Alpha-stable low-rank plus residual decomposition for speech enhancement

Umut Simsekli, Halil Erdogan, Simon Leglaive, Antoine Liutkus, Roland Badeau, Gael Richard

ALPHA-STABLE LOW-RANK PLUS RESIDUAL DECOMPOSITION FOR SPEECH ENHANCEMENT

Umut Şimşekli[1], *Halil Erdoğan*[2], *Simon Leglaive*[1], *Antoine Liutkus*[3], *Roland Badeau*[1], *Gaël Richard*[1]

1: LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
2: Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul, Turkey
3: Inria and LIRMM, Montpellier, France

ABSTRACT

In this study, we propose a novel probabilistic model for separating clean speech signals from noisy mixtures by decomposing the mixture spectrograms into a structured speech part and a more flexible residual part. The main novelty in our model is that it uses a family of heavy-tailed distributions, so called the α-stable distributions, for modeling the residual signal. We develop an expectation-maximization algorithm for parameter estimation and a Monte Carlo scheme for posterior estimation of the clean speech. Our experiments show that the proposed method outperforms relevant factorization-based algorithms by a significant margin.

Index Terms— Alpha-stable distributions, Audio source separation, Speech enhancement, Monte Carlo Expectation-Maximization

1. INTRODUCTION

Speech enhancement is one of the central problems in audio signal processing. The aim in this problem is to recover clean signals after observing noisy mixture signals. It is often formulated as a source separation problem, where the clean speech and the noise are considered as latent sources to be estimated [1].

One of the popular approaches for model-based speech enhancement is based on non-negative matrix factorization (NMF) models [2, 1]. In such approaches, the *time-frequency* representations of the latent sources are modeled in such a way that their power spectral densities (PSD) are assumed to admit a low-rank structure. There have been several extensions to NMF-based speech enhancement approaches, to name a few [3, 4, 5].

One of the main limitations of NMF-based enhancement techniques is that they are usually based on certain Gaussianity assumptions, which turn out to be restrictive for audio signals [6]. As a result, non-Gaussian models have started receiving increasing attention in the audio processing community [7, 8, 9, 10]. The main goal in these approaches is to better capture the variability of audio signals by replacing the Gaussian models with *heavy-tailed* models. It has been shown that the use of such heavy-tailed models can be advantageous for speech enhancement [11, 12, 8].

In this study, we propose a novel probabilistic model for *single-channel, unsupervised* speech enhancement. The proposed model is based on a rather simple assumption that the observed mixture is composed of two components whose statistical properties are significantly different. The *target* signal (i.e. the clean speech) is conventionally modeled in the time-frequency domain by using a centered Gaussian distribution, whose variance admits a low-rank structure. The *residual* signal, however, is not required to have a low-rank structure and is modeled (in the time-frequency domain as well) by using a family of heavy-tailed distributions.

The proposed approach shares similarities with the 'low rank plus sparse' (LRS) decomposition algorithms [13, 5, 14] that decomposes the observed spectra into a low-rank and a sparse part. However, instead of placing explicit sparsity assumptions, we exploit the statistical differences between the two latent signals, which renders our approach to be adapted to more general scenarios as opposed to LRS approaches, which are more suitable transient noise environments. On the other hand, as we are in a probabilistic setting, we can develop a theoretically principled way of obtaining the posterior estimates of the latent signals, whereas the LRS approaches often need to resort to certain heuristics.

Even though its construction would seem simple at a first sight, making inference in the proposed model turns out to be a challenging task. For parameter estimation, we develop a Monte Carlo expectation-maximization (MCEM) algorithm. We further develop a novel 'Wiener-like' filter for estimating the posterior expectations of the latent sources. We evaluate our approach on a challenging speech enhancement problem. Our experiments show that the proposed approach outperforms relevant factorization-based algorithms by a significant margin. We also show that the performance of our approach can be further improved by combining it with an existing speech enhancement algorithm.

