *Gaussian case* to heavy tails as in the *Cauchy case*. Indeed, α-stable distributions are the only ones which admit a central limit theorem and stability by summation [20]. Various studies have been carried out on audio modeling using alpha-stable processes [23, 16]. Especially in the TF domain, a generalization of wide-sense stationary (WSS) processes [17] has been established in the $\alpha-$stable case [16] and applied to noise reduction [12]. More precisely, in [24] it was considered that the target source is Gaussian and the residual noise is α-stable, in order to get a greater flexibility on noise modeling.

This paper introduces a generalization of [24] to the multichannel case. The goal is to develop a Gaussian NMF model for speech that assumes a low-rank structure for speech covariances [8], while the noise part is taken as an $\alpha-$stable process. Parameters are estimated through a combination of the multichannel extension of Itakura Saito NMF (IS-NMF) [21] for speech and a Markov Chain Monte Carlo (MCMC) strategy for estimating the noise part. The proposed method is evaluated for multichannel denoising, and compared to other state-of-the-art algorithms.

2 Probabilistic and Filtering models

2.1 Mixture model

Let $\boldsymbol{x} \in \mathbb{C}^{F\times T\times K}$ be the observed data in the short-time Fourier transform (STFT) domain where F, T and K denote the number of frequency bands, time frames and microphones, respectively. The observation $\boldsymbol{x}$ will be assumed to be the sum of two latent audio sources: the first one is written $\boldsymbol{y} \in \mathbb{C}^{F\times T\times K}$ and accounts for the *speech signal.* The second one is written $\boldsymbol{r} \in \mathbb{C}^{F\times T\times K}$ and called the *residual component.* We have:

$$\boldsymbol{x}_{ft} = \boldsymbol{y}_{ft} + \boldsymbol{r}_{ft}, \tag{1}$$

where each term belongs to $\mathbb{C}^K$. The main goal in this paper is to estimate $\boldsymbol{y}$ and $\boldsymbol{r}$ knowing $\boldsymbol{x}$, by using a probabilistic model described below.

2.2 Source model

At short time scales, the speech signal may be assumed stationary and does not feature strong impulsiveness. This motivates modeling it as a locally stationary

Gaussian process [17]. Furthermore, we also assume that the different channels for $\boldsymbol{y}_{ft}$ are correlated, accounting for the *spatial* characteristics of the signal. Consequently, we assume that each $\boldsymbol{y}_{ft}$ is an isotropic complex Gaussian vector[1] of mean $\mathbf{0}$ and covariance matrix $\boldsymbol{C}_{ft}^{\boldsymbol{y}} \triangleq \boldsymbol{R}_f v_{f,t}$, where the *spatial covariance matrix* $\boldsymbol{R}_f \in \mathbb{C}^{K\times K}$ encodes the time-invariant correlations of speech in the different channels and v_{ft} is the PSD of the speech signal [8]. To exploit the redundancy of speech, we further decompose v_{ft} through NMF and obtain:

$$\forall f,t \quad \boldsymbol{y}_{ft} \sim \mathcal{N}_c\left(\boldsymbol{y}_{ft}; 0, \boldsymbol{R}_f v_{ft} \triangleq \boldsymbol{R}_f \sum_{l=1}^{L} w_{fl} h_{lt}\right). \tag{2}$$

where $\triangleq$ means "equals by definition"and $\boldsymbol{W} \in \mathbb{R}_+^{F\times L}, \boldsymbol{H} \in \mathbb{R}_+^{L\times T}$ are matrices which respectively contain all non-negative scalars w_{fl} and h_{lt}. While $\boldsymbol{W}$ is understood as L spectral bases, $\boldsymbol{H}$ stands for their activations over time. To make notations simpler, let $\boldsymbol{\Theta} \triangleq \{\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}\}$ be the parameters to be estimated with $\boldsymbol{R} \triangleq \{\boldsymbol{R}_f\}_f$.

In contrast to the speech signal, the model of the residual component should allow for outliers and impulsiveness. To do so, the residual part is modeled by an heavy-tailed distribution in the time domain. Recent works proposed a stationary model called $\alpha-$harmonizable process with $\alpha \in (0,2]$ in the single-channel case. It is shown in [20, 16] that such a model is equivalent to assuming that the signal at every time-frequency bin f,t follows a complex isotropic symmetric $\alpha-$stable distribution. With the aim of extending the previous model to a multichannel one, we take all $\boldsymbol{r}_{ft}$ as distributed with respect to an *elliptically contoured multivariate stable distribution* (ECMS, denoted $\mathcal{E}\alpha S$) and independent of one another. These distributions, which are a particular case of $\alpha-$stable distributions, have the particularity of requiring only two parameters [20, 15]:

1. A *characteristic exponent* $\alpha \in (0,2]$: the smaller α, the heavier the tails of the distribution.
2. A positive definite Hermitian *scatter matrix* in $\mathbb{C}^{K\times K}$.

