Alpha-stable low-rank plus residual decomposition for speech enhancement
Umut Şimşekli, Halil Erdogan, Simon Leglaive, Antoine Liutkus, Roland Badeau, Gael Richard

To cite this version:

HAL Id: hal-01714909
https://inria.hal.science/hal-01714909
Submitted on 22 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 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—α-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-time Fourier transform (STFT) domain as $X \equiv \{x_{fn}\}_{f,n} \in \mathbb{C}^{F \times N}$, 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 $S \equiv \{s_{fn}\}_{f,n} \in \mathbb{C}^{F \times N}$ and $R \equiv \{r_{fn}\}_{f,n} \in \mathbb{C}^{F \times N}$. 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 $S$ and $R$, after observing $X$.

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 $S$ are independent and distributed as:

$$s_{fn} \sim \mathcal{N}_c(0, \hat{v}_{fn} \triangleq \sum_k w_{fk} h_{kn}),$$

(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.
Given the observed mixture $X$, our first goal will be to estimate the ‘most likely’ factor matrices $W$ and $H$, in other words, to obtain the maximum likelihood estimate that is defined as follows:

$$Q_t(\Theta) = \sum_{fn} \left( \mathbb{E}[\log(p(\phi_{fn}, \sigma + \hat{v}_{fn})) + \mathbb{E}\left[ \frac{|x_{fn}|^2}{\phi_{fn}\sigma + \hat{v}_{fn}} \right] \right), \quad (9)$$

where $x_{fn}$ are the observed data, $\phi_{fn}$ are the impulse variables, and $\sigma + \hat{v}_{fn}$ are the noise terms. The EM algorithm iteratively computes the following steps:

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

M-Step:  
$$\Phi = \arg\max_{W, H} Q_t(W, H), \quad (8)$$

where $\Phi = \{ \phi_{fn} \}_{fn} \in \mathbb{R}^{F \times T}$, $t$ denotes the iteration number, and the expectation is taken with respect to the posterior distribution of $\Phi$, i.e. $p(\Phi | X, W^{(t-1)}, H^{(t-1)})$. The common scale parameter $\sigma > 0$ is assumed to be known. The strength of the $\mathcal{SO}_\alpha$ distribution is that it has sufficiently large tails that would allow such a simple model to capture the possibly large variations of $R$. A similar modeling scheme has proved useful in source localization applications [21,22].

3. PARAMETER ESTIMATION VIA MCEM

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

$$Q_t(\Theta) = \arg\max_{W, H} \log p(X | W, H). \quad (3)$$

Unfortunately, we cannot attempt to solve this problem by using classical optimization algorithms, since the $\alpha$-stable pdfs cannot be written in closed-form analytical expressions. Therefore, we follow a similar approach to the ones presented in [23,24,25] 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 $\alpha$-stable densities [26], as follows:

$$\phi_{fn} \sim \mathcal{P}_\alpha S(\phi_{fn}; \cos \frac{\pi \alpha}{4})^{2/\alpha} \quad \text{(4)}$$

$$r_{fn} | \phi_{fn} \sim \mathcal{N}_r(r_{fn}; 0, \phi_{fn} \sigma), \quad \text{(5)}$$

where the law of $s_{fn}$ is still the same as defined in (1). Here, $\phi_{fn}$ is called the impulse variable and it modulates the variance of the conditional distribution of $r_{fn}$ given in (5). Its distribution $\mathcal{P}_\alpha S$ is positive $\alpha$ stable distribution, that is a right-skewed heavy-tailed distribution defined for positive random variables [27], as illustrated in Fig. 1. When marginalized over $\phi_{fn}$, this model reduces to the original model defined in (1) [29].

As the sum of two Gaussian random variables is also Gaussian distributed, given $\phi_{fn}$, we can directly express the conditional distribution of $x_{fn}$ by combining (1) and (2), given as follows:

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

where the prior of $\phi_{fn}$ is given in (4). With the new formulation of the model given in [29], we can treat the impulse variables $\phi_{fn}$ as latent variables and develop an expectation-maximization (EM) algorithm, which iteratively maximizes a lower-bound to the log-likelihood $\log p(X | W, H)$. The EM algorithm iteratively computes the following steps:

3. PARAMETER ESTIMATION VIA MCEM

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

$$(W^*, H^*) = \arg\max_{W, H} \log p(X | W, H). \quad (3)$$

Unfortunately, we cannot attempt to solve this problem by using classical optimization algorithms, since the $\alpha$-stable pdfs cannot be written in closed-form analytical expressions. Therefore, we follow a similar approach to the ones presented in [23,24,25] 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 $\alpha$-stable densities [26], as follows:

$$\phi_{fn} \sim \mathcal{P}_\alpha S(\phi_{fn}; \cos \frac{\pi \alpha}{4})^{2/\alpha} \quad \text{(4)}$$

$$r_{fn} | \phi_{fn} \sim \mathcal{N}_r(r_{fn}; 0, \phi_{fn} \sigma), \quad \text{(5)}$$

where the law of $s_{fn}$ is still the same as defined in (1). Here, $\phi_{fn}$ is called the impulse variable and it modulates the variance of the conditional distribution of $r_{fn}$ given in (5). Its distribution $\mathcal{P}_\alpha S$ is positive $\alpha$ stable distribution, that is a right-skewed heavy-tailed distribution defined for positive random variables [27], as illustrated in Fig. 1. When marginalized over $\phi_{fn}$, this model reduces to the original model defined in (1) [29].

