Student's t Source and Mixing Models for Multichannel Audio Source Separation

Simon Leglaive, Roland Badeau, Gael Richard

Student's *t* Source and Mixing Models for Multichannel Audio Source Separation

Simon Leglaive, Roland Badeau, *Senior Member, IEEE,* Gaël Richard *Fellow, IEEE*

***Abstract*—This paper presents a Bayesian framework for under-determined audio source separation in multichannel reverberant mixtures. We model the source signals as Student's *t* latent random variables in a time-frequency domain. The specific structure of musical signals in this domain is exploited by means of a non-negative matrix factorization model. Conversely, we design the mixing model in the time domain. In addition to leading to an exact representation of the convolutive mixing process, this approach allows us to develop simple probabilistic priors for the mixing filters. Indeed, as those filters correspond to room responses they exhibit a simple characteristic structure in the time domain that can be used to guide their estimation. We also rely on the Student's *t* distribution for modeling the impulse response of the mixing filters. From this model, we develop a variational inference algorithm in order to perform source separation. The experimental evaluation demonstrates the potential of this approach for separating multichannel reverberant mixtures.**

***Index Terms*—Audio source separation, multichannel reverberant mixtures, Student's *t* distribution, statistical room acoustics, non-negative matrix factorization, variational inference.**

I. Introduction

AUDIO source separation consists in recovering a set of audio source signals from the observation of a mixture signal. In music, we usually encounter stereophonic mixtures (recorded with two microphones) involving more than two sources, leading to an *under-determined* source separation problem. Moreover, musical acoustic recordings are usually done in a reverberant environment, the mixing process is said to be *convolutive* in that case. This work aims to address the under-determined source separation problem for multichannel musical reverberant mixtures.

In this section, before introducing our contributions that will be further developed in the rest of the paper, we first review some of the related source and mixing models that have been proposed in the literature for solving the audio source separation problem.

A. Source Modeling

Audio source separation methods commonly work with a time-frequency (TF) representation of the source signals. Audio signals are indeed much sparser in the TF domain than in the original time-domain, which facilitates the development of a source model [1].

S. Leglaive, R. Badeau and G. Richard are with LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France (e-mail: firstname.lastname@telecom-paristech.fr).

The research leading to this paper was partly supported by the French National Research Agency (ANR) as a part of the EDISON 3D project (ANR-13-CORD-0008-02).

Sparse component analysis methods [2] for instance make use of sparsity inducing penalties [3], [4] or super-Gaussian distributions [5], [6] for modeling the source signals, assuming independent and identically distributed TF source coefficients.

In order to take more complex TF structures into account, another approach referred to as the *variance modeling framework* in [7] was introduced. It consists in modeling the TF source coefficients by means of a non-stationary probability distribution. In the Gaussian case and under a local stationarity assumption of the source signals [8], this approach amounts to modeling the short-term power spectral density (PSD) of the source signals, represented by the variance of the TF source coefficients.

For instance, a piece of music is generally composed of musical events that repeat over time such as tonal musical notes or percussive sounds. This redundant structure can be represented within the variance modeling framework, by means of a compositional model [9] such as non-negative matrix factorization (NMF). A Gaussian generative model based on a non-negative decomposition of the short-term PSD of audio signals was first introduced in the single-channel source separation method [10]. This model was then further developed in [11] where the equivalence between maximum likelihood estimation of the parameters and NMF with the Itakura-Saito divergence was proven. In the recent literature many different probability distributions have been studied for audio signal modeling based on NMF [12]–[16].

Even though those NMF-based probabilistic models were introduced for single-channel audio source separation, they were rapidly used for multichannel audio source separation as well, such as in [17]–[21].

We can also mention that deep neural networks were recently introduced within the variance modeling framework for multichannel audio source separation [22], [23]. This work illustrates the promising approach of building methods using supervised source modeling and unsupervised convolutive mixture modeling.

B. Convolutive Mixture Modeling

Under-determined source separation becomes even more challenging when the mixture is reverberant. When a punctual source is emitting sound in an enclosed space, the signal recorded by the microphone is equal to the convolution of the source signal with a *room impulse response* (RIR). This RIR characterizes the acoustic path between the two points in the room. In the context of audio source separation, the RIRs associated with every pair source-microphone are referred to as the *mixing filters*.

More formally, let $s_j(t) \in \mathbb{R}$, $t = 0, ..., L_s - 1$, $j = 1, ..., J$, be the j-th source signal and $a_{ij}(t) \in \mathbb{R}$, $t = 0, ..., L_a - 1$, $i = 1, ..., I$, the impulse response of the mixing filter between source j and microphone i. We define the j-th *source image* as seen by the i-th microphone for $t = 0, ..., T - 1$ with $T = L_s + L_a - 1$ as follows:

$$y_{ij}(t) = [a_{ij} \star s_j](t) = \sum\nolimits_{\tau=0}^{L_a-1} a_{ij}(\tau) s_j(t - \tau), \quad (1)$$

where the source signal is zero outside of the support $\{0, ..., L_s - 1\}$. The signal recorded by microphone i is defined as the sum of the source images:

$$x_i(t) = \sum\nolimits_{j=1}^{J} y_{ij}(t). \quad (2)$$

A large number of source separation methods also work with a TF representation of the mixture signal. It is very common to rely on the short-time Fourier transform (STFT) because it allows us to approximate the time-domain convolution as a simple multiplication in each frequency band of the STFT:

$$y_{ij,fn} = a_{ij,f} s_{j,fn}, \quad (3)$$

where $a_{ij,f} \in \mathbb{C}$ is the frequency response of the mixing filter and $s_{j,fn}, y_{ij,fn} \in \mathbb{C}$ are the STFT coefficients of $s_j(t)$ and $y_{ij}(t)$ respectively, at the TF point (f, n) where f denotes the frequency index and n the time-frame index. Such an approach has for example been followed in [17] along with a Gaussian NMF-based source model. However, the approximation in (3) is only valid when the mixing filters are shorter than the STFT analysis window [24], [25]. The length of the mixing filters is usually considered to be equal to the reverberation time, denoted by T_{60} and defined as the time it takes for the sound energy to decrease by 60 dB after extinction of the source. The reverberation time in domestic or office rooms usually varies between 0.1 and 0.8 s, while for concert halls it can be as high as 2 s [26]. On the contrary the length of the STFT window is typically short (between 30 and 120 ms) in order to assume the local stationarity of the source signals over the temporal support of the STFT window.

Therefore, the approximation in (3) is limited to very weakly reverberant mixtures. This is precisely the reason why other mixture models have been investigated in the literature. A Gaussian multichannel source image model based on a *spatial covariance matrix* was developed in [27] to represent non-punctual sources. It was shown to overcome to some extent the limitations of the approximation (3). This spatial covariance matrix model was used along with an NMF-based source model in the multichannel source separation methods [18], [19], [28].

In [29] it was demonstrated that the time-domain convolution can be exactly expressed as a two-dimensional filtering in a TF domain based on any perfect reconstruction filter bank. This representation is called *cross-band filtering* (CBF) in [30] when the filter bank corresponds to the STFT. In [31], the time-domain convolution is approximated as a sub-band filtering in the STFT domain. This approach is equivalent to neglecting the relationships between the frequency bands in the CBF representation. It is referred to as the convolutive transfer function (CTF) model in the recent source separation methods [32] and [33].

Finally, the methods in [4] and [34] directly worked with the time-domain mixture signals in order to exactly represent the convolutive mixing process, but in a limited semi-blind scenario where the mixing filters were known. We also followed this time-domain approach in previous works. In [35], the sources were represented in the Modified Discrete Cosine Transform (MDCT) domain. Each TF coefficient was modeled as a centered Gaussian random variable whose variance was structured by means of an NMF model. The time-domain mixing filters were treated as deterministic parameters that had to be estimated, however they were initialized in an oracle fashion. In [36] we extended this approach by using a Student's *t* distribution for modeling the source MDCT coefficients[1]. We considered a semi-blind setting assuming that the mixing filters were known and fixed. Finally, in [37] we generalized our approach in [35] to a source representation based on the STFT. Again the experiments were performed knowing the true mixing filters.

C. Contributions

Apart from leading to an exact representation of the convolutive mixing process, working with time-domain mixing filters allows us to design simple probabilistic priors. Indeed, as the mixing filters correspond to RIRs, they exhibit a simple specific structure in the time domain that can be used to guide their estimation. In an under-determined context, being able to exploit some additional information is a very important aspect. For example, it has been shown that taking this specific temporal structure into account is useful for the estimation of multiple RIRs in a non-blind (knowing the source signals) and potentially under-determined (more RIR samples than observed mixture samples) setting [38] [39, Ch. 5].

As previously mentioned, the source separation methods in the literature that represented the convolutive mixing process in the time domain depended on the true mixing filters. This scenario was restrictive because in a real-world application it is difficult to have this knowledge. The main contribution of the present paper is to propose a new Bayesian framework where the mixing filters are treated as latent random variables in the time domain. We introduce a probabilistic modeling of the mixing filters using the Student's *t* distribution. It generalizes a widely used Gaussian model in statistical room acoustics. We also formulate the source model so that different TF resolutions can be used to characterize different sources in the mixture. Finally the experiments are carried out in a more realistic scenario as we do not assume that the mixing filters are known.

The rest of this paper is organized as follows: Section II introduces the source and mixing models. The variational inference algorithm is derived in Section III. We detail the experimental results in Section IV and finally conclude in Section V.

[1]Note that the Student's *t* source model in the present paper is slightly different from the one in this previous work.

II. Models

A. Source Time-Frequency Representation

The source signals are represented by a set of TF synthesis coefficients $\{s_{j,fn} \in \mathbb{R}\}_{(f,n)\in\mathcal{B}_j}$ with $\mathcal{B}_j = \{0,...,F_j-1\} \times \{0,...,N_j-1\}$ as follows:

$$s_j(t) = \sum\nolimits_{(f,n)\in\mathcal{B}_j} s_{j,fn}\phi_{j,fn}(t). \tag{4}$$

We work with the MDCT, such that for all $(f,n) \in \mathcal{B}_j$:

$$\phi_{j,fn}(t) = \sqrt{\frac{4}{M_j}} w_j(t-nH_j) \cos\left(\frac{2\pi}{M_j}\left(t-nH_j+\frac{1}{2}+\frac{M_j}{4}\right)\left(f+\frac{1}{2}\right)\right), \tag{5}$$

where $w_j(t)$ is a sine window of length M_j, $H_j = M_j/2$ is the hop size and $F_j = M_j/2$ is the number of frequency bins. Note that in this representation the window length is source-dependent. This specificity allows us to adapt the TF resolution according to the sources in the mixture. For example we can choose a short window for drums, resulting in a good time resolution, and a longer window for tonal instruments such as the guitar, resulting in a good frequency resolution.