In this article, the scatter matrices for all $\boldsymbol{r}_{ft}$ are taken equal to $\sigma_f \boldsymbol{I}_K$, where $\boldsymbol{I}_K \in \mathbb{R}^{K\times K}$ is the identity matrix and $\sigma_f > 0$ is a positive scalar that does not depend on time. We have:

$$\forall f,t \quad \boldsymbol{r}_{ft} \sim \mathcal{E}\alpha S^K\left(\sigma_f \boldsymbol{I}_K\right). \tag{3}$$

2.3 Filtering model

As mentioned in subsection 2.1, we aim to reconstruct the sources $\boldsymbol{y}$ and $\boldsymbol{r}$ from the observed data $\boldsymbol{x}$. From a signal processing point of view, when parameters $\boldsymbol{\sigma}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}$ are known, one would like to compute the Minimum Mean Squared

[1] The probability density function (PDF) of an isotropic complex Gaussian vector is $\mathcal{N}_C(\boldsymbol{x}|\mu, \boldsymbol{C}) = \frac{1}{\pi^K \det \boldsymbol{C}} \exp\left(-\left(\boldsymbol{x}-\mu\right)^\star \boldsymbol{C}^{-1}\left(\boldsymbol{x}-\mu\right)\right)$.

Error (MMSE) estimates of both sources. In our probabilistic context, these can be expressed as the posteriori expectations $\mathbb{E}\left(\boldsymbol{y}_{ft}|\boldsymbol{x}_{ft},\boldsymbol{\Theta},\boldsymbol{\sigma}\right)$.

To compute such estimates, a property specific to ECMS distributions can be exploited to represent $\boldsymbol{r}$ as a complex normal distribution $\mathcal{N}_c$ of dimension K, whose variance is randomly multiplied by a positive random *impulse variable* ϕ_{ft} distributed as $\mathcal{P}\frac{\alpha}{2}S\left(2\cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}\right)$, where $\mathcal{P}\frac{\alpha}{2}S$ is the *positive* $\alpha/2$*-stable distribution* (see [23] for more details):

$$\forall f,t \quad \boldsymbol{r}_{ft}|\phi_{ft} \sim \mathcal{N}_c\left(\boldsymbol{r}_{ft};0,\phi_{ft}\sigma_f\boldsymbol{I}_k\right), \tag{4}$$

If we assume for now that $\boldsymbol{\Phi} \triangleq \{\phi_{ft}\}_{f,t}$ are known in (4), we get the distribution of the mixture as:

$$\forall f,t \quad \boldsymbol{x}_{ft}|\phi_{ft} \sim \mathcal{N}_c\left(\boldsymbol{x}_{ft};\,0,\boldsymbol{C}^{\boldsymbol{x}|\phi}_{ft}\right), \tag{5}$$

where $\boldsymbol{C}^{\boldsymbol{x}|\phi}_{ft} \triangleq \boldsymbol{R}_f\sum_{l=1}^{L} w_{fl}h_{lt} + \phi_{ft}\sigma_f\boldsymbol{I}_k$. This in turns allows to build a multichannel Wiener filter as [3]:

$$\mathbb{E}\left(\boldsymbol{y}_{ft}|\boldsymbol{x}_{ft},\boldsymbol{\Phi},\boldsymbol{\Theta},\boldsymbol{\sigma}\right) = \boldsymbol{C}^{\boldsymbol{y}}_{ft}\left(\boldsymbol{C}^{\boldsymbol{x}|\phi}_{ft}\right)^{-1}\boldsymbol{x}_{ft}, \tag{6}$$

with $.^{-1}$ standing for matrix inversion.

Now, the strategy we adopt here is to marginalize this expression over $\boldsymbol{\Phi} \mid x$, to get:

$$\hat{\boldsymbol{y}}_{ft} = \mathbb{E}_{\boldsymbol{\Phi}|x}\left[\mathbb{E}\left[\boldsymbol{y}_{ft}|\boldsymbol{x}_{ft},\boldsymbol{\Phi},\boldsymbol{\Theta},\boldsymbol{\sigma}\right]\right] = \boldsymbol{G}_{ft}\boldsymbol{x}_{ft},$$

where

$$\boldsymbol{G}_{ft} \triangleq \boldsymbol{C}^{\boldsymbol{y}}_{ft}\boldsymbol{\Xi}_{ft} \tag{7}$$

is the marginal Wiener filter, and $\boldsymbol{\Xi}_{ft} \triangleq \mathbb{E}_{\boldsymbol{\Phi}|x}\left[\left(\boldsymbol{C}^{\boldsymbol{x}|\phi}_{ft}\right)^{-1}\right]$ is the average inverse mixture covariance matrix. We will explain how to compute $\boldsymbol{\Xi}$ later in section 3.3.