2. THE PROPOSED MODEL

Notation: We consider a single-channel observed audio signal, called the *mixture*, and expressed in the short-term Fourier transform (STFT) domain as $\mathbf{X} \equiv \{x_{fn}\}_{f,n} \in \mathbb{C}^{F\times N}$, where $n = 1, \ldots, N$ denotes the time-frames and $f = 1, \ldots, F$ denotes the frequency bands. In this study, we assume that the mixture is the sum of two latent signals $\mathbf{S} \equiv \{s_{fn}\}_{f,n} \in \mathbb{C}^{F\times N}$ and $\mathbf{R} \equiv \{r_{fn}\}_{f,n} \in \mathbb{C}^{F\times N}$. While the first is referred to as the *target*, the second is called the *residual*. The mixture model thus simply becomes: $x_{fn} = s_{fn} + r_{fn}$. In a speech enhancement application, our objective will be to estimate $\mathbf{S}$ and $\mathbf{R}$, after observing $\mathbf{X}$.

Target signal model: The target signal s_{fn} is assumed to feature *redundancies* and *structure*. We take it as a locally stationary Gaussian process [15], with a PSD that obeys an NMF model. This very classical model [16] assumes that all the entries of $\mathbf{S}$ are independent and distributed as:

$$s_{fn} \sim \mathcal{N}_c\big(s_{fn}; 0, \hat{v}_{fn} \triangleq \sum\nolimits_k w_{fk} h_{kn}\big), \qquad (1)$$

This work is partly supported by the French National Research Agency (ANR) as a part of the FBIMATRIX (ANR-16-CE23-0014), and KAMoulox (ANR-15-CE38-0003-01) projects. This research was conducted while Halil Erdoğan was an intern at Télécom ParisTech.

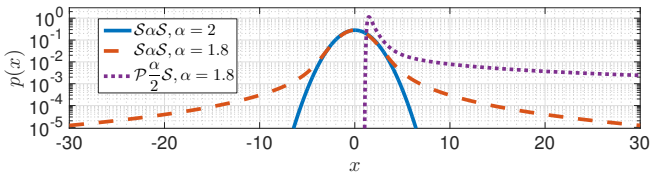

Fig. 1. Approximately computed pdfs of $\mathcal{S}\alpha\mathcal{S}$ and $\mathcal{P}\frac{\alpha}{2}\mathcal{S}$ for $x \in \mathbb{R}$.

where $\mathcal{N}_c$ denotes the isotropic complex Gaussian distribution [17], and $\mathbf{W} \equiv \{w_{fk}\}_{f,k} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \equiv \{h_{kn}\}_{k,n} \in \mathbb{R}_+^{K\times N}$ are the model parameters with $K < \min(F, N)$.

From the perspective of this study, the important feature of this model is that it is both rich, i.e. contains sufficient expressive power for modeling real signals, and it is also *restrictive* at the same time, since the Gaussian distributions have light tails, meaning that it is unlikely for s_{fn} to explore much further than just a few standard deviations $\sqrt{\hat{v}_{fn}}$. Consequently, the target signal should closely match its model expressed in (1).

Residual model: Contrarily to the target signal, the probabilistic model that we develop for the residual signal $\mathbf{R}$ should be very *flexible*, i.e. should have as few parameters as possible, while being extremely *permissive*, meaning that the residual signal is allowed to have large dynamics. The rationale for such a model is that the residual should stand for everything that the target is *not*.

Recent research in audio modeling introduced the α-harmonizable models as good alternatives to their Gaussian counterparts for modeling locally stationary but *heavy-tailed* (impulsive) signals [18, 6]. In short, they amount to simply replacing the isotropic Gaussians by complex isotropic (symmetric) α-stable distributions, denoted as $\mathcal{S}\alpha\mathcal{S}_c$. Just like its Gaussian counterpart, this distribution is parameterized by a scale parameter $\sigma > 0$, but also by a *characteristic exponent* $\alpha \in (0, 2]$, which controls the heaviness of the tails: smaller values of α result in heavier tails. The probability density function (pdf) of an $\mathcal{S}\alpha\mathcal{S}_c$ distribution cannot be written in closed-form except for certain special cases, such as the Gaussian distribution, which appears as a limiting case ($\alpha = 2$). The pdfs of $\mathcal{S}\alpha\mathcal{S}_c$ is illustrated in Fig. 1. The $\mathcal{S}\alpha\mathcal{S}_c$ distributions have already gathered some interest in the audio signal processing literature [19, 6, 8, 20].