Any TF dictionary could be chosen in the source representation (4) (MDCT, STFT, unions of MDCT, ...). Here we choose to work with the MDCT because it is critically sampled. Compared with the STFT which is redundant, it allows us to reduce the number of source TF coefficients to be estimated and therefore to reduce the computational load. Moreover we experimentally showed in [37] that using the STFT does not improve the source separation performance, according to a perceptually motivated objective measure. We can also mention that in [40] it has been shown that the MDCT is more suitable than the STFT for assuming the independence of the TF points, which is a common assumption in probabilistic approaches for audio source separation.

According to this TF representation, a source image as defined in (1) can be rewritten with $g_{ij,fn}(t) = [a_{ij} \star \phi_{j,fn}](t)$ as follows:

$$y_{ij}(t) = \sum\nolimits_{(f,n)\in\mathcal{B}_j} s_{j,fn}\, g_{ij,fn}(t). \tag{6}$$

B. The Student's t Distribution

Under-determined reverberant audio source separation is an ill-posed problem in the sense that the solution is not unique; an infinite number of solutions for the source signals and the mixing filters can explain the same observed mixture. It is therefore necessary to introduce some information in order to guide the source separation process. In a Bayesian perspective, it consists in defining probabilistic priors over the source coefficients $\mathbf{s} = \{s_{j,fn}\}_{j,f,n}$ and the mixing filters $\mathbf{a} = \{a_{ij}(t)\}_{i,j,t}$. The priors should be informative in order to express some specific belief about these unknown quantities. Compared with deterministic models which would strictly restrict the possible values for the unobserved set of variables $\mathbf{s}$ and $\mathbf{a}$, using probabilistic priors allows us to take uncertainty about the model into account. As both priors on the TF source coefficients and the time-domain mixing filters will rely on the Student's *t* distribution, we first introduce it in this subsection before presenting the priors in the following ones.

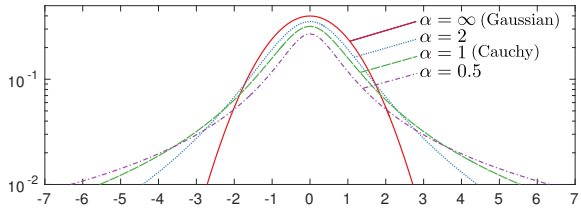

Fig. 1. Probability density function of the $\mathcal{T}_\alpha(0,1)$ distribution (in logarithmic scale) for different values of the shape parameter α.

Let $\mathcal{T}_\alpha(\mu,\lambda)$ denote the Student's *t* distribution over a real-valued random variable (r.v.) where $\mu \in \mathbb{R}$ is the location parameter, $\alpha \in]0,+\infty[$ is the shape parameter (also called degrees of freedom) and $\lambda \in]0,+\infty[$ is the scale parameter. Its probability density function (pdf) is defined in Appendix A, equation (48). The Student's *t* distribution can be expressed as a scale mixture of Gaussians (SMoG) [41] by introducing a positive real-valued inverse-gamma r.v. v:

$$x \sim \mathcal{T}_\alpha(\mu,\lambda) \Leftrightarrow \begin{cases} v & \sim \mathcal{IG}\left(\frac{\alpha}{2},\frac{\alpha}{2}\right); \\ x|v & \sim \mathcal{N}(\mu, v\lambda^2). \end{cases} \tag{7}$$

The pdfs of the Gaussian and inverse-gamma distributions are defined in Appendix A, equations (50) and (51) respectively.

The Student's *t* distribution is a super-Gaussian distribution [42] which exhibits thicker tails than the Gaussian as can be seen in Figure 1. As α tends to infinity the Student's *t* distribution approaches the Gaussian one, while smaller values of α yield heavier tails. The particular case $\alpha = 1$ corresponds to the Cauchy distribution. Informally speaking, the SMoG representation of the Student's *t* distribution in (7) can be interpreted in the following way: realizations of x are globally expected to be distributed according to a Gaussian distribution. However the overall variance λ^2 is scaled by realizations of the inverse-gamma variable v, which most of the time remain small (the mode of the distribution being $\alpha/(\alpha+2)$), but may sometimes get significantly large, accounting for outliers or high uncertainty in the model with respect to the Gaussian assumption.

C. Prior for the Sources

In the generative model [11], the TF source coefficients are assumed to independently follow a centered Gaussian distribution: $s_{j,fn} \sim \mathcal{N}(0,\lambda_{j,fn}^2)$. The variances are further structured by means of an NMF model:

$$\lambda_{j,fn}^2 = [\mathbf{W}_j\mathbf{H}_j]_{fn}, \tag{8}$$

with $\mathbf{W}_j \in \mathbb{R}_+^{F_j\times K_j}$ and $\mathbf{H}_j \in \mathbb{R}_+^{K_j\times N_j}$. K_j is the rank of the factorization and is chosen such that $K_j(F_j+N_j) \ll F_jN_j$. $\mathbf{W}_j$ is a matrix containing spectral templates and $\mathbf{H}_j$ contains the activations of these templates over the time frames.

This model was originally proposed for a source representation based on the complex-valued STFT. The source

coefficients $s_{j,fn}$ were consequently represented as circularly symmetric complex Gaussian random variables [43]. However it can be used similarly for a source representation based on the real-valued MDCT as already proposed in [44] for informed source separation. Note that if we observe a set of realizations of this Gaussian source model denoted by $\{\tilde{s}_{j,fn}\}_{f,n}$, it can be shown that the maximum likelihood estimation of the NMF parameters is obtained by minimizing the Itakura-Saito divergence [11] between $|\tilde{s}_{j,fn}|^2$ and $[\mathbf{W}_j\mathbf{H}_j]_{fn}$.

This generative Gaussian source model has then led to several extensions based on other probability distributions. In particular, heavy-tailed distributions such that the Cauchy [13], the symmetric alpha-stable [14] and the Student's *t* [15] have been investigated. The flexibility induced by these heavy-tailed distributions allows the expected source power to strongly deviate from the NMF source model. Indeed, in the Gaussian case the source coefficients $s_{j,fn}$ are expected to lie just within a few standard deviations $\lambda_{j,fn}$ from the mean. On the contrary, a heavy-tailed source model allows for larger deviations due to the thicker tails of the distribution. This flexibility can be useful when the NMF source model is a misfit due to inaccurate parameter estimation or because of the nature of the source which is not well represented by an NMF model. The superiority of these heavy-tailed distributions for modeling audio signals along with NMF has already been reported in this literature, in particular for single channel audio separation of rank-1 sources using a Student's *t* NMF observation model [15].

We propose here to model the individual source signals with this Student's *t* NMF model. We assume that the TF source coefficients $s_{j,fn}$ independently follow a Student's *t* distribution:

$$s_{j,fn} \sim \mathcal{T}_{\alpha_v}(0, \lambda_{j,fn}). \tag{9}$$

As explained in the previous subsection, this model is equivalent to the following hierarchical one:

$$\begin{cases} s_{j,fn}|v_{j,fn} & \sim \mathcal{N}(0, v_{j,fn}\lambda_{j,fn}^2); \\ v_{j,fn} & \sim \mathcal{IG}\left(\frac{\alpha_v}{2}, \frac{\alpha_v}{2}\right). \end{cases} \tag{10}$$

We further assume that the squared scale parameters are structured by means of an NMF model as in (8). The choice of α_v regarding the source separation results will be discussed in the experimental part of this paper.

D. Prior for the Mixing Filters

1) Statistical Room Acoustics: The mixing filters introduced in (1) for representing a source image are actually RIRs that are supposed to be related to a same acoustic space (a room with its specific acoustical properties). Therefore we drop the source and microphone indices in this subsection and an RIR in the acoustical environment where the mixture is recorded will be denoted by $a(t)$. It will allow us to clarify the model presentation.

We represent in the top plot of Figure 2 the first 400 ms of an RIR $\tilde{a}(t)$ from the MIRD database [45]. This RIR was recorded in a room with a reverberation time of 610 ms. The source-to-microphone distance was approximately 2 m.

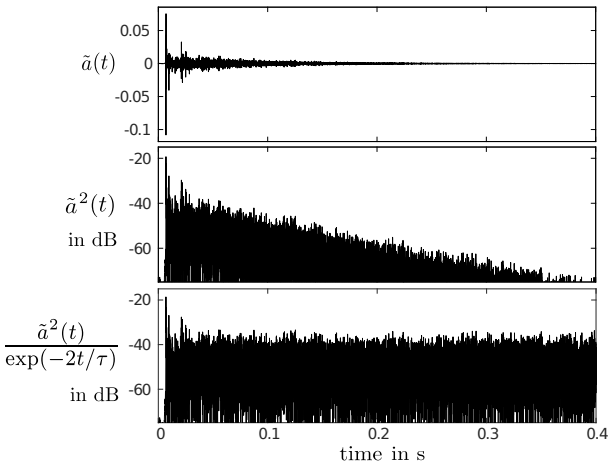

Fig. 2. Amplitude (top), power (middle) and normalized power (bottom) of the first 400 ms of an RIR from the MIRD database [45]. The RIR was recorded in a room with a reverberation time of 610 ms and with a source-to-microphone distance of 2 m.

An RIR is usually divided into two main parts: (1) the early contributions that include the direct path between the source and the microphone and the early echoes coming from the first significant reflections on the boundaries and objects in the room; (2) the late reverberation. The time instant that separates the early contributions from the late reverberation is usually called *mixing time*. It is common to assume that late reverberation corresponds to a stage of the propagation where the sound field is diffuse, i.e. the energy density is uniformly distributed across the room and over all directions [46].

The middle plot of Figure 2 shows the power in decibels (dB) of this RIR. We see that globally, the power decays exponentially over time (or equivalently linearly in dB units). In statistical room acoustics, an RIR is usually modeled as a centered random process such that $a(t)$ is a random variable. For a proper choice of probability distribution, the exponential decay property then relates to the variance $\mathbb{E}[a^2(t)]$ [47], [48]. It is important to mention that different realizations from a statistical RIR model can be interpreted as different observations with varying source and microphone positions in a same room. Mathematical expectation has thus to be understood as spatial averaging.

2) Gaussian Model with Exponential Decay: Moorer noticed in [49] that concert halls impulse responses sound very similarly to white noise with an exponentially decaying envelope. In [50], Polack introduced a widely used model in statistical room acoustics that formalized this observation. In this model an RIR is represented as a non-stationary centered Gaussian random process such that:

$$a(t) \sim \mathcal{N}(0, r^2(t)) \tag{11}$$

with

$$r(t) = \sigma_r \exp(-t/\tau), \qquad \tau = \frac{T_{60} f_s}{3\ln(10)}. \tag{12}$$

σ_r is a global scale parameter related to the total energy of the RIR and τ is the exponential decay factor defined by the reverberation time T_{60} in seconds and the sampling rate f_s in Hertz. The choice of the Gaussian distribution relates to

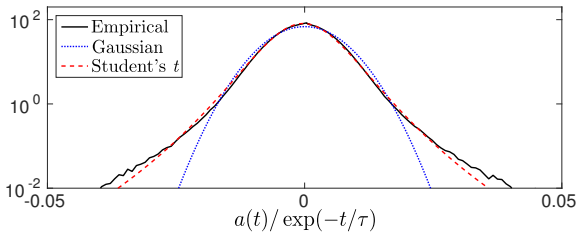

Fig. 3. Empirical pdf (black solid line) and pdf of the Gaussian (blue dotted line) and Student's t (red dashed line) distributions computed using 624 normalized RIRs from the MIRD database [45]. The RIRs were recorded in a room with a reverberation time of 610 ms.

the linearity of the Fourier transform and Schroeder's model of room frequency response [48], [51]. In this model, the room frequency response is represented as a random process whose real and imaginary parts are independent centered and stationary Gaussian processes with the same variance.

