Multichannel Audio Modeling with Elliptically Stable Tensor Decomposition

Mathieu Fontaine, Fabian-Robert Stöter, Antoine Liutkus, Umut Şimşekli, Romain Serizel, Roland Badeau

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 [2], 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 [5, 13]. A common approach to speech enhancement is the spectral subtraction method [6]. 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 substracted from 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 .

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 [3]. 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 [22, 19].

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 [16]. Various studies have been carried out on audio modeling using alpha-stable processes [19, 12]. Especially in the TF domain, a generalization of wide-sense stationary (WSS) processes [13] has been established in the $\alpha-$stable case [12] and applied to noise reduction [8]. More precisely, in [20] 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 [20] to the multichannel case. The goal is to develop a Gaussian NMF model for speech that assumes a low-rank structure for speech covariances [5], 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) [17] 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 represented by two tensors: 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 the tensors $\boldsymbol{y}$ 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 [13]. 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 $\boldsymbol{C}^{\boldsymbol{y}}_{ft} \triangleq \boldsymbol{R}_f v_{f,t}$, 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 [5]. 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 $\boldsymbol{R} \triangleq \{\boldsymbol{R}_f\}_f$. Note that the decomposition of v_{ft} is not unique: it is defined up to multiplicative constant.

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 [16, 12] 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 [16, 11]:

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 [19] 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 [2]:

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

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 $\boldsymbol{W},\boldsymbol{H}$:

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

where $\boldsymbol{\Phi}$ is a latent variable and $\log\mathbb{P}\left(\boldsymbol{x},\boldsymbol{\Phi}\mid\boldsymbol{\Theta},\boldsymbol{\sigma}\right)$ is the log-likelihood. As in [20], we propose an EM algorithm. This method aims to minimize an upper-bound of $\mathcal{L}_n\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R}\right) = -\log\mathbb{P}\left(\boldsymbol{x},\boldsymbol{\Phi}\mid\boldsymbol{\Theta},\boldsymbol{\sigma}\right)$. 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 [17]:

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

where $\tilde{\boldsymbol{X}}_{ft} \triangleq \boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^{\star}$ and $.^{\star}$ stands for the Hermitian transposition. A positive auxiliary function $\mathcal{L}_n^{+}\left(\boldsymbol{W},\boldsymbol{H},\boldsymbol{R},\boldsymbol{U},\boldsymbol{V}\right) = \sum_{f,t}\left[\sum_l \frac{\operatorname{tr}\left(\tilde{\boldsymbol{X}}_{ft}\boldsymbol{U}_{lft}\left(\boldsymbol{C}_{lft}^{\boldsymbol{x}|\phi}\right)^{-1}\boldsymbol{U}_{lft}\right)}{w_{fl}h_{lt}} + \frac{\operatorname{tr}\left(\tilde{\boldsymbol{X}}_{ft}\boldsymbol{U}_{lft}^2\right)}{\sigma_f\phi_{ft}} + \log\det\boldsymbol{V}_{ft} + \frac{\det\boldsymbol{C}_{ft}^{\boldsymbol{x}|\phi} - \det\boldsymbol{V}_{ft}}{\det\boldsymbol{V}_{ft}}\right]$ 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 [17]. 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:

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

The form in (14) admits partial derivatives that will be useful as part of a multiplicative update [7] 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:

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

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 [5], further using the trick proposed in [14] to use a weighted update, resulting in:

$$\boldsymbol{R}_f \leftarrow \left(\sum_t v_{ft}\right)^{-1} \times \sum_t \left(\boldsymbol{C}_{ft}^{\boldsymbol{y}\boldsymbol{y}^\star|\boldsymbol{x}}\right), \tag{16}$$

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 [4] 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:

$$\sigma_f \leftarrow \mathbb{M}\left(\| \sum_t \boldsymbol{x}\left(f,t\right) \|^{\alpha/2}\right)^2. \tag{17}$$

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:

$$\overline{\boldsymbol{\Xi}}_{ft} \simeq \frac{1}{I}\sum_{i=1}^{I}\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\varphi_i}\right)^{-1};\ \overline{\boldsymbol{P}}_{ft} \simeq \frac{1}{I}\sum_{i=1}^{I}\left(\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\varphi_i}\right)^{-1}\tilde{\boldsymbol{X}}_{ft}\left(\boldsymbol{C}_{ft}^{\boldsymbol{x}|\varphi_i}\right)^{-1}\right) \tag{18}$$

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

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

– 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* $\boldsymbol{B}_f \in \mathbb{C}^K$ [21]:

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

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 $\boldsymbol{B}_f^{\star}\boldsymbol{y}_{ft} \mid \boldsymbol{x}$:

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

This is solved by taking $\boldsymbol{B}_f$ as the eigenvector associated to the largest eigenvalue of the Hermitian matrix $\frac{1}{T}\sum_t\left(\boldsymbol{C}_{ft}^{\boldsymbol{y}\boldsymbol{y}^{\star}|\boldsymbol{x}}\right)$ [5].

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 The proposed Alpha Residual component. We take $N = 10$ iterations for the EM and pick $\alpha = 1.9$.

MWF The classic multi-channel Wiener filter [5] which assumes Gaussianity for both noise and speech.

GEVD The generalized eigenvalue decomposition [18] is based on a low-rank approximation of the autocorrelation matrix of the speech signal.

5.1 Experimental setup

The corpus for evaluation is made up of mono speech excerpts from Librispeech [15] and three different environmental noises taken from Aurora [10]: babble noise, restaurant and train. A groundtruth voice activity detection (VAD) is used on all three methods.