3 Parameter Estimation

3.1 Expectation-Maximization (EM) algorithm

Assuming that the observations $\boldsymbol{x}$ and the impulse variable $\boldsymbol{\phi}$ are known, we first aim to estimate the parameters $\boldsymbol{\Theta}$. We choose a maximum likelihood estimator in order to get the most likely source NMF parameters $\boldsymbol{W},\boldsymbol{H}$:

$$\left(\boldsymbol{W}^\star,\boldsymbol{H}^\star,\boldsymbol{R}^\star\right) = \arg\max_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{R}} \log\mathbb{P}\left(\boldsymbol{x},\boldsymbol{\Phi}\,|\,\boldsymbol{\Theta},\boldsymbol{\sigma}\right), \tag{8}$$

where $\boldsymbol{\Phi}$ is a latent variable and $\log\mathbb{P}\left(\boldsymbol{x},\boldsymbol{\Phi}\,|\,\boldsymbol{\Theta},\boldsymbol{\sigma}\right)$ is the log-likelihood. As in [24], we propose an EM algorithm. This method aims to minimize an upper-bound of $\mathcal{L}_n\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R}\right) = -\log\mathbb{P}\left(\boldsymbol{x},\boldsymbol{\Phi}\,|\,\boldsymbol{\Theta},\boldsymbol{\sigma}\right)$. This approach is summarized in the following two steps:

$$\text{E-Step:} \qquad \mathcal{Q}_n\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R}\right) = -\mathbb{E}_{\boldsymbol{\Phi}|\boldsymbol{x},\boldsymbol{W}^{(n-1)},\boldsymbol{H}^{(n-1)}}\left[\mathcal{L}_n\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R}\right)\right], \tag{9}$$

$$\text{M-Step:} \qquad \left(\boldsymbol{W}^{(n)},\boldsymbol{H}^{(n)},\boldsymbol{R}^{(n)}\right) = \arg\max_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{R}} \mathcal{Q}_n\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R}\right). \tag{10}$$

E-Step: We first introduce a positive function that upper-bounds the negative log-likelihood $\mathcal{L}_n\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R}\right)$, which is equal to [21]:

$$\mathcal{L}_n\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R}\right) = \sum_{f,t}\left[\operatorname{tr}\left(\tilde{\boldsymbol{X}}_{ft}\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\phi}\right)^{-1}\right) + \log\det \boldsymbol{C}_{ft}^{\boldsymbol{x}|\phi}\right] \tag{11}$$

where $\tilde{\boldsymbol{X}}_{ft} \triangleq \boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^{\star}$ and $.^{\star}$ stands for the Hermitian transposition. A positive auxiliary function $\mathcal{L}_n^{+}\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R},\boldsymbol{U},\boldsymbol{V}\right) = \sum_{f,t}\Big[\sum_l \frac{\operatorname{tr}\left(\tilde{\boldsymbol{X}}_{ft}\boldsymbol{U}_{lft}\left(\boldsymbol{C}_{lft}^{\boldsymbol{x}|\phi}\right)^{-1}\boldsymbol{U}_{lft}\right)}{w_{fl}h_{lt}} + \frac{\operatorname{tr}\left(\tilde{\boldsymbol{X}}_{ft}\boldsymbol{U}_{rft}^{2}\right)}{\sigma_f\phi_{ft}} + \log\det\boldsymbol{V}_{ft} + \frac{\det\boldsymbol{C}_{ft}^{\boldsymbol{x}|\phi} - \det\boldsymbol{V}_{ft}}{\det\boldsymbol{V}_{ft}}\Big]$ which satisfies:

$$\mathcal{L}_n^{+}\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R},\boldsymbol{U},\boldsymbol{V}\right) \geq \mathcal{L}_n\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R}\right) \tag{12}$$

is introduced in [21]. Using (12) and the definition of $\mathcal{Q}_n$ in (9), we obtain:

$$\mathbb{E}_{\boldsymbol{\Phi}|\boldsymbol{x}}\mathcal{L}_n\left(.\right) \leq \mathbb{E}_{\boldsymbol{\Phi}|\boldsymbol{x}}\mathcal{L}_n^{+}\left(.\right) \triangleq \mathcal{Q}_n^{+}\left(.\right) \tag{13}$$

with:

$$\begin{aligned}\mathcal{Q}_n^{+}\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R},\boldsymbol{U},\boldsymbol{V}\right) = \sum_{f,t}\Bigg[\sum_l \frac{\mathbb{E}_{\boldsymbol{\Phi}|\boldsymbol{x}}\left(tr\left[\tilde{\boldsymbol{X}}_{ft}\boldsymbol{U}_{lft}\left(\boldsymbol{C}_{lft}^{\boldsymbol{x}|\phi}\right)^{-1}\boldsymbol{U}_{lft}\right]\right)}{w_{fl}h_{lt}} \\ +\mathbb{E}_{\boldsymbol{\Phi}|\boldsymbol{x}}\left(tr\left[\tilde{\boldsymbol{X}}_{ft}\boldsymbol{U}_{rft}^{2}\right]\right)\sigma_f^{-1}\phi_{ft}^{-1} + \mathbb{E}_{\boldsymbol{\Phi}|\boldsymbol{x}}\left(\log\det\boldsymbol{V}_{ft} + \det\left(\boldsymbol{V}_{ft}^{-1}\boldsymbol{C}_{lft}^{\boldsymbol{x}|\phi}\right) - 1\right)\Bigg]\end{aligned} \tag{14}$$

The form in (14) admits partial derivatives that will be useful as part of a multiplicative update [11] in the M-Step.