3) Preliminary Experiment: In [52] Polack concluded that RIRs can be represented by means of statistical processes only after the mixing time, when the echo density is large enough. The Gaussian model (11) is therefore theoretically valid only for the diffuse part of the RIR which corresponds to late reverberation.

We represent in the bottom plot of Figure 2 the normalized power $\tilde{a}^2(t)/\exp(-2t/\tau)$ in order to compensate for the exponential decay. The decay factor τ is fixed according to (12) using the knowledge of the reverberation time. We observe that the normalized power is indeed almost constant over time. However there are some strong deviations at the beginning of the RIR due to the direct path and early echoes.

We now propose to experimentally verify the validity of the Gaussian model (11)-(12). For that purpose we use RIRs from the MIRD database [45], that were recorded in a room with a reverberation time of 610 ms using 3 arrays of 8 microphones for 26 different source positions. It results in a total number of 624 RIRs, that we first normalize by $\exp(-t/\tau)$ in order to compensate for the exponential decay (we know the reverberation time). The empirical pdf of this set of normalized RIR coefficients is represented in Figure 3 (black solid line). According to the standard Gaussian model (11)-(12) we have $a(t)/\exp(-t/\tau) \sim \mathcal{N}(0, \sigma_r^2)$, so those normalized RIR coefficients are supposed to be independent and identically distributed (i.i.d) realizations of $\mathcal{N}(0, \sigma_r^2)$. Therefore, from this set of i.i.d data points, we estimate the mean and the variance of a Gaussian distribution in the maximum likelihood sense. We represent with a blue dotted line in Figure 3 the pdf of this Gaussian distribution, using the estimated values of the parameters (the estimated σ_r is 0.006 and the mean is almost zero). We clearly observe that the empirical distribution has thicker tails than the Gaussian. This is due to the direct path and the early echoes of the RIRs. Statistically speaking, they act as outliers with respect to the Gaussian model (11)-(12) as already noticed from the bottom plot of Figure 2.

4) Student's t Model with Exponential Decay: This is precisely the reason why we propose to use the Student's t distribution. Indeed its robustness with respect to outliers allows us to take those strong deviations into account. To experimentally demonstrate this point, we also estimate the parameters of a Student's t distribution, using the exact same data as before. The maximum likelihood estimation leads to $\sigma_r = 0.005$ and a shape parameter that is equal to 7.2, the location parameter is again very close to zero. The pdf of the Student's t distribution, using the estimated parameters, is represented in Figure 3 with a red dashed line. We clearly observe that the Student's t is a better choice for modeling RIRs.

5) Proposed Prior for the Mixing Filters: According to the previous discussion and preliminary experiment, we assume that the mixing filter coefficients $a_{ij}(t)$ independently follow a Student's t distribution:

$$a_{ij}(t) \sim \mathcal{T}_{\alpha_u}(0, r(t)), \tag{13}$$

where $r(t)$ models the exponential decay as defined in (12). This model admits Polack's Gaussian one in (11) as a special case when $\alpha_u \to \infty$. It can be rewritten in a hierarchical manner as follows:

$$\begin{cases} a_{ij}(t)|u_{ij}(t) & \sim \mathcal{N}\left(0, u_{ij}(t) r^2(t)\right); \\ u_{ij}(t) & \sim \mathcal{IG}\left(\frac{\alpha_u}{2}, \frac{\alpha_u}{2}\right). \end{cases} \tag{14}$$

The choice of α_u regarding the source separation results will be discussed in the experimental part of this paper.

E. Likelihood

To complete the model definition we need to relate the latent variables to the observations. For that purpose we simply consider an error term in the mixture equation (2) represented by a white Gaussian additive noise such that:

$$x_i(t)|\mathbf{s}, \mathbf{a}; \sigma_i^2 \sim \mathcal{N}\left(\sum\nolimits_{j=1}^{J} y_{ij}(t), \sigma_i^2\right), \tag{15}$$

where $y_{ij}(t)$ is defined according to $\mathbf{s}$ and $\mathbf{a}$ in (6). The whole model is represented as a Bayesian network [53, Chapter 8] in Figure 4.

III. Variational Inference

Let us define the following sets of latent variables: $\mathbf{s} = \{s_{j,fn}\}_{j,f,n}$; $\mathbf{v} = \{v_{j,fn}\}_{j,f,n}$; $\mathbf{a} = \{a_{ij}(t)\}_{i,j,t}$; $\mathbf{u} = \{u_{ij}(t)\}_{i,j,t}$ and $\mathbf{z} = \{\mathbf{s}, \mathbf{v}, \mathbf{a}, \mathbf{u}\}$. Let $\mathbf{x} = \{x_i(t)\}_{i,t}$ be the set of observed variables and $\boldsymbol{\theta} = \{\boldsymbol{\lambda} = \{\lambda_{j,fn}^2\}_{j,f,n}, \boldsymbol{\sigma} = \{\sigma_i^2\}_i\}$ the set of deterministic model parameters (other parameters which will be supposed to be known are not indicated).

Performing inference aims to compute the posterior distribution of the latent variables $p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^\star)$, where $\boldsymbol{\theta}^\star$ denotes an estimate of the model parameters (e.g. in the maximum likelihood sense). However exact posterior inference with the proposed model is analytically intractable. We thus resort to approximate inference using a variational approach [53].

Let $\mathcal{F}$ denote a set of pdfs over the latent variables $\mathbf{z}$. For any parameter set $\boldsymbol{\theta}$, variational inference aims to find the pdf $q^\star$ in the variational family $\mathcal{F}$ that minimizes the Kullback-Leibler (KL) divergence from the true posterior:

$$q^\star(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathcal{F}}{\arg\min}\, D_{\mathrm{KL}}(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})), \tag{16}$$

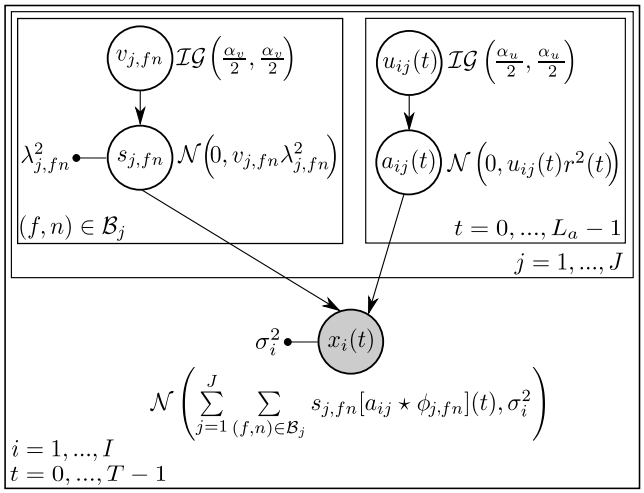

Fig. 4. Bayesian network [53, Chapter 8] corresponding to the proposed model. Latent random variables are represented with empty circles, observations with shaded circles, and the deterministic model parameters with dots. Rectangles denote the plate notation. The sub-graph contained in each rectangle is repeated according to the indicated indices. Any link that crosses a plate boundary is also replicated.

where $D_{\text{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) = \langle \ln\left(q(\mathbf{z})/p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})\right)\rangle_q$ and $< \cdot >_q$ denotes the mathematical expectation taken with respect to q. From the definition of the KL divergence we can show that:

$$D_{\text{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) = \ln p(\mathbf{x};\boldsymbol{\theta}) - \mathcal{L}(q;\boldsymbol{\theta}), \tag{17}$$

where $\ln p(\mathbf{x};\boldsymbol{\theta})$ is the marginal log-likelihood and $\mathcal{L}(q;\boldsymbol{\theta})$ is the variational free energy (VFE), also called evidence lower bound, defined as:

$$\mathcal{L}(q;\boldsymbol{\theta}) = \langle \ln\left(p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})/q(\mathbf{z})\right)\rangle_q . \tag{18}$$

$\ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$ is referred to as the complete data log-likelihood. From (17) we see that minimizing the KL divergence with respect to $q(\mathbf{z})$ is equivalent to maximizing the VFE. Moreover, as the KL divergence is always non-negative, we see that the VFE is a lower bound of the marginal log-likelihood. In this work we will use the variational expectation-maximization (VEM) algorithm [53] which consists in iterating two steps until convergence: the E-step where we compute $q^\star = \arg\max_{q\in\mathcal{F}} \mathcal{L}(q;\boldsymbol{\theta}^\star)$ and the M-step where we compute $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q^\star;\boldsymbol{\theta})$.

A. Mean-Field Approximation

In practice we have to assume some constraints on the variational family $\mathcal{F}$ for solving the E-step. Here we use the mean-field approximation assuming that the variational family corresponds to the set of pdfs that can be factorized as:

$$q(\mathbf{z}) = \prod_{j=1}^{J}\left[\prod_{(f,n)\in\mathcal{B}_j} q^s_{jfn}(s_{j,fn})q^v_{jfn}(v_{j,fn})\right] \left[\prod_{i=1}^{I}\prod_{t=0}^{L_a-1} q^a_{ijt}(a_{ij}(t))q^u_{ijt}(u_{ij}(t))\right]. \tag{19}$$

For simplicity of notation we will drop the superscripts and indices when referring to one of the factors in (19).