The recent research on α-harmonizable models for audio modeling has focused on imposing an NMF structure on the scale parameter [8, 10]. In this study, we follow a different direction and adopt a *marginal model*, where we assume that all r_{fn} follow the following probabilistic model:

$$r_{fn} \sim \mathcal{S}\alpha\mathcal{S}_c\big(r_{fn}; \sigma\big). \tag{2}$$

The common scale parameter $\sigma > 0$ is assumed to be known. The strength of the $\mathcal{S}\alpha\mathcal{S}_c$ distribution is that it has sufficiently large tails that would allow such a simple model to capture the possibly large variations of $\mathbf{R}$. A similar modeling scheme has proved useful in source localization applications [21, 22].

3. PARAMETER ESTIMATION VIA MCEM

Given the observed mixture $\mathbf{X}$, our first goal will be to estimate the 'most likely' factor matrices $\mathbf{W}$ and $\mathbf{H}$, in other words, to obtain the maximum likelihood estimate that is defined as follows:

$$(\mathbf{W}^\star, \mathbf{H}^\star) = \arg\max_{\mathbf{W},\mathbf{H}} \log p(\mathbf{X}|\mathbf{W}, \mathbf{H}). \tag{3}$$

Unfortunately, we cannot attempt to solve this problem by using classical optimization algorithms, since the α-stable pdfs cannot be written in closed-form analytical expressions. Therefore, we follow a similar approach to the ones presented in [23, 8, 24, 25] by coming up with an extended, conditionally Gaussian model that will be equivalent to the proposed model in (1) and (2). We define the extended model by making use of the product properties of the symmetric α-stable densities [26], as follows:

$$\phi_{fn} \sim \mathcal{P}\frac{\alpha}{2}\mathcal{S}\Big(\phi_{fn}; 2\big(\cos\frac{\pi\alpha}{4}\big)^{2/\alpha}\Big) \tag{4}$$

$$r_{fn}|\phi_{fn} \sim \mathcal{N}_c\Big(r_{fn}; 0, \phi_{fn}\sigma\Big), \tag{5}$$

where the law of s_{fn} is still the same as defined in (1). Here, ϕ_{fn} is called the *impulse* variable and it modulates the variance of the conditional distribution of r_{fn} given in (5). Its distribution $\mathcal{P}\frac{\alpha}{2}\mathcal{S}$ is *positive* $\frac{\alpha}{2}$ stable distribution, that is a right-skewed heavy-tailed distribution defined for positive random variables [27], as illustrated in Fig. 1. When marginalized over ϕ_{fn}, this model reduces to the original model defined in (2) [26].

As the sum of two Gaussian random variables is also Gaussian distributed, given ϕ_{fn}, we can directly express the conditional distribution of x_{fn} by combining (1) and (5), given as follows:

$$x_{fn}|\phi_{fn} \sim \mathcal{N}_c\Big(x_{fn}; 0, \phi_{fn}\sigma + \hat{v}_{fn}\Big), \tag{6}$$

where the prior of ϕ_{fn} is given in (4). With the new formulation of the model given in (6), we can treat the impulse variables ϕ_{fn} as latent variables and develop an expectation-maximization (EM) algorithm, which iteratively maximizes a lower-bound to the log-likelihood $\log p(\mathbf{X}|\mathbf{W}, \mathbf{H})$. The EM algorithm iteratively computes the following steps:

$$\text{E-Step:} \qquad Q_t(\mathbf{W}, \mathbf{H}) = \mathbb{E}[\log p(\mathbf{X}, \mathbf{\Phi}|\mathbf{W}, \mathbf{H})], \tag{7}$$

$$\text{M-Step:} \qquad (\mathbf{W}^{(t)}, \mathbf{H}^{(t)}) = \arg\max_{\mathbf{W},\mathbf{H}} Q_t(\mathbf{W}, \mathbf{H}), \tag{8}$$

where $\mathbf{\Phi} \equiv \{\phi_{fn}\}_{fn} \in \mathbb{R}_+^{F\times N}$, t denotes the iteration number, and the expectation is taken with respect to the posterior distribution of $\mathbf{\Phi}$, i.e. $p(\mathbf{\Phi}|\mathbf{X}, \mathbf{W}^{(t-1)}, \mathbf{H}^{(t-1)})$.

E-Step: By using (6), we observe that the E-step reduces to the computation the following expression:

$$Q_t(\cdot) =^+ -\sum_{fn}\Big(\mathbb{E}[\log(\phi_{fn}\sigma + \hat{v}_{fn})] + \mathbb{E}\Big[\frac{|x_{fn}|^2}{\phi_{fn}\sigma + \hat{v}_{fn}}\Big]\Big), \tag{9}$$

where $=^+$ denotes equality up to an additive constant that does not depend on either $\mathbf{W}$ or $\mathbf{H}$. However, computing this lower-bound is intractable since the required expectations do not admit an analytical expression. Therefore, we need to resort to approximate algorithms.

