An Inverse-Gamma Source Variance Prior with Factorized Parameterization for Audio Source Separation

Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Sharon Gannot, Radu Horaud

AN INVERSE-GAMMA SOURCE VARIANCE PRIOR WITH FACTORIZED PARAMETERIZATION FOR AUDIO SOURCE SEPARATION

Dionyssos Kounades-Bastian[1], *Laurent Girin*[1,2], *Xavier Alameda-Pineda*[3], *Sharon Gannot*[4], *Radu Horaud*[1]

[1]INRIA Grenoble Rhône-Alpes, [2]GIPSA-Lab, Univ. Grenoble Alpes
[3]University of Trento, [4]Faculty of Engineering, Bar-Ilan University

ABSTRACT

In this paper we present a new statistical model for the power spectral density (PSD) of an audio signal and its application to multichannel audio source separation (MASS). The source signal is modeled with the local Gaussian model (LGM) and we propose to model its variance with an inverse-Gamma distribution, whose scale parameter is factorized as a rank-1 model. We discuss the interest of this approach and evaluate it in a MASS task with underdetermined convolutive mixtures. For this aim, we derive a variational EM algorithm for parameter estimation and source inference. The proposed model shows a benefit in source separation performance compared to a state-of-the-art LGM NMF-based technique.

Index Terms— Audio modeling, local Gaussian model, PSD model, audio source separation.

1. INTRODUCTION

For the past decade, the statistical modeling of audio signals in the time-frequency (TF) domain has been thoroughly investigated. Among the proposed models, the local Gaussian model (LGM) [1] has become very popular because, among other reasons, it can be naturally coupled with models of the signal power spectral density (PSD) (which identifies with the signal variance at each TF bin for a zero-mean signal). An important example is the use of non-negative matrix factorization (NMF), which imposes a low-rank structure on the PSD matrix, namely the product of a spectral pattern matrix and a temporal activation matrix [2]. Intuitively, NMF is meant to efficiently represent the structure of the audio signal power in the TF domain with a reduced number of parameters. The NMF factors were first treated as parameters [3, 4, 5, 2], and then as latent variables within a Bayesian framework [6, 2, 7, 8, 9].

These models have been successfully applied to audio source separation. In MASS configurations, the source signal models are combined with a mixing model, accounting for the source-to-sensor channels, e.g. [10, 11, 12, 13, 14]. In general, an EM algorithm is derived to estimate the source and channel parameters, which are then used to construct demixing Wiener filters. Besides, a general framework for inserting prior information about the sources in TF-domain MASS has been proposed in [15].

In current Bayesian NMF PSD models, the source PSD is first modeled with NMF and, second, the NMF factors are assigned a prior distribution. In this paper, we propose to change the order of things: The source is still seen as a sum of components, but we first assign a prior distribution to the component PSD and then assume a factorized model, reminiscent of NMF, on the parameters of this distribution. More precisely, we model the component PSD with an inverse-Gamma (IG) distribution and assume that the scale parameter of this IG follows a rank-1 NMF. As explained in Section 2, this enables to add flexibility in the modeling of source PSD matrix, compared to conventional NMF, while preserving the ability to model structured source PSD. We apply the proposed model to MASS from underdetermined convolutive mixtures.

The proposed model is presented in Section 2. The associated variational EM (VEM) algorithm that we derived to estimate the model parameters and infer the source signals is described in Section 3. Experimental evaluation reported in Section 4 shows competitive performances in comparison to the state-of-the-art LGM-NMF MASS method of [10].

2. MODELS

2.1. The mixing model

As usually done in the MASS literature, we work under the narrow-band assumption, which allows us to write a time-invariant convolutive mixture in the short-term Fourier transform (STFT) domain as:

$$\mathbf{x}_{f\ell} = \mathbf{A}_f \mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}, \tag{1}$$

where $\mathbf{x}_{f\ell} = [x_{1,f\ell}, \ldots, x_{I,f\ell}]^\top \in \mathbb{C}^I$ is the I channel observation vector, $\mathbf{s}_{f\ell} = [s_{1,f\ell}, \ldots, s_{J,f\ell}]^\top \in \mathbb{C}^J$ is the source vector to be inferred, $\mathbf{A}_f \in \mathbb{C}^{I\times J}$ is a mixing matrix parameter to be estimated, and $\mathbf{b}_{f\ell} \in \mathbb{C}^I$ is the sensor noise. We assume $p(\mathbf{b}_{f\ell}) = \mathcal{N}_c(\mathbf{b}_{f\ell}; \mathbf{0}, \mathrm{v}_f \mathbf{I}_I)$,[1] with $\mathrm{v}_f \in \mathbb{R}_+$

This research has received funding from the EU-FP7 STREP project EARS (#609465) and ERC Advanced Grant VHIA (#340113).