Under this approximation, the j-th TF source estimate is given by $\hat{s}_{j,fn} = \langle s_{j,fn}\rangle_{q(s_{j,fn})}$. The time-domain source estimate $\hat{s}_j(t)$ is then reconstructed by inverse MDCT:

$$\hat{s}_j(t) = \sum\nolimits_{(f,n)\in\mathcal{B}_j} \hat{s}_{j,fn}\phi_{j,fn}(t). \tag{20}$$

Similarly, the estimate of a mixing filter is given by $\hat{a}_{ij}(t) = \langle a_{ij}(t)\rangle_{q(a_{ij}(t))}$. Let us also define the variable $\hat{g}_{ij,fn}(t) = [\hat{a}_{ij} \star \phi_{j,fn}](t)$. Finally, the estimate of the j-th source image as seen by microphone i is given by:

$$\begin{aligned}\hat{y}_{ij}(t) =& [\hat{a}_{ij} \star \hat{s}_j](t) \\ =& \sum\nolimits_{(f,n)\in\mathcal{B}_j} \hat{s}_{j,fn}\hat{g}_{ij,fn}(t).\end{aligned} \tag{21}$$

B. Conjugate-Exponential Model

The model defined in Section II and represented as a Bayesian network in Figure 4 is said to be *conjugate-exponential* [54]. Indeed, the distribution of each latent variable, conditionally on its parents, belongs to the exponential family and the distributions of its parents are conjugate with respect to it. As a consequence of this exponential-conjugacy property, it can be shown that under the mean-field approximation, the optimal variational distribution $q^\star(z)$, $z \in \mathbf{z}$, will have the same form as the prior distribution of z conditionally on its parents [54]. For our model we consequently have:

$$q^\star(v_{j,fn}) = IG(v_{j,fn};\, \nu_v, \beta_{j,fn}); \tag{22}$$

$$q^\star(s_{j,fn}) = N(s_{j,fn};\, \hat{s}_{j,fn}, \gamma_{j,fn}); \tag{23}$$

$$q^\star(u_{ij}(t)) = IG(u_{ij}(t);\, \nu_u, d_{ij}(t)); \tag{24}$$

$$q^\star(a_{ij}(t)) = N\left(a_{ij}(t);\, \hat{a}_{ij}(t), \rho_{ij}(t)\right). \tag{25}$$

The E-step of the VEM algorithm finally aims to maximize the VFE with respect to the variational parameters, i.e. the parameters of the distributions (22) to (25). In order not to vainly overload this section with equations, we choose to detail the expression of the VFE in Appendix B.

C. E-Step

Multiple choices are possible for solving this optimization problem. The most popular approach is referred to as *coordinate ascent variational inference* (CAVI) [53], [55]. It consists in maximizing the VFE by cycling over each individual scalar variational parameter. We can indeed show that the optimal distribution $q^\star(z)$, $z \in \mathbf{z}$, should satisfy [53]:

$$\ln q^\star(z) = \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_{q(\mathbf{z}\backslash z)} + \text{constant}, \tag{26}$$

where $\mathbf{z}\backslash z$ denotes the set of all latent variables but z. This technique aims to identify the distribution of the individual factors $q^\star(z)$ along with the expression of its parameters by developing the right hand side of (26). Every term that does not depend on z can be injected in the constant. In this way it can be shown that the expectation in (26) only involves the random variables belonging to the Markov blanket $MB(z)$ of z, i.e. its parents, children and co-parents [54]. This inference

method therefore leads to a set of coupled solutions, as the variational parameters for $q^\star(z)$ will depend on the variational parameters of the distributions of the variables in $MB(z)$. Due to a lot of potential coupling, this cyclical approach can lead to a very slow inference procedure. One possible way of speeding up the algorithm is to directly update groups of variational parameters by using gradient-based approaches.

Because of the Gaussian nature of the variational distributions (23) and (25), we have to solve two quadratic and convex optimization problems to individually update the sets of variational parameters $\{\hat{s}_{j,fn}\}_{j,f,n}$ and $\{\hat{a}_{ij}(t)\}_{i,j,t}$ (see the form of the VFE defined in Appendix B). For that purpose we will use the preconditioned conjugate gradient (PCG) method [56], which was already considered for variational inference in [57]. For all other variational parameters, that are ν_v, ν_u, $\{\beta_{j,fn}, \gamma_{j,fn}\}_{j,f,n}$ and $\{d_{ij}(t), \rho_{ij}(t)\}_{i,j,t}$, we will use the standard coordinate-wise updating procedure, relying on (26), which except for ν_v and ν_u is equivalent to canceling the partial derivatives of the VFE[2].

The updates of the variational parameters for solving the E-step are detailed below.

1) E-V Step: We can develop (26) for the latent variable $v_{j,fn}$ as follows:

$$\ln q^\star(v_{j,fn}) \stackrel{c}{=} -\ln(v_{j,fn})\left(\frac{\alpha_v+1}{2}+1\right) - \frac{1}{v_{j,fn}}\left(\frac{\alpha_v}{2}+\frac{\langle s_{j,fn}^2\rangle_{q(s_{j,fn})}}{2\lambda_{j,fn}^2}\right), \quad (27)$$

where $\stackrel{c}{=}$ denotes equality up to an additive constant. From this equation we indeed recognize that the optimal distribution $q^\star(v_{j,fn})$ is Inverse-Gamma as already indicated in (22). Moreover we can identify its parameters:

$$\nu_v = \frac{\alpha_v+1}{2}; \quad (28)$$

$$\beta_{j,fn} = \frac{\alpha_v}{2}+\frac{\hat{s}_{j,fn}^2+\gamma_{j,fn}}{2\lambda_{j,fn}^2}. \quad (29)$$

2) E-U Step: Similarly, using (26) we have:

$$\ln q^\star(u_{ij}(t)) \stackrel{c}{=} -\ln(u_{ij}(t))\left(\frac{\alpha_u+1}{2}+1\right) - \frac{1}{u_{ij}(t)}\left(\frac{\alpha_u}{2}+\frac{\langle a_{ij}^2(t)\rangle_{q(a_{ij}(t))}}{2r^2(t)}\right), \quad (30)$$

where the parameters of this distribution as introduced in (24) are given by:

$$\nu_u = \frac{\alpha_u+1}{2}; \quad (31)$$

$$d_{ij}(t) = \frac{\alpha_u}{2}+\frac{\hat{a}_{ij}^2(t)+\rho_{ij}(t)}{2r^2(t)}. \quad (32)$$

[2]Due to the gamma and digamma functions involved in the VFE, we cannot cancel the partial derivative of the VFE for updating ν_v and ν_u.

3) E-S Step: Straightforward differentiation of the VFE with respect to the variance $\gamma_{j,fn}$ leads to the following update:

$$\gamma_{j,fn} = \left[\frac{\nu_v}{\beta_{j,fn}\lambda_{j,fn}^2}+\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\left(\sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau)+\|\hat{\mathbf{g}}_{ij,fn}\|_2^2\right)\right]^{-1}, \quad (33)$$

where $\hat{\mathbf{g}}_{ij,fn} = [\hat{g}_{ij,fn}(0), ..., \hat{g}_{ij,fn}(T-1)]^\top$ and $\cdot^\top$ denotes transposition.

Let us now focus on the update of the mean parameters $\{\hat{s}_{j,fn}\}_{j,f,n}$. Let $\hat{\mathbf{s}} \in \mathbb{R}^{JFN}$ denote the column vector of entries $\hat{s}_{j,fn}$ and $\Delta\hat{\mathbf{s}}$ the gradient of the VFE with respect to $\hat{\mathbf{s}}$. Let $\mathbf{x}_i = [x_i(0), ..., x_i(T-1)]^\top$ and $\hat{\mathbf{G}}_i \in \mathbb{R}^{T\times JFN}$ be the matrix formed by concatenating the column vectors $\hat{\mathbf{g}}_{ij,fn}$. We can show that the gradient is given as follows:

$$\Delta\hat{\mathbf{s}} = \mathbf{\Lambda}_{\hat{\mathbf{s}}}\hat{\mathbf{s}} - \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\hat{\mathbf{G}}_i^\top\mathbf{x}_i, \quad (34)$$

where for any arbitrary mapping[3] between the triplet (j,f,n) and $b \in \{1, ..., B\}$ with $B = \sum_{j=1}^{J} F_j N_j$, the matrix $\mathbf{\Lambda}_{\hat{\mathbf{s}}} \in \mathbb{R}^{JFN\times JFN}$ is defined as:

$$\mathbf{\Lambda}_{\hat{\mathbf{s}}} = \mathrm{diag}\left(\left\{\frac{\nu_v}{\beta_{j,fn}\lambda_{j,fn}^2}+\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau)\right\}_b\right) + \sum_{i=1}^{I}\frac{1}{\sigma_i^2}\hat{\mathbf{G}}_i^\top\hat{\mathbf{G}}_i. \quad (35)$$

$\mathrm{diag}(\{c(b)\}_b)$ denotes the diagonal matrix constructed from the sequence of coefficients $c(b)$. It is clear from (34) and (35) that canceling the gradient with respect to $\hat{\mathbf{s}}$ is equivalent to solving a symmetric positive semidefinite linear system of equations. For that purpose we employ the PCG method [56]. The preconditioning matrix is equal to the diagonal part of $\mathbf{\Lambda}_{\hat{\mathbf{s}}}$. We can show that $[\mathbf{\Lambda}_{\hat{\mathbf{s}}}]_{b,b} = \gamma_{j,fn}^{-1}$. It is interesting to note that each entry of the gradient will therefore be scaled by the variance $\gamma_{j,fn}$. This scaling compensates for the fact that changing the mean of a Gaussian distribution with a small variance has a much stronger impact than when the variance is large. Moreover we can mention that $\gamma_{j,fn}^{-1}$ is equal to the Fisher information defined by $\mathcal{I}(\hat{s}_{j,fn}) = \left\langle -\partial^2\left(\ln q^\star(s_{j,fn})\right)/(\partial\hat{s}_{j,fn}^2)\right\rangle_{q^\star(s_{j,fn})}$. The Fisher information is indeed used for characterizing the geometry of an optimization problem related to probability distributions [57] and it is involved in the computation of the natural gradient.

4) E-A Step: This step is very similar to the E-S step. Straightforward differentiation of the VFE with respect to the variance $\rho_{ij}(t)$ leads to the following update:

$$\rho_{ij}(t) = \left[\frac{\nu_u}{d_{ij}(t)r^2(t)}+\frac{1}{\sigma_i^2}\left(\|\hat{\mathbf{s}}_j\|_2^2+\sum_{(f,n)\in\mathcal{B}_j}\gamma_{j,fn}\right)\right]^{-1}, \quad (36)$$

where $\hat{\mathbf{s}}_j = [\hat{s}_j(0), ..., \hat{s}_j(L_s-1)]^\top$.

[3]This mapping must however be consistent with the construction of the vectors $\hat{\mathbf{s}}$, $\Delta\hat{\mathbf{s}} \in \mathbb{R}^{JFN}$ and the matrix $\hat{\mathbf{G}}_i \in \mathbb{R}^{T\times JFN}$.

Let us define $\hat{\mathbf{a}}_{ij} = [\hat{a}_{ij}(0), ..., \hat{a}_{ij}(L_a-1)]^\top$ and $\Delta\hat{\mathbf{a}}_{ij}$ the gradient of the VFE with respect to $\hat{\mathbf{a}}_{ij}$. We also introduce the following definitions:

- $T_{L_a}\{c(k)\}$: A symmetric Toeplitz [56] matrix constructed from the sequence $\{c(k)\}_{k=0}^{L_a-1}$;
- $\hat{r}_j^{ss}(k) = \sum_{t=0}^{L_s-1-k} \hat{s}_j(t)\hat{s}_j(t+k)$;
- $\hat{r}_{j,fn}^{\phi\phi}(k) = \sum_{t=0}^{L_s-1-k} \phi_{j,fn}(t)\phi_{j,fn}(t+k)$;
- $\epsilon_{ij}(t) = x_i(t) - \sum_{j'\neq j} \hat{y}_{ij'}(t)$;
- $\hat{r}_{ij}^{s\epsilon}(k) = \sum_{t=0}^{L_s-1} \hat{s}_j(t)\epsilon_{ij}(t+k)$.