We now focus on (9) and attempt to obtain a simplified, alternative bound. As $x \longmapsto \log(x)$ is a concave function, we can apply Jensen's inequality to the first term in (9) and obtain an alternative lower-bound to the log-likelihood, given as follows:

$$\begin{aligned} Q_t(\cdot) &\geq -\sum_{fn}\Big(\log\left(\sigma\mathbb{E}\left[\phi_{fn}\right] + \hat{v}_{fn}\right) + \mathbb{E}\left[\frac{|x_{fn}|^2}{\phi_{fn}\sigma + \hat{v}_{fn}}\right]\Big) \\ &\triangleq -L_t(\mathbf{W}, \mathbf{H}). \end{aligned} \tag{10}$$

Since $-L_t$ is still a valid lower-bound, the theoretical guarantees of the EM algorithm still hold in this modified scheme. Even though L_t is still intractable due to the required expectations, it contains simpler terms, which will facilitate the resulting algorithm.

M-Step: By using the definition of the new lower-bound given in (10), we modify the M-step so that it aims to solve the following optimization problem:

$$(\mathbf{W}^{(t)}, \mathbf{H}^{(t)}) = \underset{\mathbf{W},\mathbf{H}}{\arg\min}\, L_t(\mathbf{W}, \mathbf{H}). \tag{11}$$

We will now develop a gradient-based algorithm for solving the optimization problem given in (11). By assuming that integration with respect to ϕ_{fn} and derivation with respect to w_{fk} can be interchangeable, we can write the partial derivatives of $L_t(\mathbf{W}, \mathbf{H})$ with respect to $\mathbf{W}$ as follows:

$$\frac{\partial L_t(\cdot)}{\partial w_{fk}} = \underbrace{\sum_n \frac{h_{kn}}{\sigma \mathbb{E}\left[\phi_{fn}\right] + \hat{v}_{fn}}}_{\mathcal{A}_1 \geq 0} - \underbrace{\sum_n \mathbb{E}\left[\frac{|x_{fn}|^2 h_{kn}}{(\phi_{fn}\sigma + \hat{v}_{fn})^2}\right]}_{\mathcal{A}_2 \geq 0}.$$

Here, we observe that the partial derivative is the difference of two non-negative terms, $\mathcal{A}_1$ and $\mathcal{A}_2$. Therefore, we can develop a multiplicative gradient descent algorithm [16, 28], that consists in updating w_{fk} by multiplying its current value by the ratio of the negative ($-\mathcal{A}_2$) and positive parts ($\mathcal{A}_1$) of its partial derivative: $w_{fk} \leftarrow w_{fk} \times (\mathcal{A}_2/\mathcal{A}_1)$. In matrix form, the multiplicative update rule for $\mathbf{W}$ is given as follows:

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left((\bar{\mathbf{P}} \circ \mathbf{V})\mathbf{H}^\top\right) \oslash \left((\sigma\bar{\mathbf{\Phi}} + \hat{\mathbf{V}})^{\circ -1}\mathbf{H}^\top\right) \tag{12}$$

where $\mathbf{A} \odot \mathbf{B}$ and $\mathbf{A} \oslash \mathbf{B}$ denote element-wise product and division of two matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A}^{\circ -1}$ denotes element-wise power, i.e. $[\mathbf{A}^{\circ -1}]_{ij} = 1/[A]_{ij}$, and $\top$ denotes the matrix transpose. Here, we have also defined $[\mathbf{V}]_{fn} = |x_{fn}|^2$, $[\hat{\mathbf{V}}]_{fn} = \hat{v}_{fn}$, $[\bar{\mathbf{\Phi}}]_{fn} = \bar{\phi}_{fn} \triangleq \mathbb{E}\left[\phi_{fn}\right]$ and $[\bar{\mathbf{P}}]_{fn} = \bar{p}_{fn} \triangleq \mathbb{E}\left[1/(\phi_{fn}\sigma + \hat{v}_{fn})^2\right]$. By using the same approach, we obtain the following update rule for $\mathbf{H}$:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\mathbf{W}^\top(\bar{\mathbf{P}} \circ \mathbf{V})\right) \oslash \left(\mathbf{W}^\top(\sigma\bar{\mathbf{\Phi}} + \hat{\mathbf{V}})^{\circ -1}\right). \tag{13}$$

In order to obtain a solution to the problem in (11), we run these multiplicative update rules until convergence at each EM iteration t.