As the sum of two Gaussian random variables is also Gaussian distributed, given $\phi_{fn}$, we can directly express the conditional distribution of $x_{fn}$ by combining (1) and (2), given as follows:

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

where the prior of $\phi_{fn}$ is given in (4). With the new formulation of the model given in [29], we can treat the impulse variables $\phi_{fn}$ as latent variables and develop an expectation-maximization (EM) algorithm, which iteratively maximizes a lower-bound to the log-likelihood $\log p(X | W, H)$. The EM algorithm iteratively computes the following steps:

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

M-Step:  
$$\Phi = \arg\max_{W, H} Q_t(W, H), \quad (8)$$

where $\Phi = \{ \phi_{fn} \}_{fn} \in \mathbb{R}^{F \times T}$, $t$ denotes the iteration number, and the expectation is taken with respect to the posterior distribution of $\Phi$, i.e. $p(\Phi | X, W^{(t-1)}, H^{(t-1)})$. The common scale parameter $\sigma > 0$ is assumed to be known. The strength of the $\mathcal{SO}_\alpha$ distribution is that it has sufficiently large tails that would allow such a simple model to capture the possibly large variations of $R$. A similar modeling scheme has proved useful in source localization applications [21,22].
Given in (11), unfortunately, they cannot be directly used in practice. To provide us a principled approach to solve the optimization problem, we run these multiplicative update rules given in (12) and (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 update wfk can be inter-changeable, we can write the partial derivatives of Lt(W, H) with respect to W as follows:

$$\frac{\partial L_t}{\partial w_{fk}} = \sum_n \sigma \frac{h_{kn}}{\phi_n} + \delta_{fn} - \sum_n \mathbb{E} \left[ \frac{x_{fn}^2 h_{kn}}{\phi_n \sigma + \delta_{fn}} \right].$$

Here, we observe that the partial derivative is the difference of two non-negative terms, A1 and A2. Therefore, we can develop a multiplicative gradient descent algorithm [16, 28], that consists in updating wfk by multiplying its current value by the ratio of the negative (A2) and positive parts (A1) of its partial derivative: w′fk ← wfk × (A2/A1).

In matrix form, the multiplicative update rule for W is given as follows:

$$W \leftarrow W \odot \left( (\Phi \odot V)H^\top \right) \odot \left( (\sigma \Phi + \hat{V})^{\odot 1} H^\top \right)$$

where A ⊙ B and A⊙B denote element-wise product and division of two matrices A and B, A⊙1 denotes element-wise power, i.e. A⊙1 lij = 1/|Aij|, and T denotes the matrix transpose. Here, we have also defined [V](fn) = [x fn]2, [V](fn) = [ϕ n] = [ϕ n] ⊆ E[ϕ n], and [P](fn) = p(fn) ⊆ E[1/ϕ nσ + δ(fn)]. By using the same approach, we obtain the following update rule for H:

$$H \leftarrow H \odot \left( W^\top (\Phi \odot V) \right) \odot \left( W^\top (\sigma \Phi + \hat{V})^{\odot 1} \right).$$

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

### 4. SOURCE RECONSTRUCTION

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

$$\hat{s}_f n \approx \mathbb{E} [s_f n | x_f n, W^*, H^*] \quad (16)$$

$$= \mathbb{E} \left[ \mathbb{E} [s_f n | x_f n, \phi_f n, W^*, H^*] \right]_{p(\phi_f n | x_f n, W^*, H^*)} \quad (17)$$

$$= x_f n \mathbb{E} [\hat{v}_f n | (\phi_f n + \hat{v}_f n)]_{p(\phi_f n | x_f n, W^*, H^*)} \quad (18)$$

$$\approx x_f n \mu_f n,$$

where 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 M ≡ {μf n}f n performs a ‘soft-masking’ operation on X, similar to classical Wiener filtering. By following the same approach, we obtain rfn = \mathbb{E} [\tau f n | p(\tau f n | x_f n, W^*, H^*)] = x_f n (1 − μf n) = x_f n − \hat{s}_f n.

Perhaps not surprisingly, the expectations that are required for computing the soft-masking matrix 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
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].

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

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR</th>
<th>0 dB</th>
<th>5 dB</th>
<th>10 dB</th>
<th>15 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>OM-LSA</td>
<td>3.99</td>
<td>8.64</td>
<td>13.04</td>
<td>17.48</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>5.02</td>
<td>9.30</td>
<td>13.35</td>
<td>17.57</td>
<td></td>
</tr>
</tbody>
</table>

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.

In our second set of experiments, we compare our approach to OM-LSA and RPCA [34] 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], 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 \( s_{\mathbf{W}} \) as a Gaussian whose variance is decomposed via NMF, we directly set \( v_{\mathbf{W}} \) 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 3. The results are given in Table 3. 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 [5].

2 We tune the parameters of RPCA and report the best results.
7. REFERENCES