We can show that the gradient is given as follows:

$$\Delta\hat{\mathbf{a}}_{ij} = \mathbf{\Lambda}_{\hat{\mathbf{a}}_{ij}}\hat{\mathbf{a}}_{ij} - \frac{1}{\sigma_i^2}\hat{\mathbf{r}}_{ij}^{s\epsilon}, \tag{37}$$

where $\hat{\mathbf{r}}_{ij}^{s\epsilon} = \left[\hat{r}_{ij}^{s\epsilon}(0), \hat{r}_{ij}^{s\epsilon}(1), \cdots, \hat{r}_{ij}^{s\epsilon}(L_a-1)\right]^\top$ and the matrix $\mathbf{\Lambda}_{\hat{\mathbf{a}}_{ij}} \in \mathbb{R}^{L_a\times L_a}$ is defined as:

$$\mathbf{\Lambda}_{\hat{\mathbf{a}}_{ij}} = \mathrm{diag}\left(\left\{\frac{\nu_u}{d_{ij}(t)r^2(t)}\right\}_t\right) + \frac{1}{\sigma_i^2}T_{L_a}\left\{\hat{r}_j^{ss}(k) + \sum_{(f,n)\in\mathcal{B}_j}\gamma_{j,fn}\hat{r}_{j,fn}^{\phi\phi}(k)\right\}. \tag{38}$$

Again we see that canceling the gradient with respect to $\hat{\mathbf{a}}_{ij}$ is equivalent to solving a symmetric positive semidefinite linear system of equations. We therefore employ the PCG method with the diagonal part of $\mathbf{\Lambda}_{\hat{\mathbf{a}}_{ij}}$ as a preconditioner. It is straightforward to verify that $[\mathbf{\Lambda}_{\hat{\mathbf{a}}_{ij}}]_{t,t} = \rho_{ij}(t)^{-1}$ because $\hat{r}_{j,fn}^{\phi\phi}(0) = 1$, the MDCT atoms forming an orthonormal basis, and $\hat{r}_j^{ss}(0) = \|\hat{\mathbf{s}}_j\|_2^2$. We can finally mention that as the update of $\hat{\mathbf{a}}_{ij}$ depends on $\{\hat{\mathbf{a}}_{ij'}\}_{j'\neq j}$, it is necessary to proceed sequentially.

D. M-Step

1) Noise Variance: In practice we will decrease the noise variance σ_i^2 progressively along the iterations. This parameter allows us to balance the contributions of the priors and the likelihood in the VFE. A relatively high variance can be interpreted as favoring the priors, which can be useful in the first iterations. On the contrary, decreasing the variance progressively will increase the contribution of the observed data in the estimation of the parameters. Such an approach has already been shown to be useful in source separation (see for example [17]). Alternatively, we can show that canceling the derivative of the VFE with respect to this parameter leads to the following update:

$$\sigma_i^2 = \frac{1}{T}\bar{e}_i, \tag{39}$$

where $\bar{e}_i$ is defined in Appendix B, equation (54).

2) NMF Parameters: We recall that $\lambda_{j,fn}^2 = [\mathbf{W}_j\mathbf{H}_j]_{fn}$. We can show that maximizing $\mathcal{L}(q^\star, \boldsymbol{\theta})$ with respect to $\mathbf{W}_j, \mathbf{H}_j$ under a non-negativity constraint is equivalent to minimizing the following cost function under the same constraint:

$$\mathcal{C}(\mathbf{W}_j, \mathbf{H}_j) = \sum\nolimits_{(f,n)\in\mathcal{B}_j} d_{IS}\left(\hat{p}_{j,fn}, [\mathbf{W}_j\mathbf{H}_j]_{fn}\right), \tag{40}$$

where $d_{IS}(x,y) = x/y - \ln(x/y) - 1$ is the Itakura-Saito (IS) divergence and

$$\hat{p}_{j,fn} = \frac{\hat{s}_{j,fn}^2 + \gamma_{j,fn}}{\beta_{j,fn}/\nu_v}. \tag{41}$$

This minimization problem can be solved by using the standard multiplicative update (MU) rules given in [11]:

$$\mathbf{H}_j \leftarrow \mathbf{H}_j \odot \frac{\mathbf{W}_j^\top\left[(\mathbf{W}_j\mathbf{H}_j)^{\odot-2}\odot\hat{\mathbf{P}}_j\right]}{\mathbf{W}_j^\top(\mathbf{W}_j\mathbf{H}_j)^{\odot-1}}; \tag{42}$$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j \odot \frac{\left[(\mathbf{W}_j\mathbf{H}_j)^{\odot-2}\odot\hat{\mathbf{P}}_j\right]\mathbf{H}_j^\top}{(\mathbf{W}_j\mathbf{H}_j)^{\odot-1}\mathbf{H}_j^\top}, \tag{43}$$

where $\odot$ denotes entry-wise multiplication and exponentiation, and $\hat{\mathbf{P}}_j \in \mathbb{R}_+^{F_j\times N_j}$ with $[\hat{\mathbf{P}}_j]_{fn} = \hat{p}_{j,fn}$ for all $(f,n)\in\mathcal{B}_j$. It has been shown in [58] that this procedure leads to a monotonic decrease of the IS divergence.

It is also interesting to note that by using (28) and (29) we can rewrite $\hat{p}_{j,fn}$ as follows:

$$\hat{p}_{j,fn} = \left(\frac{\alpha_v(\hat{s}_{j,fn}^2+\gamma_{j,fn})^{-1} + \lambda_{j,fn}^{-2}}{\alpha_v+1}\right)^{-1}. \tag{44}$$

This equation corresponds to a weighted harmonic mean between the posterior mean of the j-th source power spectrogram $\langle s_{j,fn}^2\rangle_q = \hat{s}_{j,fn}^2 + \gamma_{j,fn}$ and the current NMF parametrization $\lambda_{j,fn}^2$. Injecting (44) in (40) we obtain an optimization problem very similar to the one presented in the recent Student's t NMF framework [15]. The only differences come from the fact that we used the Student's *t* distribution defined over a real-valued r.v. (because we work in the MDCT domain), and we have an additional expectation $\langle s_{j,fn}^2\rangle_q$ because the true source power spectrogram is not directly observed. It is finally interesting to note that in the limit case where α_v tends to infinity (the Student's *t* tends to the Gaussian distribution), $\hat{p}_{j,fn}$ tends to $\hat{s}_{j,fn}^2 + \gamma_{j,fn}$ and we obtain the same optimization problem as in the popular IS-NMF framework [11].

IV. Experimental Evaluation

A. Baseline Methods

We present below the methods of the literature that will be used for comparison in the experimental evaluation.

1) Deterministic Time-Domain Mixing Filters: The first baseline method corresponds to our previous work [35] that has already been described in the last paragraph of Section I-B. In the following, the results using this method will be denoted by "Unconstrained time-domain filters".

2) Spatial Covariance Matrix Model: The two other baseline methods [28] and [18] exploit the following Gaussian multichannel source image model:

$$\mathbf{y}_{j,fn} \sim \mathcal{N}(0, \lambda_{j,fn}^2\mathbf{R}_{j,f}), \tag{45}$$

where $\mathbf{y}_{j,fn} = [y_{1j,fn}, ..., y_{Ij,fn}]^\top \in \mathbb{C}^I$ is the j-th multichannel source image, expressed in the STFT domain at TF bin (f,n), $\lambda_{j,fn}^2 \in \mathbb{R}_+$ is the short-term PSD of the source, which is further parametrized with an NMF model as in (8), and

$\mathbf{R}_{j,f} \in \mathbb{C}^{I\times I}$ is the frequency-dependent *spatial covariance matrix* (SCM) accounting for the spatial properties of the source. The full rank (i.e. rank 2 for stereophonic mixtures) SCM model was originally proposed in [28] for representing non-punctual sources. When the SCM is constrained to be rank 1, this model is equivalent to the punctual convolutive mixture model that was used for example in [17], [59].

The method proposed in [18] estimates the NMF parameters and the SCM by means of multiplicative update rules, derived by using the auxiliary function technique. In the following, this method will be denoted by "Sawada et al. - SCM rank 2".

In [28], the SCM is further constrained to have the following structure:

$$\mathbf{R}_{j,f} = \mathbf{A}_{j,f}\mathbf{A}_{j,f}^H, \tag{46}$$

where $\mathbf{A}_{j,f} \in \mathbb{C}^{I\times R_j}$ is a matrix of rank $0 < R_j \leq I$ and $(\cdot)^H$ denotes conjugate transpose. The punctual model corresponds to $R_j = 1$ while the non-punctual one corresponds to the full rank case $R_j = 2$ for stereophonic mixtures. In this method, the parameters are estimated by means of an expectation-maximization (EM) algorithm. We use here the implementation provided by the authors and available at [60]. In the following the results using this framework when the SCM is of rank 1 and 2 will be denoted by "Ozerov et al. - SCM rank 1" and "Ozerov et al. - SCM rank 2" respectively.

The source separation algorithms for the proposed and the baseline methods are run for 200 iterations. At each iteration of the proposed VEM algorithm, the PCG method in the E-S and E-A steps is run for 10 iterations while the MU rules for updating the NMF activation matrices are run for 20 iterations.

B. Experimental Setup

1) Database: The experiments are performed using musical audio source signals provided by the MTG MASS database [61]. We created 8 stereo mixtures sampled at 16 kHz using RIRs from the MIRD database [45]. These RIRs were measured in a room with adjustable reverberation level in order to obtain three different reverberation times: 160, 360 and 610 ms. Each mixture contains between 3 and 5 spatially disjoint sources and its duration ranges from 12 to 28 seconds. The musical instruments involved in this database are drums, piano, bass, guitar and voice.

2) Semi-blind Scenario: The main contribution of this paper is to propose a new Bayesian source separation framework where the mixing filters are treated as latent random variables in the time domain. This is why we are mostly interested in the evaluation of the mixing model. Therefore, we will assume some prior knowledge on the source signals. In order to keep a realistic scenario, we only learn for each source the NMF spectral templates (the matrix $\mathbf{W}_j$ using a rank of $K_j = 10$) from the true source signals. These "ideal pre-trained dictionaries" are then fixed during the VEM algorithm, only the activation matrices $\mathbf{H}_j$ are updated at the M-Step. We also assume that the reverberation time is known in order to fix the exponential decay profile $r(t)$ in (12). All other parameters are blindly estimated. This semi-blind scenario is used for the proposed method and the baselines.

In [18], the NMF parameters are shared across the sources and a latent variable is introduced in the source model in order to indicate to which source each NMF basis belongs. As we are here working in a semi-blind setting, this clustering information is known in advance.