[1] $\mathcal{N}_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\pi\boldsymbol{\Sigma}|^{-1} \exp\left(-[\mathbf{x}-\boldsymbol{\mu}]^{\mathrm{H}} \boldsymbol{\Sigma}^{-1} [\mathbf{x}-\boldsymbol{\mu}]\right)$ is the proper complex Gaussian distribution with $\mathbf{x} \in \mathbb{C}^I$, $\boldsymbol{\mu} \in \mathbb{C}^I$ and $\boldsymbol{\Sigma} \in \mathbb{C}^{I\times I}$.

being a variance parameter to be estimated ($\mathbf{I}_I$ is the identity matrix of dimension I). The above assumption implies $p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}) = \mathcal{N}_c(\mathbf{x}_{f\ell}; \mathbf{A}_f\mathbf{s}_{f\ell}, \mathrm{v}_f\mathbf{I}_I)$. Note that the above mixture can be underdetermined, i.e. we can have $I < J$.

2.2. The source model

We embrace the LGM framework [1, 10] where $s_{j,f\ell} \in \mathbb{C}$ is assumed to follow a zero-mean proper complex Gaussian distribution. Moreover, $s_{j,f\ell}$ is assumed to be the sum of elementary components $c_{k,f\ell} \in \mathbb{C}$, also zero-mean proper complex Gaussian:

$$s_{j,f\ell} = \sum_{k \in \mathcal{K}_j} c_{k,f\ell} \Leftrightarrow \mathbf{s}_{f\ell} = \mathbf{G}\mathbf{c}_{f\ell}, \tag{2}$$

where $\mathcal{K}_j$ is a subset of a nontrivial partition $\mathcal{K} = \{\mathcal{K}_j\}_{j=1}^J$ (known in advance) of the K components into the J sources, $\mathbf{c}_{f\ell} = [c_{1,f\ell}, \ldots, c_{K,f\ell}]^\top \in \mathbb{C}^K$ is the vector of component coefficients, and $\mathbf{G} \in \mathbb{N}^{J\times K}$ is a binary matrix with entries $G_{jk} = 1$ if $k \in \mathcal{K}_j$ and $G_{jk} = 0$ otherwise. Finally, as in [10], we assume that all $\{c_{k,f\ell}\}_{f,\ell,k=1}^{F,L,K}$ are independent, with:

$$p(c_{k,f\ell}|u_{k,f\ell}) = \mathcal{N}_c(c_{k,f\ell}; 0, u_{k,f\ell}). \tag{3}$$

In particular, the source PSD at each TF bin is the sum of individual component PSDs at that bin.

2.3. The component PSD model

Traditionally, the component PSD (or component variance) $u_{k,f\ell}$ is typically assumed to factorise over f and ℓ, i.e. an NMF model is applied on the source PSD directly. In a Bayesian framework, the NMF factors are assigned a prior distribution. The main contribution of this paper is to reverse the traditional order: We first assume a prior distribution for the component variance and then impose a nonnegative factorized structure on its parameters. More precisely, we assume that each entry $u_{k,f\ell}$ of the component PSD matrix follows an inverse Gamma (IG) distribution[2]:

$$p(u_{k,f\ell}) = \mathcal{IG}\left(u_{k,f\ell}; \gamma_k, \delta_{k,f\ell}\right) \text{ with } \delta_{k,f\ell} = w_{fk}h_{k\ell}, \tag{4}$$

where $\gamma_k, w_{fk}, h_{k\ell} \in \mathbb{R}_+$. The choice for the IG distribution emerges naturally, as it is the conjugate prior of the variance of a Gaussian. The factorization of the scale parameter $\delta_{k,f\ell}$ into a rank-1 model is a key point of our model. Indeed, modeling the *parameters* of the component PSD prior (for instance $\delta_{k,f\ell}$) with a rank-1 model instead of the component PSD $u_{k,f\ell}$ itself allows the latter not to be constrained to have a low-rank structure. Therefore, with the proposed model, both the component PSD and the source PSD can be full-rank, as opposed to conventional NMF. In the meantime, the proposed model keeps a limited number of parameters and an ability to represent structured signals, in the spirit of conventional NMF. Finally, we postulate that the proposed model has the potential to better represent natural audio signals, such as speech. As for the IG shape parameter γ_k, it intuitively acts as a measure of the relevance of the k-th component: high (resp. low) values of γ_k decrease (resp. increase) the contribution of the k-th component.