M-Step: Solving the M-Step in (10) is equivalent to zeroing the partial derivatives $\frac{\partial\mathcal{Q}_n^{+}}{\partial w_{fl}}$ and $\frac{\partial\mathcal{Q}_n^{+}}{\partial h_{lt}}$ and to set $\boldsymbol{U},\boldsymbol{V}$ such that the equality in (13) is verified. A multiplicative update approach yields:

$$w_{fl} \leftarrow w_{fl}\sqrt{\frac{\sum_t h_{lt}\operatorname{tr}\left(\boldsymbol{R}_f\boldsymbol{P}_{ft}\right)}{\sum_t h_{lt}\operatorname{tr}\left(\boldsymbol{R}_f\boldsymbol{\Xi}_{ft}\right)}} \tag{15}$$

$$h_{lt} \leftarrow h_{lt}\sqrt{\frac{\sum_f w_{fl}\operatorname{tr}\left(\boldsymbol{R}_f\boldsymbol{P}_{ft}\right)}{\sum_f w_{fl}\operatorname{tr}\left(\boldsymbol{R}_f\boldsymbol{\Xi}_{ft}\right)}} \tag{16}$$

where the quantity $\boldsymbol{\Xi}_{ft} = \mathbb{E}_{\boldsymbol{\Phi}|\boldsymbol{x}}\left[\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\varphi_i}\right)^{-1}\right]$ has been used above in (7) and $\boldsymbol{P}_{ft} = \mathbb{E}_{\boldsymbol{\Phi}|\boldsymbol{x}}\left[\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\varphi_i}\right)^{-1}\tilde{\boldsymbol{X}}_{ft}\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\varphi_i}\right)^{-1}\right]$. We will explain how to compute these expectations in subsection 3.3.

3.2 Estimation of spatial covariance matrices and noise gains $\boldsymbol{\sigma}$

We update the spatial covariance matrix $\boldsymbol{R}$ in the M-step as in [8], further using the trick proposed in [18] to use a weighted update, resulting in:

$$\boldsymbol{R}_f \leftarrow \left(\sum_t v_{ft}\right)^{-1} \times \sum_t \left(\boldsymbol{C}_{ft}^{\boldsymbol{y}\boldsymbol{y}^\star|\boldsymbol{x}}\right), \tag{17}$$

where: $\boldsymbol{C}_{ft}^{\boldsymbol{y}\boldsymbol{y}^\star|\boldsymbol{x}} \triangleq \boldsymbol{G}_{ft}\tilde{\boldsymbol{X}}_{ft}\boldsymbol{G}_{ft} + \boldsymbol{C}_{ft}^{\boldsymbol{y}} - \boldsymbol{G}_{ft}\boldsymbol{C}_{ft}^{\boldsymbol{y}}$ is the total posterior variance for the speech source.

Concerning the estimation of the noise gain $\boldsymbol{\sigma}$ in (3), we exploit a result in [5] that if $z \sim \mathcal{E}\alpha S\left(\sigma\right)$, then $\mathbb{E}\left[\|z\|^p\right]^{\frac{\alpha}{p}} \propto \sigma$, for $p < \alpha$, with $\propto$ standing for proportionality. The strategy we adopt is to apply this estimation only once at the beginning of the algorithm to the mixture itself, by taking a robust estimation like the median $\mathbb{M}$ instead of the average, to account for the fact that not all TF bins pertain to the noise, but that a small portion also pertain to speech. We thus pick $p = \alpha/2$ and take:

$$\sigma_f \leftarrow \mathbb{M}\left(\|\sum_t \boldsymbol{x}\left(f,t\right)\|^{\alpha/2}\right)^2. \tag{18}$$

3.3 Expectation estimation using Metropolis-Hastings algorithm

We still have to calculate the expectations $\boldsymbol{\Xi}_{ft}$ and $\boldsymbol{P}_{ft}$. Unfortunately, they cannot be calculated analytically. To address this issue, we set up a Markov Chain Monte Carlo (MCMC) algorithm in order to approximate the expectations for each iteration. We are focusing on the Metropolis-Hastings algorithm through an empirical estimation of $\boldsymbol{\Xi}_{ft}$ and $\boldsymbol{P}_{ft}$ as follows:

$$\overline{\boldsymbol{\Xi}}_{ft} \simeq \frac{1}{I}\sum_{i=1}^{I}\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\varphi_i}\right)^{-1} \tag{19}$$

$$\overline{\boldsymbol{P}}_{ft} \simeq \frac{1}{I}\sum_{i=1}^{I}\left(\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\varphi_i}\right)^{-1}\tilde{\boldsymbol{X}}_{ft}\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\varphi_i}\right)^{-1}\right) \tag{20}$$

with $\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\varphi_i}\right)^{-1} = \left[\sum_l \left(\boldsymbol{R}_{fl}w_{fl}h_{lt}\right) + \varphi_{ft,i}\sigma_f\boldsymbol{I}_k\right]^{-1}$ and $\varphi_{ft,i}$ are sampled as follows:

First Step (Sampling process): Generate a sampling via the prior distribution $\varphi'_{ft} \sim \mathcal{P}\frac{\alpha}{2}S\left(2\cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}\right)$.