3) Performance Measures: We evaluate the quality of the separation in terms of reconstructed stereo source images. We use standard energy ratios defined in [62] and expressed in decibels (dB): the signal-to-distortion (SDR), artifact (SAR) and interference (SIR) ratios and the source image-to-spatial distortion ratio (ISR). These measures are computed using the BSS Eval toolbox [63]. We also consider the Overall Perceptual Score (OPS) which is a perceptually motivated objective measure [64], [65]. It is expressed in percentage and computed using the PEASS toolbox [66].

C. Parameters Initialization

We present below how the parameters of the proposed and baseline methods are initialized.

1) Variational Parameters of the Proposed Method: Let us first focus on the proposed method. ν_v and ν_u are fixed according to (28) and (31) respectively. The other variational parameters are then initialized as follows: $\beta_{j,fn} = \alpha_v/2$ and $\gamma_{j,fn} = \tilde{v}_{j,fn}\lambda_{j,fn}^2$ where $\tilde{v}_{j,fn}$ is a realization from $\mathcal{IG}(\nu_v, \beta_{j,fn})$. $\hat{s}_{j,fn}$ is then initialized as a realization from $\mathcal{N}(0, \gamma_{j,fn})$. $d_{ij}(t) = \alpha_u/2$ and $\rho_{ij}(t) = \tilde{u}_{ij}(t)r^2(t)$ where $\tilde{u}_{ij}(t)$ is a realization from $\mathcal{IG}(\nu_u, d_{ij}(t))$. $\hat{a}_{ij}(t)$ is then initialized as a realization from $\mathcal{N}(0, \rho_{ij}(t))$. It is important to mention that we did not "optimize" this initialization procedure. It is just a simple way of initializing the approximate posterior distributions (22)-(25) similarly to the corresponding prior distributions in Section II.

2) Spatial Parameters of the Baselines: The initial spatial parameters of the baseline methods are set using the initialization of $\hat{a}_{ij}(t)$ (see previous paragraph) in order to provide the same "initial information". For our previous method [35] we exactly set the deterministic time-domain mixing filters to $\hat{a}_{ij}(t)$. For the two baseline methods using an SCM model (rank 1 or 2), we first compute $\hat{a}_{ij,f}$, the discrete Fourier transform of $\hat{a}_{ij}(t)$ with a number of points that is equal to the length of the STFT window. We then initialize the SCM of the j-th source as a rank-1 matrix built from the outer-product of $\hat{\mathbf{a}}_{j,f} = [\hat{a}_{1j,f}, ..., \hat{a}_{Ij,f}]^\top \in \mathbb{C}^I$ with itself.

3) NMF Parameters: As already mentioned we work in a semi-blind scenario where the dictionary matrices $\mathbf{W}_j$ are computed from the true source TF coefficients denoted by $\{\tilde{s}_{j,fn}\}_{f,n}$. Let us first focus on the proposed Student's t source model where $\tilde{s}_{j,fn} \in \mathbb{R}$. Similarly to [15] we can use a majorize-minimize approach to show that maximizing the likelihood with respect to the NMF parameters under a non-negativity constraint is equivalent to minimizing (40) under the same constraint with

$$\hat{p}_{j,fn} = \left(\frac{\alpha_v \tilde{s}_{j,fn}^{-2} + [\mathbf{W}_j\mathbf{H}_j]_{fn}^{-1}}{\alpha_v + 1}\right)^{-1}, \tag{47}$$

which is defined using the current value of the NMF parameters. For the baseline methods where the source model is

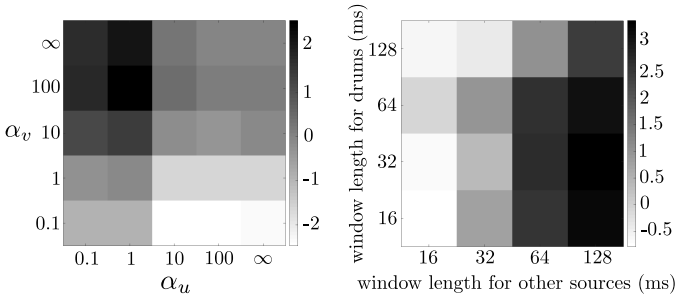

Fig. 5. Average SDR according to the shape parameters α_v and α_u (left plot) and according to the window length for the drums and other sources (right plot). The reverberation time of the mixtures is equal to 360 ms.

Gaussian (defined in the STFT or MDCT domains), we have $\hat{p}_{j,fn} = |\tilde{s}_{j,fn}|^2$. This minimization problem can be solved with the MU rules[4] given in [11] and recalled in (42)-(43). Finally, once the dictionary matrix $\mathbf{W}_j$ is estimated, we discard the obtained activation matrix $\mathbf{H}_j$ and re-initialize it with an all-one matrix.

D. Model Hyperparameters and Algorithm Parameters

The specific choices for model hyperparameters and algorithm parameters are described in this subsection. The experiments in this subsection have been performed using the mixtures associated with a reverberation time of 360 ms only, in order not to overfit on all mixing conditions.

1) Noise Variance: As explained in Section III-D1, instead of using (39) we decrease the noise variance σ_i^2 over the iterations from 10^{-2} to 10^{-6} by using the same schedule as in [17].

2) Scale Parameter of the Exponential Decay: Due to the well known scaling indeterminacy between the mixing filters and the source signals, the scale parameter σ_r in (12) can be arbitrarily fixed. We here choose $\sigma_r = 0.1$ so that the overall energy of the mixing filters is not too high.

3) Student's t Shape Parameters: We now study the influence on the source separation results of the shape parameters for the source and mixing Student's *t* models. For that purpose we consider a grid of values: $(\alpha_v, \alpha_u) \in \{0.1, 1, 10, 100, \infty\}^2$ (infinity actually corresponds to 4.5×10^{15}). We recall that when the shape parameter tends to infinity, the Student's *t* distribution is equivalent to the Gaussian one. We compute the SDR averaged over all the separated sources in the database for all the values in this grid. The results are represented in the left plot of Figure 5. It can be seen that the optimal values are $(\alpha_v, \alpha_u) = (100, 1)$. The fact that α_v is quite high tells us that the Gaussian assumption for the source model seems to be a reasonable choice. On the contrary we observe that it is very important to choose a small value for α_u, confirming that the Student's *t* is much more appropriate than the Gaussian for modeling RIRs.

4) Time-frequency Resolution: The MDCT window length was fixed to 64 ms for all sources in the experiment detailed in the previous paragraph. However the fact that we consider time-domain observations for inferring TF latent sources allows us to adapt the TF transform to each source in the mixture (see Section II-A). In particular we can adapt the MDCT window length according to the nature of the source. We investigate here a simple scenario: we consider a window length for the drums different from that of any other instrument. Once again we resort to a grid search which is presented in the right plot of Figure 5. The shape parameters for the Student's *t* source and mixing models were set according to the optimal values obtained in the previous paragraph. As could be expected, a short window is suitable for the drums (32 ms) while a long one is more appropriate for the other instruments (128 ms). This specificity of the model could be further investigated using for example unions of TF dictionaries with different resolutions for each source such as in [67]. However this is left for future work as this paper mostly focuses on proposing a new mixing model for reverberant mixtures.

[4]In order to provide a fair comparison of all the methods we kept the same random seed for initializing the NMF parameters before running the MU rules.

E. Comparison with the Baselines

We compare in this subsection the source separation performance obtained with the proposed method and the baselines introduced in Section IV-A for three different reverberation times: 160, 360 and 610 ms. The Student's *t* shape parameters for the proposed model are fixed according to the preliminary results detailed in Section IV-D3. In order to provide a fair comparison we will detail the results with the proposed method when the TF window length is the same as that of the baseline methods (64 ms for all sources) and when it is set according to the previous results: 32 ms for drums and 128 ms for other instruments. Those two approaches are denoted by "Proposed - w/o adapted TF window" and "Proposed - w/ adapted TF window" respectively in the presentation of the results.

The results are detailed in Table I. As can be seen, according to all the performance measures, our method performs better than the one proposed in [28], whether it uses an SCM of rank 1 or 2. This baseline method is algorithmically speaking close to ours, as both methods are based on a (V)EM algorithm. From this perspective, those results show that modeling reverberant mixtures in the time-domain, with suitable priors for the mixing filters, can help improving the source separation results.

However, we see that the method proposed in [18] performs better than ours in terms of SIR, and also in terms of SDR and ISR for the mixtures with a reverberation time of 160 and 610 ms. According to the SIR, this baseline method results in a lower amount of interferences between the estimated sources. However, as indicated by the SAR, this approach introduces more artifacts than the proposed one. Moreover, according to the OPS which is an objective measure of the global source separation quality designed to be correlated with the human perception, our method always performs better. The reader is invited to listen to the audio examples available online [68].

It is also interesting to note that the method [18] performs better than [28], even though the models are in essence equivalent (in the case of an SCM of rank 2). This may be due to the use of multiplicative updates derived from the auxiliary function method instead of the EM algorithm.

TABLE I
AVERAGE SOURCE SEPARATION RESULTS.

	SDR	ISR	SIR	SAR	OPS
Reverberation time: 160 ms					
Ozerov et al. [28] - SCM rank 1	2.5	6.4	3.6	9.8	18.8
Ozerov et al. [28] - SCM rank 2	2.4	6.2	3.5	9.7	18.8
Sawada et al. [18] - SCM rank 2	**4.3**	**9.1**	**8.4**	9.6	22.2
Unconstrained time-domain filters [35]	1.2	6.2	3.2	8.7	20.8
Proposed - w/o adapted TF window	2.8	6.9	4.3	10.6	20.8
Proposed - w/ adapted TF window	3.5	8.0	6.1	**12.0**	**27.6**
Reverberation time: 360 ms					
	SDR	ISR	SIR	SAR	OPS
Ozerov et al. [28] - SCM rank 1	1.9	5.7	2.4	9.4	16.6
Ozerov et al. [28] - SCM rank 2	1.8	5.6	2.5	9.5	16.6
Sawada et al. [18] - SCM rank 2	3.2	8.6	**8.2**	8.1	21.2
Unconstrained time-domain filters [35]	0.6	6.3	2.6	6.9	20.6
Proposed - w/o adapted TF window	2.6	7.4	3.5	10.7	22.4
Proposed - w/ adapted TF window	**3.4**	**8.7**	5.0	**12.5**	**28.2**
Reverberation time: 610 ms					
	SDR	ISR	SIR	SAR	OPS
Ozerov et al. [28] - SCM rank 1	1.8	5.4	2.4	9.4	14.6
Ozerov et al. [28] - SCM rank 2	1.7	5.3	2.2	9.5	14.7
Sawada et al. [18] - SCM rank 2	**4.1**	**8.6**	**8.0**	9.0	20.6
Unconstrained time-domain filters [35]	0.3	5.8	1.8	5.3	19.0
Proposed - w/o adapted TF window	3.2	8.2	4.9	10.3	23.8
Proposed - w/ adapted TF window	3.1	8.1	4.1	**11.6**	**24.8**

Our previous approach [35] with unconstrained time-domain mixing filters obtains the worst results according to the SDR, while the OPS tends to show that on average it performs better than [28] but worse than [18]. By listening to the separated sources, we believe that the results are indeed not satisfactory especially for high reverberation times, which agrees with the tendency of the SDR to decrease when the reverberation time increases. Even though the time-domain convolutive mixing process is exactly represented in this method, the mixing filters are only estimated from the data, without constraints. We believe that this lack of constraint precisely explains the poor results for long mixing filters. We noticed in particular that due to some inherent ambiguities in the modeling of the source images in (1), some parts of the source signals were sometimes contained in the estimated mixing filters. On the contrary, we do not have this issue with the proposed method, thanks to the prior on the mixing filters. This is also illustrated with online audio examples [68].