3.1. Estimating the Expectations via MCMC

Even though the multiplicative update rules given in (12) and (13) provide us a principled approach to solve the optimization problem given in (11), unfortunately, they cannot be directly used in practice since the terms $\bar{\phi}_{fn}$ and $\bar{p}_{fn}$ cannot be computed analytically.

In this study, we develop a Markov Chain Monte Carlo (MCMC) algorithm, namely the Metropolis-Hastings (MH) algorithm [29] for approximating the intractable expectations by using sample averages. It consist in computing:

$$\bar{\phi}_{fn} \approx \frac{1}{M}\sum_{m=1}^{M} \varphi_{fn}^{(t,m)}, \;\; \bar{p}_{fn} \approx \frac{1}{M}\sum_{m=1}^{M} \frac{1}{(\varphi_{fn}^{(t,m)}\sigma + \hat{v}_{fn})^2}, \tag{14}$$

where m denotes the iteration number of the MH algorithm and $\varphi_{fn}^{(t,m)}$ denotes the samples that are (asymptotically) drawn from the posterior distribution $p(\mathbf{\Phi}|\mathbf{X}, \mathbf{W}^{(t-1)}, \mathbf{H}^{(t-1)})$.

At the m-th iteration of the MH algorithm, we generate the sample $\varphi_{fn}^{(t,m)}$ in two steps. In the first step, we generate a sample φ'_{fn} from the prior distribution $\varphi'_{fn} \sim \mathcal{P}\frac{\alpha}{2}\mathcal{S}\big(2(\cos\frac{\pi\alpha}{4})^{2/\alpha}\big)$ by using the algorithm in [30]. In the second step, we compute an acceptance probability that is defined as follows:

$$\text{acc}(\varphi_{fn} \rightarrow \varphi'_{fn}) = \min\Big\{1, \frac{\mathcal{N}_c(x_{fn}; 0, \varphi'_{fn}\sigma + \hat{v}_{fn})}{\mathcal{N}_c(x_{fn}; 0, \varphi_{fn}\sigma + \hat{v}_{fn})}\Big\}. \tag{15}$$

Algorithm 1: MCEM for maximum likelihood estimation.

input: $\mathbf{W}^{(0)}$,$\mathbf{H}^{(0)}$, σ, T, M
for $t = 1, \cdots, T$ **do**
  // E-Step
  for $m = 1, \cdots, M$ **do**
    Draw $\varphi_{fn}^{(t,m)}$ via Metropolis-Hastings (Eq. 15)
  Approximately compute $\bar{\mathbf{\Phi}}$, $\bar{\mathbf{P}}$ (Eq. 14)
  // M-Step
  Set $\mathbf{W} = \mathbf{W}^{(t-1)}$ and $\mathbf{H} = \mathbf{H}^{(t-1)}$
  while *not converged* **do**
    Update $\mathbf{W}$ (Eq. 12)
    Update $\mathbf{H}$ (Eq. 13)
  Set $\mathbf{W}^{(t)} = \mathbf{W}$ and $\mathbf{H}^{(t)} = \mathbf{H}$

We then draw a uniform random number u in $[0, 1]$. If $u < \text{acc}(\varphi_{fn}^{(t,m-1)} \rightarrow \varphi'_{fn})$, we set $\varphi_{fn}^{(t,m)} = \varphi'_{fn}$ (acceptance), otherwise we set $\varphi_{fn}^{(t,m)} = \varphi_{fn}^{(t,m-1)}$ (rejection). After replacing $\bar{\mathbf{\Phi}}$ and $\bar{\mathbf{P}}$ in (12) and (13) with their approximations, we obtain the ultimate MCEM algorithm, as given in Algorithm 1.