[2]The Inverse Gamma distribution is defined as $\mathcal{IG}(u;\gamma,\delta) = \frac{(\delta)^\gamma}{\Gamma(\gamma)}u^{-(\gamma+1)}\exp\left(-\frac{\delta}{u}\right)$, with support $u \in \mathbb{R}_+$, shape parameter $\gamma \in \mathbb{R}_+$, scale parameter $\delta \in \mathbb{R}_+$, and $\Gamma(\cdot)$ being the Gamma function.

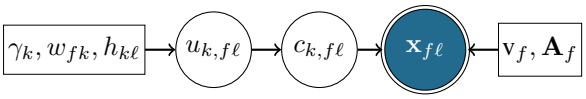

Fig. 1. Graphical reprsentation of the probabilisitc model. Latent variables are represented with circles, observations with double circles, deterministic parameters with rectangles.

3. VARIATIONAL INFERENCE

We propose an EM algorithm to perform inference of the hidden variables $\mathcal{H} = \{\mathbf{c}_{f\ell}, u_{k,f\ell}\}_{f,\ell,k=1}^{F,L,K}$ and estimation of the parameters $\theta = \{\mathbf{A}_f, \mathrm{v}_f, \gamma_k, w_{fk}, h_{k\ell}\}_{f,\ell,k=1}^{F,L,K}$. As the E-step does not admit a closed form solution, we use variational inference: Let $q(\mathcal{H}_0) = p\big(\mathcal{H}_0|\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L};\theta\big)$ denote the posterior distribution of a variable $\mathcal{H}_0 \in \mathcal{H}$. First $q(\mathcal{H})$ is imposed to factorise as $q(\mathcal{H}) \approx \prod_{f,\ell=1}^{F,L} q(\mathbf{c}_{f\ell}) \prod_{f,\ell,k=1}^{F,L,K} q(u_{k,f\ell})$. Then $q(\mathcal{H}_0) = p(\mathcal{H}_0|\,\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L};\theta)$ is infered with:

$$q(\mathcal{H}_0) \propto \exp\Big(\mathbb{E}_{q(\mathcal{H}/\mathcal{H}_0)}\big[\log p(\mathcal{H}, \{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L};\theta)\big]\Big), \tag{5}$$

where $\mathbb{E}_{q(\mathbf{z})}[f(\mathbf{z})]$ is the expectation of functional $f(\mathbf{z})$ w.r.t. the distribution $q(\mathbf{z})$ over the support of the variable $\mathbf{z}$, and where $q(\mathcal{H}/\mathcal{H}_0)$ is the joint posterior distribution of all hidden variables except $\mathcal{H}_0$. The resulting E-step is the alternating inference of $q(\mathbf{c}_{f\ell})$ (E-C), and $q(u_{k,f\ell})$ (E-U), $\forall f, \ell, k$.

3.1. E-step

Let the superscript (r) denote the (V)EM iteration index, i.e. $\theta^{(r)}$ are the parameters computed at the r^{th} iteration.

E-U-step: First we consider the inference of $q(u_{k,f\ell})$. Using (5), one can easily identify $q(u_{k,f\ell})$ to be also an IG:

$$\begin{aligned} q(u_{k,f\ell}) &\propto p(u_{k,f\ell}) \exp\left(\mathbb{E}_{q(\mathbf{c}_{f\ell})}[\log p(c_{k,f\ell}|u_{k,f\ell})]\right) \\ &= \mathcal{IG}\left(u_{k,f\ell}; g_k^{(r)}, d_{k,f\ell}^{(r)}\right), \end{aligned} \tag{6}$$

with posterior parameters $g_k^{(r)}, d_{k,f\ell}^{(r)} \in \mathbb{R}_+$ calculated as:

$$g_k^{(r)} = \gamma_k^{(r-1)} + 1,\ d_{k,f\ell}^{(r)} = \delta_{k,f\ell}^{(r-1)} + \Sigma_{kk,f\ell}^{\mathbf{c}(r-1)} + \big|\hat{c}_{k,f\ell}^{(r-1)}\big|^2, \tag{7}$$

where $\Sigma_{kk,f\ell}^{\mathbf{c}(r-1)} \in \mathbb{R}_+$ is the k^{th} diagonal entry of $\mathbf{\Sigma}_{f\ell}^{\mathbf{c}(r-1)}$, and $\hat{c}_{k,f\ell}^{(r-1)} \in \mathbb{C}$ is the k^{th} entry of $\hat{\mathbf{c}}_{f\ell}^{(r-1)}$, both being calculated below.

E-C-step: Using (5), $q(\mathbf{c}_{f\ell})$ can be identified to be complex-Gaussian:

$$q(\mathbf{c}_{f\ell}) \propto p(\mathbf{x}_{f\ell}|\mathbf{G}\mathbf{c}_{f\ell}) \prod_{k=1}^{K} \exp\left(\mathbb{E}_{q(u_{k,f\ell})}[\log p(c_{k,f\ell}|u_{k,f\ell})]\right)$$
$$= \mathcal{N}_c\left(\mathbf{c}_{f\ell}; \hat{\mathbf{c}}_{f\ell}^{(r)}, \mathbf{\Sigma}_{f\ell}^{\mathbf{c}(r)}\right). \quad (8)$$

The posterior covariance matrix $\mathbf{\Sigma}_{f\ell}^{\mathbf{c}(r)} \in \mathbb{C}^{K\times K}$ and component vector estimate $\hat{\mathbf{c}}_{f\ell}^{(r)} \in \mathbb{C}^K$ are given by:

$$\mathbf{\Sigma}_{f\ell}^{\mathbf{c}(r)} = \left[\mathrm{diag}_K\left(\frac{g_k^{(r-1)}}{d_{k,f\ell}^{(r-1)}}\right) + \frac{\left(\mathbf{A}_f^{(r-1)}\mathbf{G}\right)^{\mathrm{H}}\mathbf{A}_f^{(r-1)}\mathbf{G}}{\mathrm{v}_f^{(r-1)}}\right]^{-1},$$

$$\hat{\mathbf{c}}_{f\ell}^{(r)} = \mathbf{\Sigma}_{f\ell}^{\mathbf{c}(r)}\left(\mathbf{A}_f^{(r-1)}\mathbf{G}\right)^{\mathrm{H}}\left(\mathbf{x}_{f\ell}/\mathrm{v}_f^{(r-1)}\right), \quad (9)$$

where $\mathrm{diag}_K(x_k)$ is the $K \times K$ diagonal matrix with entries $x_1, \ldots, x_K$. Eq. (9) corresponds to the Wiener filtering of the component, thus a similar result as in [10] except for the construction of the component posterior covariance matrix.

Estimating the source coefficients: Now, using (2), it is easy to calculate the source posterior distribution, which as one expects, is a complex Gaussian with mean $\hat{\mathbf{s}}_{f\ell}^{(r)} \in \mathbb{C}^J$, and 2nd-order moment $\mathbf{R}_{f\ell}^{\mathbf{s}(r)} \in \mathbb{C}^{J\times J}$ calculated as:

$$\hat{\mathbf{s}}_{f\ell}^{(r)} = \mathbf{G}\hat{\mathbf{c}}_{f\ell}^{(r)}, \quad \mathbf{R}_{f\ell}^{\mathbf{s}(r)} = \mathbf{G}\mathbf{\Sigma}_{f\ell}^{\mathbf{c}(r)}\mathbf{G}^\top + \hat{\mathbf{s}}_{f\ell}^{(r)}\left(\hat{\mathbf{s}}_{f\ell}^{(r)}\right)^{\mathrm{H}}. \quad (10)$$

3.2. M-step

As for the M step, the parameters maximizing the expected complete-data log-likelihood are computed.

M-$\mathbf{A}_f$ step: The optimal value for the filters is:

$$\mathbf{A}_f^{(r)} = \left(\frac{1}{L}\sum_{\ell=1}^{L}\mathbf{x}_{f\ell}\left(\hat{\mathbf{s}}_{f\ell}^{(r)}\right)^{\mathrm{H}}\right)\left(\frac{1}{L}\sum_{\ell=1}^{L}\mathbf{R}_{f\ell}^{\mathbf{s}(r)}\right)^{-1}, \quad (11)$$

which is a standard form of least square estimator [10].