Second Step (Acceptance):

- Draw $u \sim \mathcal{U}([0,1])$ where $\mathcal{U}$ denotes the uniform distribution.
- Compute the following acceptance probability:

$$\mathrm{acc}\left(\varphi_{ft} \to \varphi'_{ft}\right) = \min\left(1, \frac{\mathcal{N}_c\left(\boldsymbol{x}_{ft}; 0,\ \varphi'_{ft}\sigma_f \boldsymbol{I}_K + \boldsymbol{C}^{\boldsymbol{y}}_{ft}\right)}{\mathcal{N}_c\left(\boldsymbol{x}_{ft}; 0,\ \varphi_{ft}\sigma_f \boldsymbol{I}_K + \boldsymbol{C}^{\boldsymbol{y}}_{ft}\right)}\right)$$

- Test the acceptance:
  - if $u < \mathrm{acc}\left(\varphi_{ft,\,i-1} \to \varphi'_{ft}\right)$, then $\varphi_{ft,\,i} = \varphi'_{ft}$ (acceptance)
  - otherwise, $\varphi_{ft,\,i} = \varphi_{ft,\,i-1}$ (rejection)

4 Single-Channel Speech Signal Reconstruction

Let $\hat{\boldsymbol{y}}$ be the multichannel signal obtained after Wiener filtering (7). In the context of speech enhancement, the desired speech is rather a single-channel signal, that we write $\widehat{\boldsymbol{s}} \in \mathbb{C}^{F\times T}$. In this study, we take $\hat{\boldsymbol{s}}$ as a linear combination of $\hat{\boldsymbol{y}}$ with a time-invariant *beamformer* $\boldsymbol{B}_f \in \mathbb{C}^K$ [26]:

$$\hat{\boldsymbol{s}}_{ft} \triangleq \boldsymbol{B}^\star_f \hat{\boldsymbol{y}}_{ft},$$

Where $.^\star$denotes the Hermitian transposition. There are many ways to devise the beamformer $\boldsymbol{B}_f$. In this study, we choose to maximize the energy of $\boldsymbol{B}^\star_f \boldsymbol{y}_{ft} \mid \boldsymbol{x}$, which means maximizing:

$$\begin{aligned}\frac{1}{T}\sum_t \mathbb{E}\left(\left|\boldsymbol{B}^\star_f \boldsymbol{y}_{ft}\right|^2 |\boldsymbol{x}_{ft}\right) &= \boldsymbol{B}^\star_f \mathbb{E}\left(\boldsymbol{y}_{ft}\boldsymbol{y}^\star_{ft}|\boldsymbol{x}\right)\boldsymbol{B}_f. \\ &= \boldsymbol{B}^\star_f \frac{1}{T}\sum_t \left(\boldsymbol{C}^{\boldsymbol{y}\boldsymbol{y}^\star|\boldsymbol{x}}_{ft}\right)\boldsymbol{B}_f.\end{aligned}$$

The solution of this optimization problem is to choose $\boldsymbol{B}_f$ as the eigenvector associated to the largest eigenvalue of the Hermitian matrix $\frac{1}{T}\sum_t \left(\boldsymbol{C}^{\boldsymbol{y}\boldsymbol{y}^\star|\boldsymbol{x}}_{ft}\right)$ [8]. The Algorithm 1 summarizes all the steps of our proposed method for denoising.

Algorithm 1 Denoising Algorithm

1. **Inputs** :
   - mixture $\boldsymbol{x}$
   - number of components L
   - numbers N of EM iterations.
2. **Initialization**
   - Compute $\boldsymbol{\sigma}$ as in (18)
   - Initialize $\boldsymbol{W}$ and $\boldsymbol{H}$ randomly
   - $\boldsymbol{R}_f \leftarrow \boldsymbol{I}_K$
   - $\phi_{ft} \sim \mathcal{P}\frac{\alpha}{2}S\left(2\cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}\right)$
3. **EM algorithm,** for $n = 1...N$
   - MH algorithm:
     (a) Draw φ_{ift} via Metropolis-Hastings algorithm (subsection 3.3)
     (b) Compute $\overline{\boldsymbol{\Xi}}$ (19) and $\overline{\boldsymbol{P}}$ (20)
   - Update $\boldsymbol{W}$ (15), $\boldsymbol{H}$ (16) and $\boldsymbol{R}$ (17)
4. **Image Source reconstruction**
   compute $\hat{\boldsymbol{y}}$ as in (7)
5. **Beamforming**
   - Set $\boldsymbol{B}_f$ as the principal eigenvector of $\frac{1}{T}\sum_t \boldsymbol{C}_{ft}^{\boldsymbol{y}\boldsymbol{y}^\star|\boldsymbol{x}}$
   - Compute $\hat{\boldsymbol{s}}_{ft} = \boldsymbol{B}_f^\star \hat{\boldsymbol{y}}_{ft}$