4. SOURCE RECONSTRUCTION

As we are in a source separation context, our ultimate purpose is to estimate the source signals given the observations $\mathbf{X}$ and the model parameters $\mathbf{W}$, $\mathbf{H}$, σ. After obtaining the estimates $\mathbf{W}^\star, \mathbf{H}^\star$ via the MCEM algorithm, we are interested in the Minimum Mean Squared Error (MMSE) estimates of the sources s_{fn} and r_{fn}. The MMSE can be expressed as a posterior expectation, given as follows:

$$\hat{s}_{fn} \triangleq \mathbb{E}\left[s_{fn}\right]_{p(s_{fn}|x_{fn}, \mathbf{W}^\star, \mathbf{H}^\star)} \tag{16}$$

$$= \mathbb{E}\left[\mathbb{E}\left[s_{fn}\right]_{p(s_{fn}|x_{fn}, \phi_{fn}, \mathbf{W}^\star, \mathbf{H}^\star)}\right]_{p(\phi_{fn}|x_{fn}, \mathbf{W}^\star, \mathbf{H}^\star)} \tag{17}$$

$$= x_{fn}\mathbb{E}\left[\hat{v}_{fn}/(\phi_{fn}\sigma + \hat{v}_{fn})\right]_{p(\phi_{fn}|x_{fn}, \mathbf{W}^\star, \mathbf{H}^\star)} \tag{18}$$

$$\triangleq x_{fn}\mu_{fn}, \tag{19}$$

where $\mathbb{E}[f(x)]_{p(x)} = \int f(x)p(x)dx$ and we have used the law of total expectation in (17). Here, the matrix $\mathbf{M} \equiv \{\mu_{fn}\}_{f,n}$ performs a 'soft-masking' operation on $\mathbf{X}$, similar to classical Wiener filtering. By following the same approach, we obtain $\hat{r}_{fn} \triangleq \mathbb{E}\left[r_{fn}\right]_{p(r_{fn}|x_{fn}, \mathbf{W}^\star, \mathbf{H}^\star)} = x_{fn}(1 - \mu_{fn}) = x_{fn} - \hat{s}_{fn}$.

Perhaps not surprisingly, the expectations that are required for computing the soft-masking matrix $\mathbf{M}$ are intractable, similar to the previous cases. However, we can easily approximate these expectations by using the same MH algorithm as described in Section 3.1.

5. EXPERIMENTS

We evaluate the proposed method on a speech enhancement task, where the aim is to recover the clean speech signal after observing a noisy mixture signal. We conduct our experiments on the NOIZEUS noisy speech corpus [31]. This dataset contains 30 sentences that are uttered by 6 speakers (3 male and 3 female) at an 8 kHz sampling rate. These sentences are artificially corrupted by using 8 different real noise signals that are collected from challenging acoustic scenes (airport, babble, car, exhibition hall, restaurant, street, train, train-station) at 4 different signal-to-noise ratio (SNR) levels. We analyze

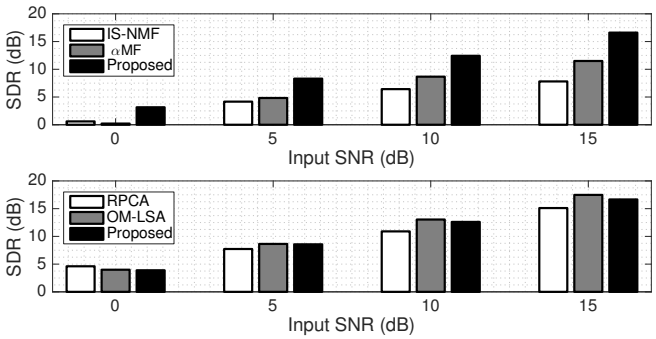

Fig. 2. Evaluation results of the proposed approach on speech enhancement. Top: comparison with semi-supervised methods, bottom: comparison with OM-LSA and RPCA.

the signals by using their STFTs with a Hamming window of length 512 samples and 75% overlap.

We compare our approach with three natural competitors: Itakura-Saito (IS) NMF [2], α-stable matrix factorization (αMF) [8], and a LRS approach based on robust principal component analysis (RPCA) [13]. We also compare our approach with one of the most effective methods for the noise environments that are considered in this study, the optimally-modified log-spectral-amplitude (OM-LSA) algorithm [32]. For evaluating the quality of the estimates we resort to the commonly-used signal-to-distortion ratio (SDR) that is computed with BSS_{EVAL} version 3.0 [33]. For perceptual evaluation, we provide audio samples in [34].