M-$\mathbf{v}_f$ step: The optimal noise variance is:

$$\mathrm{v}_f^{(r)} = \frac{1}{LI}\sum_{\ell=1}^{L}\left(\mathbf{x}_{f\ell}^{\mathrm{H}}\mathbf{x}_{f\ell} - 2\mathfrak{Re}\left\{\mathbf{x}_{f\ell}^{\mathrm{H}}\mathbf{A}_f^{(r)}\hat{\mathbf{s}}_{f\ell}^{(r)}\right\} + \mathrm{tr}\left\{\mathbf{R}_{f\ell}^{\mathbf{s}(r)}\left(\mathbf{A}_f^{(r)}\right)^{\mathrm{H}}\mathbf{A}_f^{(r)}\right\}\right), \quad (12)$$

where $\mathrm{tr}\{.\}$ is the trace operator.

M-IG step: The IG parameters $(\gamma_k, w_{fk}, h_{k\ell})$ are coupled in the objective function and thus an alternating optimization strategy is required, i.e. fixing two parameters to estimate the third. The updates for $w_{fk}, h_{k\ell}$ are:

$$w_{fk}^{(r)} = \frac{L\gamma_k^{(r-1)}}{g_k^{(r)}\sum_{\ell=1}^{L}\frac{h_{k\ell}^{(r-1)}}{d_{k,f\ell}^{(r)}}}, \quad h_{k\ell}^{(r)} = \frac{F\gamma_k^{(r-1)}}{g_k^{(r)}\sum_{f=1}^{F}\frac{w_{fk}^{(r-1)}}{d_{k,f\ell}^{(r)}}}. \quad (13)$$

Algorithm 1 Separation of J static sound sources

input $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$, binary matrix $\mathbf{G}$, initial parameters $\theta^{(0)}$.
initialise IG parameters: $\left\{g_k^{(0)}, d_{k,f\ell}^{(0)}\right\}_{f,\ell,k=1}^{F,L,K}$, **set** $r = 1$.
repeat
  E-C step: Compute $\mathbf{\Sigma}_{f\ell}^{\mathbf{c}(r)}$ and $\hat{\mathbf{c}}_{f\ell}^{(r)}$ with (9). Compute $\hat{\mathbf{s}}_{f\ell}^{(r)}$ and $\mathbf{R}_{f\ell}^{\mathbf{s}(r)}$ with (10).
  E-U step: Calculate $g_k^{(r)}$ and $d_{k,f\ell}^{(r)}$, with (7).
  M-$\mathbf{A}_f$ step: Update $\mathbf{A}_f^{(r)}$ with (11).
  M-$\mathbf{v}_f$ step: Update $\mathrm{v}_f^{(r)}$ with (12).
  M-IG step: Update $w_{fk}^{(r)}, h_{k\ell}^{(r)}$ with (13). Calculate $\delta_{k,f\ell}^{(r)} = w_{fk}^{(r)}h_{k\ell}^{(r)}$. Update $\gamma_k^{(r)}$ with (15).
  set $r = r + 1$.
until convergence
return the estimated source images.

Then we set $\delta_{k,f\ell}^{(r)} = w_{fk}^{(r)}h_{k\ell}^{(r)}$, and the update for $\gamma_k^{(r)}$ is the solution w.r.t $\gamma_k^{(r)}$ to:

$$\psi\left(g_k^{(r)}\right) - \psi\left(\gamma_k^{(r)}\right) = \frac{1}{FL}\sum_{f=1}^{F}\sum_{\ell=1}^{L}\log\left(\frac{d_{k,f\ell}^{(r)}}{\delta_{k,f\ell}^{(r)}}\right), \quad (14)$$

where $\psi(.)$ is the digamma function. Since (14) has no closed-form solution, we propose to approximate $g_k^{(r)}$ with $\gamma_k^{(r)} + 1$, relying on (7), and use the recurrence relation of the digamma function $\psi(x+1) = \psi(x) + \frac{1}{x}$. This leads to the following update rule:

$$\gamma_k^{(r)} = \frac{1}{\frac{1}{FL}\sum_{f=1}^{F}\sum_{\ell=1}^{L}\log\left(\frac{d_{k,f\ell}^{(r)}}{\delta_{k,f\ell}^{(r)}}\right)}. \quad (15)$$