Mixtures were obtained for two 15 cm separated microphones, with the Roomsimove simulator with room dimensions of $5\times4\times3$ meters and RT60=0 ms and 500 ms. The sources are taken $1\,m$ from the microphones, with different SNR values of -5, 0, 5, $10\,dB$ and an angular distance of 30°or 90°. This results in 48 experiments.

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) [9].

The SIW-SD measure is defined as:

$$\mathrm{SIW-SD} = \sum_i I_i \mathrm{SD}_i \tag{19}$$

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,

$$\mathrm{SD}_i = \frac{1}{(2^{1/6} - 2^{-1/6}) f_i^c} \int_{2^{-1/6} f_i^c}^{2^{1/6} f_i^c} |10 \log_{10} G^y(f)| df \tag{20}$$

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

$$G^y(f) = \frac{P_{\boldsymbol{y}}(f)}{P_{\hat{\boldsymbol{y}}}(f)} \tag{21}$$

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

$$\Delta \mathrm{SNR}_{\mathrm{intellig}} = \sum_i I_i (\mathrm{SNR}_{i,\mathrm{out}} - \mathrm{SNR}_{i,\mathrm{in}}) \tag{22}$$

where $\mathrm{SNR}_{i,\mathrm{out}}$ and $\mathrm{SNR}_{i,\mathrm{in}}$ 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

Results are displayed on Figure 1 and present the SIW-SNR and SIW-SD scores averaged over noise types and spatial scenarios, against the input SNR.

We first investigate the impact of reverberation. While we see that ARC is outperformed by other methods in the anechoic case, we see it is much less sensitive to reverberation and becomes competitive compared to the other algorithms in terms of SIW-SD at low input SNR.

6 Conclusion

We proposed a new method ARC for denoising that is more robust to reverberation than competing approaches, although less effective in the anechoic case. It is based on modeling the speech signal as a Gaussian process and noise as an α-stable sub-Gaussian process. Interestingly, that approach can be combined with existing methods, which could be an interesting avenue for future work.

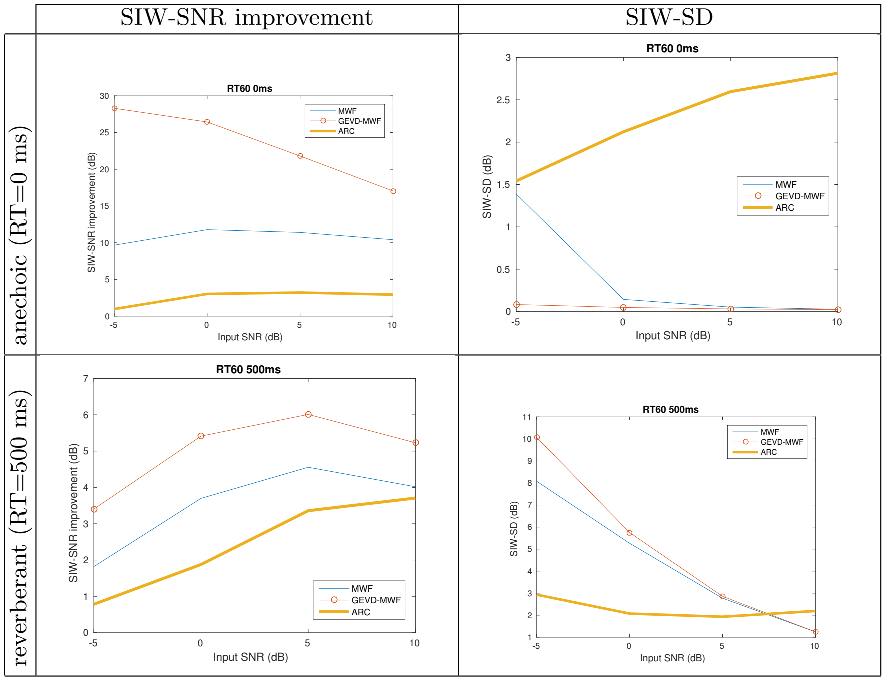

Fig. 1. SIW (left, higher is better) and SNR & SIW-SD (right, lower is better) for: (top) an anechoic scenario and (bottom) a reverberent room.

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. 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)
3. Boutsidis, C., Gallopoulos, E.: SVD based initialization: A head start for nonnegative matrix factorization. Pattern Recognition 41(4), 1350–1362 (2008)
4. Cambanis, S., Keener, R., Simons, G.: On α-symmetric multivariate distributions. Journal of Multivariate Analysis 13(2), 213–233 (1983)
5. 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)
6. 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)

7. Févotte, C., Idier, J.: Algorithms for nonnegative matrix factorization with the β-divergence. Neural computation 23(9), 2421–2456 (2011)
8. 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)
9. 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)
10. 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)
11. 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)
12. 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)
13. Liutkus, A., Badeau, R., Richard, G.: Gaussian processes for underdetermined source separation. IEEE Trans. Signal Process. 59(7), 3155–3167 (2011)
14. 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)
15. 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)
16. Samoradnitsky, G., Taqqu, M.: Stable non-Gaussian random processes: stochastic models with infinite variance, vol. 1. CRC Press (1994)
17. 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)
18. 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)
19. Şimşekli, U., Liutkus, A., Cemgil, A.: Alpha-stable matrix factorization. IEEE Signal Process. Lett. 22(12), 2289–2293 (2015)
20. 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)
21. Van Veen, B.D., Buckley, K.M.: Beamforming: A versatile approach to spatial filtering. IEEE assp magazine 5(2), 4–24 (1988)
22. 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)