In our first set of experiments, we compare the proposed approach with IS-NMF and αMF. Since both of these competitors are semi-supervised algorithms, they need to be trained on a subset of the dataset and tested on the rest of the dataset. Here, we consider the same setting as explained in [8], where both models are trained on the first 20 clean speech signals (2 female and 2 male speakers) and tested on the remaining 80 different noisy mixtures (corresponding to 10 clean speech signals corrupted with 8 different noise signals). For a fair comparison, we also evaluate the proposed algorithm on the same test data. For the rank of the factorization K, we have investigated values ranging from 1 to 40, and observed that the results do not improve substantially for $K \geq 10$, therefore, we set $K = 10$. We have also investigated the choice of α and observed that the best results are obtained when $\alpha \in [1.5, 1.9]$, and the results start degrading when α is chosen outside of this range. We therefore fix $\alpha = 1.9$. In this setting, we set the outer MCEM iterations to $T = 40$ and the number of MCMC iterations to $M = 40$. At each MCMC run we discard the first 30 samples as the burn-in period, and use the last 10 samples to approximate the expectations. For obtaining the MMSE speech estimates, we run the MH algorithm for 100 iterations, where we discarded the first 40 samples as burn-in. In our experiments, we have observed that the optimal value of σ differs for each input SNR. Accordingly, we choose $\sigma = \{3.4, 0.4, 0.1, 0.02\}$, for input SNRs $\{0, 5, 10, 15\}$ dB, respectively, indicating that the gain of the residual will be lower for higher SNRs.

The results of the first set of experiments are shown in Fig. 2(top)[1]. Even though the proposed approach is completely unsupervised, the results show that it outperforms IS-NMF and αMF by a significant margin. On average, our approach provides an SDR improvement of 5.38 dB when compared to the semi-supervised IS-NMF, and 3.82 dB SDR improvement when compared to the semi-supervised αMF. In addition, our approach requires only $K = 10$ columns in $\mathbf{W}$ and requires fewer iterations since it converges rapidly due to smallness of the parameter space. On the other hand, IS-NMF and αMF require much larger number of columns in their dictionaries in order to be able to capture the properties of a large speech corpus. Therefore, our approach also provides a significant computational advantage over the two competitors.

Table 1. Evaluation results (SDR in dB) of the combination of proposed method combined with OM-LSA.

Method \ SNR	0 dB	5 dB	10 dB	15 dB
OM-LSA	3.99	8.64	13.04	17.48
Combined	**5.02**	**9.30**	**13.35**	**17.57**

In our second set of experiments, we compare our approach to OM-LSA and RPCA[2] on the whole dataset, i.e. 240 noisy mixtures for each input SNR. The results are given in Fig. 2(bottom). We observe that, our method and OM-LSA outperform RPCA except when the input SNR is 0 dB. The proposed approach yields 1.5 dB improvement over RPCA when the input SNR is 15 dB. The results also show that for 0 and 5 dB input SNR, the proposed algorithm and OM-LSA perform similarly, whereas OM-LSA provides a slight improvement when the input SNR is 10 and 15 dB. These results are strongly encouraging since we see that our approach can perform almost as well as OM-LSA, even though OM-LSA exploits the temporal structure of the clean speech, whereas our approach is solely based on some statistical hypotheses and does not take into account the temporal information. Given the fact that IS-NMF-based approaches might outperform OM-LSA when they incorporate temporal information [3, 4], we can conclude that our approach has a strong potential for speech enhancement since it is already yielding comparable performance to OM-LSA.

Encouraged by these results, in our last set of experiments we extend our approach by combining it with OM-LSA. More precisely, instead of modeling the target signal s_{fn} as a Gaussian whose variance is decomposed via NMF, we directly set $\hat{v}_{fn}$ to the PSD of the OM-LSA's speech estimate. Since all the parameters are known in this setting, we can directly obtain the MMSE estimates by using the approach described in Section 4. The results are given in Table 5. We observe that this scheme consistently improves upon OM-LSA in all cases, where the improvement is more prominent for lower input SNRs. These results validate the use of α-stable distributions for modeling the residual signals in a speech enhancement task.

6. CONCLUSION

We proposed a novel probabilistic model based on the heavy-tailed α-stable distributions for separating clean speech from noisy mixtures. We developed MCMC-based algorithms for inference and parameter estimation. Our results showed that the proposed method outperforms relevant algorithms by a significant margin. We also showed that the performance of our method can be further improved by combining it existing speech enhancement algorithms.

[1] As the experimental setting is identical to [8], we directly use the evaluation results that are reported in [8].