3.3. Estimation of source images

Considering the inherent scale indeterminacy of the source separation problem, we rather measure the separation performance using the (time domain) source *images*, i.e. the estimates of the source signals as recorded at the microphones [11, 16]. These are calculated by applying inverse STFT with overlap-add on $\{\mathbf{a}_{j,f}\hat{s}_{j,f\ell}\}_{f,\ell=1}^{F,L}$ ($\mathbf{a}_{j,f}$ is the j-th column of $\mathbf{A}_f$). The complete VEM procedure can be found in Algorithm 1.

4. EXPERIMENTS

To asses the performance of the proposed algorithm, we simulated the challenging task of separating $J = 3$ sources from a convolutive stereo mixture ($I = 2$). Source signals were 2s-speech signals randomly chosen from the TIMIT database [17]. As mixing filters, we used binaural room

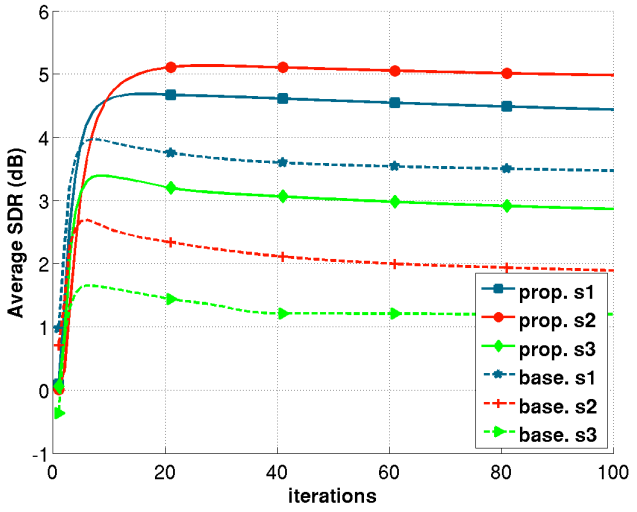

Fig. 2. Average SDR score as a function of (V)EM iterations (Mix-1,$R = 0$dB).

impulse responses (BRIR) from [18] truncated to 512 taps with reverberation time of $\text{RT}_{60} \approx 0.68$s. Two sets of BRIRs were used, corresponding respectively to azimuths $-85°, -20°, 60°$ (Mix-1), and azimuths $-45°, 75°, 10°$ (Mix-2). Standard sound separation measures, namely signal-to-distortion ratio (SDR) and signal-to-interference-ratio (SIR) [19] were computed. All reported results are average measures over 8 sets of utterances (for each mix).

To ensure a fair comparison with the baseline method [10], we provided both algorithms with the same initial information. The mixing filters were blindly initialized to $\mathbf{A}_f^{(0)} = \mathbf{1}$ (a matrix filled with ones) and we set $\text{v}_f^{(0)} = \frac{10^3}{FLI} \sum_{f\ell} (\mathbf{x}_{f\ell})^{\text{H}} \mathbf{x}_{f\ell}$, $\forall f$. As for the NMF parameters, each source was corrupted with the sum of the two other sources at two different SNRs ($R = 10$dB or 0dB). An initial NMF decomposition $\{w_{fk}^{init}, h_{k\ell}^{init}\}_{f,\ell,k}$ was then computed for each corrupted source PSD using the KL-NMF algorithm [2], with $|\mathcal{K}_j| = 20$ components per source. $\{w_{fk}^{init}, h_{k\ell}^{init}\}_{f,\ell,k}$ were also used to initialize the NMF parameters of the baseline method. We set $\delta_{k,f\ell}^{(0)} = w_{fk}^{init} h_{k\ell}^{init}$, $\gamma_k^{(1)} = 1$ and $d_{k,f\ell}^{(0)} = \delta_{k,f\ell}^{(0)}$ (thus $g_k^{(0)} = 2$ and $\mathbb{E}_{\mathcal{IG}(u_{k,f\ell}; g_k^{(0)}, d_{k,f\ell}^{(0)})}[u_{k,f\ell}] = w_{fk}^{init} h_{k\ell}^{init}$). We run 100 iterations.