5 Evaluation

We investigate both the quality of speech enhancement and the audio source separation performance. Our proposed approach will be compared to two baseline methods:

ARC Our proposed method: alpha residual component (ARC) which mixes a Gaussian component and an $\alpha-$stable noise. We will run $N = 10$ iterations for the EM part and select $\alpha = 1.9$.

MWF The classic multi-channel Wiener filter (MWF) [2, 6] which assumes that both noise and speech are Gaussian in the time-frequency domain. The multichannel Wiener filter is defined as the best estimator minimizing the mean squared error (MSE) between the estimated and the ground truth source.

GEVD Introduced in [22], the generalized eigenvalue decomposition (GEVD) multichannel Wiener filter is based on a low-rank approximation of the autocorrelation matrix of the speech signal in order to provide a more robust noise reduction.

5.1 Experimental setup

The corpus for evaluation is made up of mono speech excerpts from Librispeech [19] with a sample rate of $16\,kHz$. They are placed end-to-end with several silence periods for a total length of 3 minutes and assembled with three different environmental noises taken from Aurora [14]: babble noise, restaurant and train. We apply on both signals an STFT using a Hann window with an FFT length of 1024 and 50% overlap. A 'perfect' voice activity detection (VAD), in the sense that the VAD is estimated on the clean speech, is used on all three methods.

Those excerpts are further convolved with different room impulse responses (RIR) provided by Roomsimove in order to get reverberant stereophonic signals. The room dimensions are $5 \times 4 \times 3$ meters and reverberation times, based on a 60dB decay (RT60), are 0 and 500 ms. The distance between the microphones is 15 cm and the center of the microphone array is located at the center of the room at 1.5 m height. The sources are 1 m from the center of the microphone array. For more challenges, two spatial settings and 4 signal-to-noise (SNR) ratios will be proposed. The different SNR values are -5, 0, 5, 10 dB and the spatial configurations are an angular difference of 30°or 90° between both sources (the speech source is always facing the microphone array). In short, a total of 48 noisy sources have been denoised by the three proposed methods.

Regarding the MWF-based algorithms, the autocorrelation matrices are estimated over the full audio segments and the filters computed from these matrices are then constant over time.

5.2 Performance measures

For the evaluation, two scores will be measured: the first one is a speech intelligibility weighted spectral distortion (SIW-SD) measure and the second one is a speech intelligibility-weighted SNR (SIW-SNR) [13].

The SIW-SD measure is defined as:

$$\mathrm{SIW-SD} = \sum_i I_i \mathrm{SD}_i \tag{21}$$

where I_i is the band importance function [1] and SD_i the average SD (in dB) in the i-th one third octave band,

$$\mathrm{SD}_i = \frac{1}{(2^{1/6} - 2^{-1/6}) f_i^c} \int_{2^{-1/6} f_i^c}^{2^{1/6} f_i^c} |10 \log_{10} G^y(f)| df \tag{22}$$

with center frequencies f_i^c and $G^y(f)$ is given by:

$$G^y(f) = \frac{P_{\boldsymbol{y}}(f)}{P_{\hat{\boldsymbol{y}}}(f)} \tag{23}$$

where $P_{\boldsymbol{y}}(f)$ and $P_{\hat{\boldsymbol{y}}}(f)$ are the power, for the frequency f, of the speech component of the input signal $\boldsymbol{y}$ and the speech component output signal $\hat{\boldsymbol{y}}$, respectively.

The SIW-SNR [13] is used here to compute the *SIW-SNR improvement* which is defined as

$$\Delta \mathrm{SNR}_{\mathrm{intellig}} = \sum_i I_i (\mathrm{SNR}_{i,\mathrm{out}} - \mathrm{SNR}_{i,\mathrm{in}}) \tag{24}$$

where $\text{SNR}_{i,\text{out}}$ and $\text{SNR}_{i,\text{in}}$ represent the output SNR of the noise reduction filter and the SNR of the signal in the first microphone in the i^{th} band, respectively.

5.3 Results

In a first experiment we study the impact of the reverberation on the performance of each algorithm. Figures 1 and 2 present the SIW-SNR improvement and the SIW-SD performance averaged over noise types and spatial scenarios and depending on the input SNR for the RIR with $RT_{60} = 0ms$ (anechoic room) and $RT_{60} = 500ms$, respectively. In the scenario where $RT_{60} = 0ms$, ARC is outperformed by MWF based algorithms both in terms of SIW-SNR improvement and SIW-SID.