We also notice that even though the model hyperparameters of the proposed method (Student's *t* shape parameters and MDCT window length) have been optimized using the mixtures with a reverberation of 360 ms, they generalize well to other mixing conditions.

To conclude this experimental evaluation, we detail in Table II the computation time of one iteration of all the source separation algorithms compared in this work. Those results are obtained with a 3.70 GHz processor, for a stereo mixture of 12 s sampled at 16 kHz, involving 3 sources and recorded with a reverberation time of 360 ms. The methods based on a time-domain representation of the convolutive mixing process, as the one proposed in this paper, are the most computationally demanding.

For the sake of reproducibility, a Matlab implementation of the proposed algorithm is available online [68].

TABLE II
COMPUTATION TIME (IN SECONDS) OF ONE ITERATION OF THE SOURCE SEPARATION ALGORITHMS.

Ozerov et al. [28] - SCM rank 1	2.5
Ozerov et al. [28] - SCM rank 2	4.3
Sawada et al. [18] - SCM rank 2	0.8
Unconstrained time-domain filters [35]	20.9
Proposed - w/o adapted TF window	16.4
Proposed - w/ adapted TF window	17.7

V. CONCLUSION

This paper introduced a new Bayesian framework for multichannel audio source separation based on a time-domain representation of the convolutive mixing process. The experiments have revealed that the Gaussian distribution seems to be a reasonable choice for modeling the source signals in the MDCT domain. On the contrary, we have shown that using a heavy-tailed distribution is very important for modeling the impulse response of the mixing filters. Indeed, the robustness of the Student's *t* distribution allowed us to take into account the direct path and the early echoes of an RIR, which act as outliers with respect to the standard Gaussian model with exponentially decaying envelope. To conclude this paper, we present below some tracks for future research.

The model presented in Section II-D assumed that the RIR coefficients were i.i.d. However we could refine it by considering that the early and late parts of the RIR are not identically distributed. We have to mention that we investigated such an approach within the proposed framework. We used a time-dependent shape parameter $\alpha_u(t)$, allowing us to assume a Gaussian distribution for late reverberation only, while a Student's *t* model with finite shape parameter was used for the early contributions. As this approach did not improve the source separation results we did not present it. We believe that even though considering a time-dependent shape parameter allows us to better fit the true statistics of the mixing filters, it does not necessarily imply that it will lead to better separation results. It is for example known in blind deconvolution of images that the optimal image prior is not the one that most closely fits the statistics of natural images. It is rather the one that discriminates the blurry and sharp images as much as possible [69].

It could also be interesting to study whether the proposed mixing model is sufficiently flexible to be robust to mismatched mixing conditions such as inaccurate reverberation time. We could also develop other temporal profiles for $r(t)$ in order to represent different mixing scenarios, corresponding to the use of delays for example, which are popular audio effects in music.

In this work we considered a semi-blind scenario using oracle NMF dictionaries. Indeed our main objective was to present a new framework with a probabilistic model of mixing filters in the time domain. Nevertheless, we will focus on developing a "fully" blind method in future works. We carefully designed the experimental setup in order to be able to extend this framework in a supervised way. It could for example

consist in pre-training NMF dictionaries using a database such as MedleyDB [70]. One could also use more sophisticated NMF-based source models such as the ones in [28] (e.g. a source/filter model). In that case only the NMF update in the M-step of the proposed algorithm should be modified. Another important perspective lies in the use of neural networks for modeling the scale parameters $\lambda_{j,fn}$ in the source model (9). Such an approach was followed for the Gaussian SCM model in [22], [23].

Finally, although this method was designed for under-determined audio source separation, it is not limited to this setting and it can also be used in the (over-)determined case. In particular, it would be interesting to compare the proposed approach with the state-of-the-art determined source separation method recently proposed in [21], which unifies independent vector analysis and NMF in a new framework called independent low-rank matrix analysis.

APPENDIX A
THE STUDENT'S *t* DISTRIBUTION

The probability density function (pdf) of the Student's *t* distribution over a real-valued r.v. is defined by:

$$T_\alpha(x;\mu,\lambda)=\frac{1}{\sqrt{\alpha\pi\lambda^2}}\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\left(1+\frac{1}{\alpha}\frac{(x-\mu)^2}{\lambda^2}\right)^{-\frac{\alpha+1}{2}}, \tag{48}$$

where $\Gamma(\cdot)$ denotes the Gamma function. It can be shown (see e.g. [53]) that the Student's t distribution results from compounding a Gaussian distribution $\mathcal{N}(\mu,v\lambda^2)$ with an inverse-gamma distribution $\mathcal{IG}\left(\alpha/2,\alpha/2\right)$ over v:

$$T_\alpha(x;\mu,\lambda)=\int_0^{+\infty} N\left(x;\mu,v\lambda^2\right) IG\left(v;\frac{\alpha}{2},\frac{\alpha}{2}\right) dv. \tag{49}$$

The pdf of the Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$ over a real-valued r.v. is defined by:

$$N(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{50}$$

and the pdf of the inverse-gamma distribution $\mathcal{IG}(\alpha,\beta)$ over a positive real-valued r.v. is given by:

$$IG(x;\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-(\alpha+1)}\exp\left(\frac{-\beta}{x}\right). \tag{51}$$

The inverse-gamma distribution $\mathcal{IG}\left(\alpha/2,\alpha/2\right)$ over the r.v. v in (49) becomes degenerate at $v=1$ as α tends to infinity (i.e. the pdf tends to a Dirac delta function centered at one). Therefore, in this limit the Student's *t* distribution reduces to the Gaussian one.

The two following properties of the inverse-gamma distribution are used in the derivation of the proposed source separation algorithm: $\mathbb{E}[x^{-1}]=\alpha/\beta$ and $\mathbb{E}[\ln(x)]=\ln(\beta)-\psi(\alpha)$ where $\psi(\cdot)$ is the digamma function.

APPENDIX B
VARIATIONAL FREE ENERGY

The variational free energy can be decomposed from its definition in (18) as follows:

$$\begin{aligned}\mathcal{L}(q^\star,\boldsymbol{\theta})&=\left\langle\ln\left(\frac{p(\mathbf{x},\mathbf{s},\mathbf{v},\mathbf{a},\mathbf{u};\boldsymbol{\theta})}{q^\star(\mathbf{s})q^\star(\mathbf{v})q^\star(\mathbf{a})q^\star(\mathbf{u})}\right)\right\rangle_{q^\star(\mathbf{s})q^\star(\mathbf{v})q^\star(\mathbf{a})q^\star(\mathbf{u})}\\&=\left\langle\ln p(\mathbf{x}|\mathbf{s},\mathbf{a};\boldsymbol{\sigma})\right\rangle_{q^\star(\mathbf{s})q^\star(\mathbf{a})}\\&\quad+\left\langle\ln p(\mathbf{s}|\mathbf{v};\boldsymbol{\lambda})-\ln q^\star(\mathbf{s})\right\rangle_{q^\star(\mathbf{s})q^\star(\mathbf{v})}\\&\quad+\left\langle\ln p(\mathbf{v})-\ln q^\star(\mathbf{v})\right\rangle_{q^\star(\mathbf{v})}\\&\quad+\left\langle\ln p(\mathbf{a}|\mathbf{u})-\ln q^\star(\mathbf{a})\right\rangle_{q^\star(\mathbf{a})q^\star(\mathbf{u})}\\&\quad+\left\langle\ln p(\mathbf{u})-\ln q^\star(\mathbf{u})\right\rangle_{q^\star(\mathbf{u})}.\end{aligned} \tag{52}$$

Note that $<-\ln q(\cdot)>$ is the differential entropy of the distribution q. Let us detail each term in the right hand side of (52) from the model presented in Section II and the variational distributions given by equations (22) to (25). The objective is to express the variational free energy according to the variational parameters of these distributions. For the sake of simplicity of notation, we will omit the variational distribution in subscript of the mathematical expectation operator.

1) Likelihood Term:

$$\left\langle\ln p(\mathbf{x}|\mathbf{s},\mathbf{a};\boldsymbol{\sigma})\right\rangle=-\frac{IT}{2}\ln(2\pi)-\frac{T}{2}\sum_{i=1}^{I}\ln(\sigma_i^2)-\frac{1}{2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\bar{e}_i, \tag{53}$$

where $\bar{e}_i$ is given by:

$$\begin{aligned}\bar{e}_i=&\left\|\mathbf{x}_i-\sum_{j=1}^{J}\hat{\mathbf{y}}_{ij}\right\|_2^2+\sum_{j=1}^{J}\left[\|\hat{\mathbf{s}}_j\|_2^2\sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau)\right]\\&+\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}_j}\gamma_{j,fn}\left[\|\hat{\mathbf{g}}_{ij,fn}\|_2^2+\sum_{\tau=0}^{L_a-1}\rho_{ij}(\tau)\right],\end{aligned} \tag{54}$$

with $\hat{\mathbf{g}}_{ij,fn}=[\hat{g}_{ij,fn}(0),...,\hat{g}_{ij,fn}(T-1)]^\top$, $\hat{\mathbf{s}}_j=[\hat{s}_j(0),...,\hat{s}_j(L_s-1)]^\top$ and $\hat{\mathbf{y}}_{ij}=[\hat{y}_{ij}(0),...,\hat{y}_{ij}(T-1)]^\top$. Note that $\hat{s}_j(t)$ and $\hat{y}_{ij}(t)$ are related to the variational parameters $\{\hat{s}_{j,fn}\}_{j,f,n}$ and $\{\hat{a}_{ij}(t)\}_{i,j,t}$ by (20) and (21) respectively.