Fig. 2 shows the average SDR obtained at each iteration for Mix-1 with $R = 0$dB. We observe a quite regular evolution of the SDR scores, which are quite stabilized after 100 iterations (after some possible decrease, since the (V)EM does not guarantee monotonic evolution of the separation scores as opposed to the likelihood). For this mix, the proposed method shows a notable improvement over the baseline: up to 3.1dB for s_2 (remind that these scores are averaged over 8 experiments with same filters but different sources). Final performance (at iteration 100) for other configurations, and for SIR scores, are reported in Table 1. There we see that for Mix-1 and $R = 0$dB (configuration of Fig. 2), the SIR improvement is in line with the SDR improvement: the proposed model outperforms the baseline by 5.8dB for s_2, while the results for the two other sources are less impressive. Such quite substantial improvement of SDR and SIR may be due to the added flexibility of the proposed PSD model compared with NMF (see Section 2.3). As for Mix-1 with $R = 10$dB, all scores are higher because the NMF initialization is closer to true source PSDs. Here, the proposed method also notably and systematically outperforms the baseline method. The results are more mitigated with the Mix-2 configuration. Here both the SDR and SIR scores of the two methods are more intricate. Note that the scores of the proposed method are remarkably similar across the two mixes, as opposed to the scores of the baseline method. This seems to indicate that the proposed method is robust to the mixing configuration, but further investigation must be conducted to conclude on this. Globally, the overall results encourage us to further investigate the potential of this full-rank PSD modeling, for source separation and beyond.

Table 1. Average SDR and SIR scores.

			SDR (dB)			SIR (dB)		
	R (dB)	Algo.	s1	s2	s3	s1	s2	s3
Mix-1	10	Prop.	9.1	7.8	7.5	12.2	12.5	11.3
		Base.	7.6	5.6	3.4	12.2	10.1	4.2
	0	Prop.	4.4	5.0	2.9	5.2	8.1	3.6
		Base.	3.5	1.9	1.2	5.3	2.3	2.4
Mix-2	10	Prop.	9.0	7.7	7.4	12.2	12.5	11.4
		Base.	9.1	7.7	7.1	12.7	12.2	10.8
	0	Prop.	4.6	5.4	2.8	5.8	8.7	4.1
		Base.	5.0	5.0	3.1	7.1	8.6	4.7

5. CONCLUSION

MASS experiments have shown the potential of the proposed model for TF-domain statistical signal modeling. Future research will concern an in-depth analysis of the proposed PSD model *per se*, i.e. to model audio signals independently of the MASS context. This should include a comparative study with conventional parametric and Bayesian NMF. Also, we will investigate the characterization of component relevance from the estimated shape parameter, and its use within a model selection task when the number of source components is unknown, see, e.g. [20]. As for the MASS task, we will explore more realistic initialization techniques, e.g. using the output of existing source separation techniques, leading to a more realistic sound separation algorithm and, again, more systematic comparison with other source PSD models plugged into the LGM-based MASS framework.

6. REFERENCES

[1] A. Liutkus, B. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155–3167, 2011.

[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] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[4] ——, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001.

[5] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2003.

[6] T. Virtanen, S. Godsill *et al.*, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 1825–1828.

[7] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 538–549, 2010.

[8] M. Hoffman, D. Blei, and P. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *International Conference on Machine Learning*, 2010, pp. 439–446.

[9] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using bayesian NMF with recursive temporal updates of prior distributions," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Kyoto, Japan, 2012.

[10] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[11] N. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[12] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *International Conference on Information Sciences, Signal Processing, and their Applications*, 2010.

[13] T. Higuchi, N. Takamune, N. Tomohiko, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based on DOA-HMM," in *IEEE International Conference on Audio, Speech and Signal Processing*, 2014.

[14] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of moving sound sources," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2015.

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

[16] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet, "Linear mixing models for active listening of music productions in realistic studio conditions," in *Convention of the Audio Engineering Society (AES)*, Budapest, Hungary, 2012.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993, linguistic Data Consortium, Philadelphia.

[18] C. Hummersone, R. Mason, and T. Brookes, "A comparison of computational precedence models for source separation in reverberant environments," *Journal of the Audio Engineering Society*, vol. 61, no. 7/8, pp. 508–520, 2013.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[20] V. Y. Tan and C. Fevotte, "Automatic relevance determination in nonnegative matrix factorization with the β-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592–1605, 2013.