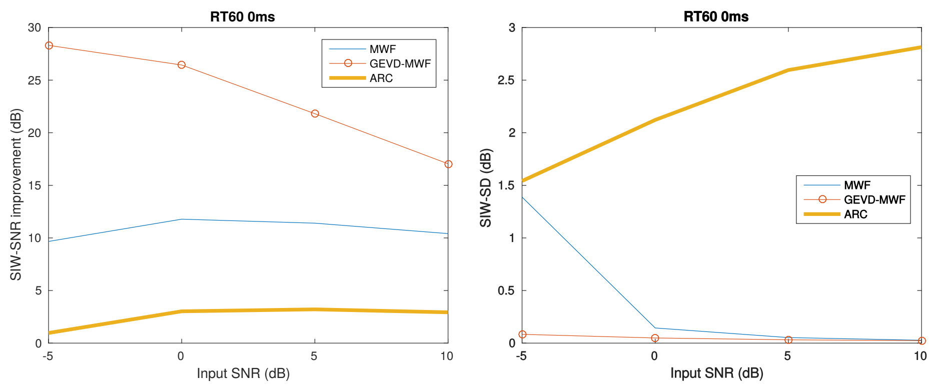

Fig. 1. SIW-SNR & SIW-SD in an anechoic scenario.

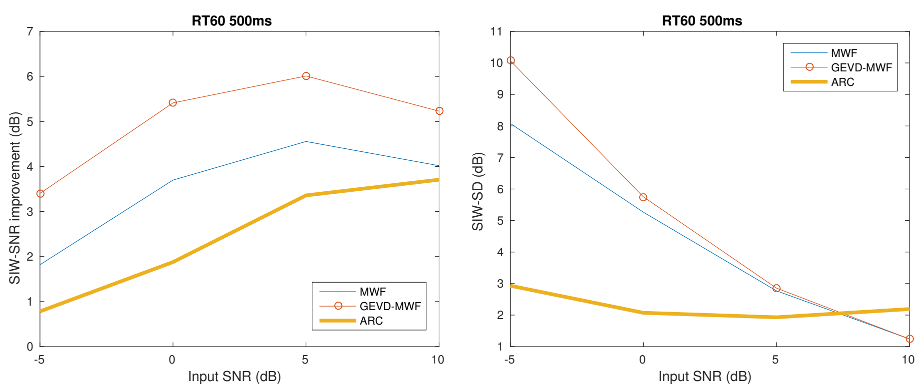

Fig. 2. SIW-SNR & SIW-SD with $RT_{60} = 500ms$.

When the reverbaration time increases (Fig. 2) the performance of the MWF-based algorithms degrades. The SIW-SNR of all three algorithms become comparable and ARC outperforms MWF-based algorithms in terms of SIW-SD at low input SNR.

In noise reduction algorithms there is a trade-off between the quantity of noise removed and the spectral distortion introduced in the speech signal. In MWF-based algorithms it is possible to tune this trade-off explicitly, this is the so-called speech distortion weighted MWF (SDW-MWF) [7, 22]. Spectral distortion can sometimes be perceived as more annoying than additive noise by the listeners and can also decrease the performance of speech recognition systems severely. It is important to be able to design noise reduction algorithms that can limit the amount of spectral distortion they introduce. Therefore, in a second experiment we set the trade-off parameter in MWF-based algorithms such that the MWF and the GEVD-MWF introduce an SIW-SD similar to the SIW-SD introduced by ARC. We then focus on the performance analysis in terms of SIW-SNR. More particularly, for each input SNR the SD-SIW introduced by the MWF, the GEVD-MWF and ARC are averaged over the noise types and the spatial scenarios. The trade-off parameter for the MWF and the GEVD-MWF is then adjusted (for each input SNR) such that every algorithm introduces a comparable amount of SIW-SD.

The SIW-SNR performance of the algorithms is presented in Table 1. ARC outperforms the standard SDW-MWF for which the SIW-SNR decreases quickly as the trade-off parameter favors the limitation of spectral distortion. At low input SNR, the trade-off parameter becomes close to 0, the filter coefficients are then close to 1 in each frequency bands and the input signal is barely affected by the MWF. On the other hand, it has been shown that the GEVD can maintain its SIW-SNR performance to some extent while limiting the spectral distortion introduced [22]. Therefore, even when both algorithms are adjusted to introduce similar SIW-SD, the GEVD-MWF still outperforms ARC.

Noise position	Input SNR	MWF	GEVD-MWF	ARC
30°	$-5\,dB$	0	2.8	0.7
	$0\,dB$	1.15	4.4	1.5
	$5\,dB$	1.4	5.3	2.7
90°	$-5\,dB$	0.1	2.9	0.8
	$0\,dB$	1	5.1	2.3
	$5\,dB$	1.8	5.6	4

Table 1. SIW-SNR improvement performance when the different systems are tuned to introduce similar SIW-SD.