2) S-term:

$$\begin{aligned}\left\langle\ln p(\mathbf{s}|\mathbf{v};\boldsymbol{\lambda})-\ln q^\star(\mathbf{s})\right\rangle=&-\frac{1}{2}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}_j}\left[\ln\left(\frac{\beta_{j,fn}}{\gamma_{j,fn}}\right)\right.\\&\left.-\psi(\nu_v)+\ln(\lambda_{j,fn}^2)+\frac{\nu_v}{\beta_{j,fn}}\frac{\hat{s}_{j,fn}^2+\gamma_{j,fn}}{\lambda_{j,fn}^2}-1\right].\end{aligned} \tag{55}$$

3) V-term:

$$\langle \ln p(\mathbf{v}) - \ln q^{\star}(\mathbf{v}) \rangle = - \sum_{j=1}^{J} \sum_{(f,n)\in\mathcal{B}_j} \Bigg[\frac{\alpha_v \nu_v}{2} \frac{1}{\beta_{j,fn}} + \frac{\alpha_v}{2} \ln(\beta_{j,fn}) + \ln \Gamma\left(\frac{\alpha_v}{2}\right) + \frac{\alpha_v}{2} \ln\left(\frac{2}{\alpha_v}\right) + \psi(\nu_v)\left(\nu_v - \frac{\alpha_v}{2}\right) - \nu_v - \ln\Gamma(\nu_v) \Bigg]. \tag{56}$$

4) A-term:

$$\langle \ln p(\mathbf{a}|\mathbf{u}) - \ln q^{\star}(\mathbf{a}) \rangle = - \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=0}^{L_a-1} \Bigg[\ln\left(\frac{d_{ij}(t)}{\rho_{ij}(t)}\right) - \psi\left(\nu_u\right) + \ln(r^2(t)) + \frac{\nu_u}{d_{ij}(t)} \frac{\hat{a}_{ij}^2(t) + \rho_{ij}(t)}{r^2(t)} - 1 \Bigg]. \tag{57}$$

5) U-term:

$$\langle \ln p(\mathbf{u}) - \ln q^{\star}(\mathbf{u}) \rangle = - \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=0}^{L_a-1} \Bigg[\frac{\alpha_u \nu_u}{2} \frac{1}{d_{ij}(t)} + \frac{\alpha_u}{2} \ln(d_{ij}(t)) + \ln \Gamma\left(\frac{\alpha_u}{2}\right) + \frac{\alpha_u}{2} \ln\left(\frac{2}{\alpha_u}\right) + \psi(\nu_u)\left(\nu_u - \frac{\alpha_u}{2}\right) - \nu_u - \ln\Gamma(\nu_u) \Bigg]. \tag{58}$$

REFERENCES

[1] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 107–115, 2014.

[2] R. Gribonval and M. Zibulevsky, "Sparse Component Analysis," in *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, P. Comon and C. Jutten, Eds., 2010, pp. 367–420.

[3] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.

[4] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1818–1829, 2010.

[5] E. Vincent, "Complex nonconvex ℓ_p norm minimization for underdetermined source separation," in *Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, 2007, pp. 430–437.

[6] C. Févotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2174–2188, 2006.

[7] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. IGI Global, 2010, pp. 162–185.

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

[9] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 125–144, 2015.

[10] L. Benaroya, L. Mcdonagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Honk Kong, 2003, pp. 613–616.

[11] 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.

[12] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to nonnegative matrix factorisation for audio signal modelling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 1825–1828.

[13] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2015, pp. 1–5.

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

[15] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, 2016, pp. 51–55.

[16] P. Magron, R. Badeau, and A. Liutkus, "Lévy NMF for robust nonnegative source separation," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, United States, 2017, pp. 259–263.

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

[18] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 971–982, 2013.

[19] S. Arberet, A. Ozerov, N. Q. 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 *Proc. IEEE Int. Conf. Information Sciences, Signal Processing and their Applications (ISSPA)*, 2010, pp. 1–4.

[20] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, 2016, pp. 1–5.

[21] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[22] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[23] ——, "Multichannel music separation with deep neural networks," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Budapest, Hungary, 2016.

[24] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, 2000.

[25] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, 2007.

[26] D. A. Bies and C. H. Hansen, *Engineering noise control: theory and practice*. CRC press, 2009.

[27] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

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

[29] R. Badeau and M. D. Plumbley, "Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 11, pp. 1670–1680, 2014.

[30] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1305–1319, 2007.

[31] H. Attias, "New EM algorithms for source separation and deconvolution with a microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 5, 2003, pp. V – 297–300.

[32] X. Li, L. Girin, and R. Horaud, "Audio source separation based on convolutive transfer function and frequency-domain lasso optimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New Orleans, LA, USA, 2017, pp. 541–545.

[33] ——, "An EM algorithm for audio source separation based on the convolutive transfer function," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, United States, Oct. 2017, pp. 56–60.

[34] S. Arberet and P. Vandergheynst, "Reverberant audio source separation via sparse and low-rank modeling," *IEEE Signal Process. Lett.*, vol. 21, no. 4, pp. 404–408, 2014.

[35] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation: variational inference of time-frequency sources from time-domain observations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New Orleans, LA, USA, 2017, pp. 26–30.

[36] ——, "Semi-blind Student's t source separation for multichannel audio convolutive mixtures," in *Proc. European Signal Processing Conference (EUSIPCO)*, Kos, Greece, 2017, pp. 2323–2327.

[37] ——, "Separating time-frequency sources from time-domain convolutive mixtures using non-negative matrix factorization," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, United States, 2017, pp. 264–268.

[38] A. Benichoux, L. S. R. Simon, E. Vincent, and R. Gribonval, "Convex regularizations for the simultaneous recording of room impulse responses," *IEEE Trans. Signal Process.*, vol. 62, no. 8, pp. 1976–1986, 2014.

[39] R. Giri, "Bayesian sparse signal recovery using scale mixtures with applications to speech," Ph.D. dissertation, UC San Diego, 2016.

[40] R. Badeau, "Preservation of whiteness in spectral and time-frequency transforms of second order processes," Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, Research Report, 2016.

[41] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 1, pp. 99–102, 1974.

[42] J. Palmer, K. Kreutz-Delgado, B. D. Rao, and D. P. Wipf, "Variational EM algorithms for non-Gaussian latent variable models," in *Proc. Adv. Neural Information Process. Syst. (NIPS)*, 2006, pp. 1059–1066.

[43] T. Adali, P. J. Schreier, and L. L. Scharf, "Complex-valued signal processing: The proper way to deal with impropriety," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5101–5125, 2011.

[44] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 8, pp. 1699–1712, 2013.

[45] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Antibes - Juan les Pins, France, 2014, pp. 313–317.

[46] T. Schultz, "Diffusion in reverberation rooms," *Journal of Sound and Vibration*, vol. 16, no. 1, pp. 17–28, 1971.

[47] M. R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *Journal of the Audio Engineering Society*, vol. 35, no. 5, pp. 299–306, 1987.

[48] ——, "Frequency-correlation functions of frequency responses in rooms," *Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1819–1823, 1962.

[49] J. A. Moorer, "About this reverberation business," *Computer music journal*, vol. 3, no. 2, pp. 13–28, 1979.

[50] J.-D. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Université du Maine, 1988.

[51] M. R. Schroeder and K. Kuttruff, "On frequency response curves in rooms. Comparison of experimental, theoretical, and Monte Carlo results for the average frequency spacing between maxima," *Journal of the Acoustical Society of America*, vol. 34, no. 1, pp. 76–80, 1962.

[52] J.-D. Polack, "Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics," *Applied Acoustics*, vol. 38, no. 2-4, pp. 235–244, 1993.

[53] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[54] J. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, no. Apr., pp. 661–694, 2005.

[55] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[56] G. H. Golub and C. F. Van Loan, *Matrix computations*. Johns Hopkins University Press, 1996.

[57] A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen, "Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes," *Journal of Machine Learning Research*, vol. 11, no. Nov., pp. 3235–3268, 2010.

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

[59] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, 2011, pp. 257–260.

[60] E. Vincent, A. Ozerov, and F. Bimbot, "Flexible Audio Source Separation Toolbox (FAAST) version 1 for Matlab," http://bass-db.gforge.inria.fr/fasst/, 2011.

[61] M. Vinyes, "MTG MASS dataset," http://mtg.upf.edu/download/datasets/mass, 2008.

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

[63] E. Vincent, "BSS Eval Toolbox Version 3.0 for Matlab," http://bass-db.gforge.inria.fr/bss_eval/, 2007.

[64] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2046–2057, 2011.

[65] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *Proc. Int. Conf. Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel Aviv, Israel, 2012, pp. 430–437.

[66] V. Emiya and E. Vincent, "PEASS Toolbox Version 2.0 for Matlab," http://bass-db.gforge.inria.fr/peass/, 2011.

[67] F. Feng and M. Kowalski, "Hybrid model and structured sparsity for under-determined convolutive audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Florence, Italy, 2014, pp. 6682–6686.

[68] S. Leglaive, "Companion website," https://sleglaive.github.io/demoSSMM-TASLP-2018.html, 2018.

[69] D. Wipf and H. Zhang, "Revisiting Bayesian blind deconvolution," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3595–3634, 2014.

[70] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research." in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2014, pp. 155–160.

Simon Leglaive received the State Engineering degree from Télécom ParisTech, Paris, France, in 2014, along with the M.Sc. degree in acoustics, signal processing and computer science applied to music (ATIAM) from the Université Pierre et Marie Curie (UPMC, Paris VI), Paris, France. In December 2017, he received the Ph.D. degree from Télécom ParisTech in the field of signal processing. Since February 2018 he is a post-doctoral researcher at Inria Grenoble Rhône-Alpes, in the Perception team. His research interests include statistical audio signal processing, statistical room acoustics, audio source separation and machine learning applied to audio signal processing.

Roland Badeau (M'02-SM'10) received the State Engineering degree from the École Polytechnique, Palaiseau, France, in 1999, the State Engineering degree from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2001, the M.Sc. degree in applied mathematics from the École Normale Supérieure (ENS), Cachan, France, in 2001, and the Ph.D. degree from the ENST in 2005, in the field of signal processing. He received the ParisTech Ph.D. Award in 2006, and the Habilitation degree from the Université Pierre et Marie Curie (UPMC), Paris VI, in 2010.

In 2001, he joined the Image, Data, Signal Department of LTCI, Télécom ParisTech, as an Assistant Professor, where he became Associate Professor in 2005. His research interests focus on statistical modeling of non-stationary signals (including adaptive high resolution spectral analysis and Bayesian extensions to NMF), with applications to audio and music (source separation, denoising, dereverberation, multipitch estimation, automatic music transcription, audio coding, audio inpainting). He is a co-author of 30 journal papers, over 100 international conference papers, and 4 patents. He is also an Associate Editor of the EURASIP Journal on Audio, Speech, and Music Processing and the IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Gaël Richard (SM'06, F'17) received the State Engineering degree from Télécom ParisTech, France in 1990, the Ph.D. degree from LIMSI-CNRS, University of Paris-XI, in 1994 in speech synthesis, and the Habilitation à Diriger des Recherches degree from the University of Paris XI in September 2001. After the Ph.D. degree, he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches for speech production. From 1997 to 2001, he successively worked for Matra, Bois d'Arcy, France, and for Philips, Montrouge, France. In particular, he was the Project Manager of several large scale European projects in the field of audio and multimodal signal processing. In September 2001, he joined Télécom ParisTech, where he is now a Full Professor in audio signal processing and Head of the Image, Data, Signal department. He is a coauthor of over 220 papers and inventor in 9 patents. He was an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing between 1997 and 2011 and one of the guest editors of the special issue on Music Signal Processing of IEEE Journal on Selected Topics in Signal Processing (2011). He currently is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee, member of the EURASIP and AES and fellow of the IEEE.