[2] We tune the parameters of RPCA and report the best results.

7. REFERENCES

[1] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE TASLP*, vol. 21, no. 10, pp. 2140–2151, 2013.

[2] 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.

[3] C. Févotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *ICASSP*, 2013, pp. 3158–3162.

[4] U. Şimşekli, J. Le Roux, and J.R. Hershey, "Non-negative source-filter dynamical system for speech enhancement," in *ICASSP*, 2014, pp. 6206–6210.

[5] Z. Chen and D. P. W. Ellis, "Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition," in *WASPAA*. IEEE, 2013, pp. 1–4.

[6] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *ICASSP*, 2015.

[7] E. E. Kuruoglu, *Signal processing in α-stable noise environments: a least lp-norm approach*, Ph.D. thesis, University of Cambridge, 1999.

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

[9] 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 *ICASSP*, 2016, pp. 51–55.

[10] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *WASPAA*, 2015.

[11] N. Sasaoka, K. Ono, and Y. Itoh, "Speech enhancement based on 4th order cumulant backward linear predictor for impulsive noise," in *ICSP*. IEEE, 2012, vol. 1, pp. 127–131.

[12] F. Deng, C. Bao, and W. B. Kleijn, "Sparse hidden Markov models for speech enhancement in non-stationary noise environments," *IEEE/ACM TASLP*, vol. 23, no. 11, pp. 1973–1987, 2015.

[13] P. S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *ICASSP*. IEEE, 2012, pp. 57–60.

[14] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence," *IEEE/ACM TASLP*, vol. 23, no. 7, pp. 1233–1242, 2015.

[15] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE TSP*, vol. 59, no. 7, pp. 3155 –3167, 2011.

[16] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[17] R. Gallager, "Circularly symmetric complex gaussian random vectors - a tutorial," Tech. Rep., Massachusetts Institute of Technology, 2008.

[18] N. Bassiou, C. Kotropoulos, and E. Koliopoulou, "Symmetric α-stable sparse linear regression for musical audio denoising," in *ISPA*. IEEE, 2013, pp. 382–387.

[19] C. Nikias and M. Shao, *Signal processing with alpha-stable distributions and applications*, Wiley-Interscience, 1995.

[20] M. Fontaine, A. Liutkus, L. Girin, and R. Badeau, "Explaining the parameterized Wiener filter with alpha-stable processes," in *WASPAA*, 2017.

[21] M. Fontaine, C. Vanwynsberghe, A. Liutkus, and R. Badeau, "Sketching for nearfield acoustic imaging of heavy-tailed sources," in *LVA/ICA*. Springer, 2017, pp. 80–88.

[22] M. Fontaine, C. Vanwynsberghe, A. Liutkus, and R. Badeau, "Scalable source localization with multichannel alpha-stable distributions," in *EUSIPCO*, 2017.

[23] S. Godsill and E. E. Kuruoglu, "Bayesian inference for time series with heavy-tailed symmetric α-stable noise processes," *Heavy Tails*, pp. 3–5, 1999.

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

[25] M. Jas, T. D. La Tour, U. Şimşekli, and A. Gramfort, "Learning the morphology of brain signals using alpha-stable convolutional sparse coding," in *NIPS*, 2017.

[26] G. Samorodnitsky and M. S. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*, vol. 1, CRC press, 1994.

[27] P. Magron, R. Badeau, and A. Liutkus, "Lévy NMF for robust nonnegative source separation," in *WASPAA*, 2017.

[28] U. Şimşekli, T. Virtanen, and A. T. Cemgil, "Non-negative tensor factorization models for Bayesian audio processing," *DSP*, vol. 47, pp. 178–191, 2015.

[29] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, 2008.

[30] J. M. Chambers, C. L. Mallows, and B. W. Stuck, "A method for simulating stable random variables," *Journal of the American statistical association*, vol. 71, no. 354, pp. 340–344, 1976.

[31] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, no. 7, pp. 588–601, 2007.

[32] I. Cohen, "On speech enhancement under signal presence uncertainty," in *ICASSP*. IEEE, 2001, vol. 1, pp. 661–664.

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

[34] U Şimşekli, H. Erdoğan, S. Leglaive, A. Liutkus, R. Badeau, and G. Richard, "Website for audio samples," `http://perso.telecom-paristech.fr/~simsekli/icassp2018`.