6 Conclusion

As a conclusion, a new method has been proposed to model the noise that is less sensitive to outliers. Despite lower scores than state-of-the-art in terms of SIW-SNR improvement, the ARC algorithm appears to be more robust to the reverberation, that can drastically decrease the performance of algorithms that rely on the estimation of simple correlation matrices (such as the MWF and GEVD-MWF algorithms presented here). In reverberant scenarios, MWF based algorithms then introduce an important amount of SIW-SD, in particular at low input SNR where the noise reduction algorithm has to face signals dominated by noise. This distortion can be perceived very negatively by human listeners and can affect speech recognition system dramatically. It is therefore important to be able to limit it. All three algorithms were thus compared while adjusted to introduce a similar amount of distortion. In the latter case, ARC outperforms the standard MWF algorithm. Future work will include a model that combines both methods GEVD-MWF and ARC in order to get an algorithm exploiting the advantages of both approaches: less spectral distortion and a better SIW-SNR improvement.

Acknowledgments. This work was partly supported by the research programme KAMoulox (ANR-15-CE38-0003-01), EDiSon3D (ANR-13-CORD-0008-01), FBIMATRIX (ANR-16-CE23-0014) funded by ANR, the French State agency for research.

References

1. ANSI: S3. 5-1997, methods for the calculation of the speech intelligibility index. New York: American National Standards Institute 19, 90–119 (1997)
2. Benesty, J., Chen, J., Huang, Y.: Noncausal (frequency-domain) optimal filters. Microphone array signal processing pp. 115–137 (2008)
3. Van den Bogaert, T., Doclo, S., Wouters, J., Moonen, M.: Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. The Journal of the Acoustical Society of America 125(1), 360–371 (2009)
4. Boutsidis, C., Gallopoulos, E.: SVD based initialization: A head start for nonnegative matrix factorization. Pattern Recognition 41(4), 1350–1362 (2008)
5. Cambanis, S., Keener, R., Simons, G.: On α-symmetric multivariate distributions. Journal of Multivariate Analysis 13(2), 213–233 (1983)
6. Doclo, S., Moonen, M.: GSVD-based optimal filtering for single and multimicrophone speech enhancement. IEEE Transactions on Signal Processing 50(9), 2230–2244 (2002)
7. Doclo, S., Spriet, A., Wouters, J., Moonen, M.: Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction. Speech Communication 49(7-8), 636–656 (2007)
8. Duong, N., Vincent, E., Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE/ACM Trans. Audio, Speech, Language Process. 18(7), 1830 –1840 (sept 2010)

9. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. 32(6), 1109–1121 (1984)
10. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. 33(2), 443–445 (1985)
11. Févotte, C., Idier, J.: Algorithms for nonnegative matrix factorization with the β-divergence. Neural computation 23(9), 2421–2456 (2011)
12. Fontaine, M., Liutkus, A., Girin, L., Badeau, R.: Parameterized Wiener filtering for single-channel denoising. In: Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (2017)
13. Greenberg, J., Peterson, P., Zurek, P.: Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. The Journal of the Acoustical Society of America 94(5), 3009–3010 (1993)
14. Hirsch, H., Pearce, D.: The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW) (2000)
15. Leglaive, S., Simsekli, U., Liutkus, A., Badeau, R., Richard, G.: Alpha-stable multichannel audio source separation. In: 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017)
16. Liutkus, A., Badeau, R.: Generalized Wiener filtering with fractional power spectrograms. In: 40th International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 266–270. IEEE (2015)
17. Liutkus, A., Badeau, R., Richard, G.: Gaussian processes for underdetermined source separation. IEEE Trans. Signal Process. 59(7), 3155–3167 (2011)
18. Nugraha, A.A., Liutkus, A., Vincent, E.: Multichannel music separation with deep neural networks. In: 24th European Signal Processing Conference (EUSIPCO), 2016. pp. 1748–1752. IEEE (2016)
19. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210. IEEE (2015)
20. Samoradnitsky, G., Taqqu, M.: Stable non-Gaussian random processes: stochastic models with infinite variance, vol. 1. CRC Press (1994)
21. Sawada, H., Kameoka, H., Araki, S., Ueda, N.: Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization. In: 37th International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 261–264. IEEE (2012)
22. Serizel, R., Moonen, M., Van Dijk, B., Wouters, J.: Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(4), 785–799 (2014)
23. Şimşekli, U., Liutkus, A., Cemgil, A.: Alpha-stable matrix factorization. IEEE Signal Process. Lett. 22(12), 2289–2293 (2015)
24. Şimşekli, U., et al.: Alpha-stable low-rank plus residual decomposition for speech enhancement. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2018)
25. Sun, M., Li, Y., Gemmeke, J.F., Zhang, X.: Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with kullback–leibler divergence. IEEE/ACM Trans. Audio, Speech, Language Process. 23(7), 1233–1242 (2015)

26. Van Veen, B.D., Buckley, K.M.: Beamforming: A versatile approach to spatial filtering. IEEE assp magazine 5(2), 4–24 (1988)
27. Yoshii, K., Itoyama, K., Goto, M.: Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 51–55. IEEE (2016)