

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

Simon Leglaive, Roland Badeau, Gaël Richard

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. In music it is very common to encounter mixtures with multiple channels as recordings are frequently done with several microphones and music is usually produced in a stereo format.

The source separation problem is generally defined according to the number of sources and microphones, and the nature of the mixture. In this paper we consider the *under-determined* case, meaning that we aim to separate more sources than the number of channels in the mixture. Moreover we focus on separating reverberant mixtures; the mixing process is said to be *convolutive* in that case. Due to the under-determined nature of the problem, model-based approaches have been used in the literature in order to incorporate information about the sources and the mixing process. In this section we introduce some of the source and mixing models that have been previously proposed before presenting our main contributions.

A. Source Modeling

The source signals are usually represented in a time-frequency (TF) domain in order to obtain a sparse representation which reveals the very specific structure of audio and

particularly musical signals [1]. This structure can then be exploited within the source model. Sparse component analysis (SCA) [2] methods for instance exploit the sparsity of the source signals in the TF domain, using for example sparsity inducing penalties [3], [4] or super-Gaussian distributions [5], [6]. They also generally take advantage of the spatial diversity of the sources, relying on anechoic propagation models [7], [8] or spatial indices such as the interaural time and loudness differences [9]. SCA methods usually assume that the TF source coefficients are independent and identically distributed, thus ignoring any specific TF structure. On the contrary, some approaches model the TF source coefficients by means of a non-stationary probability distribution. For example they can be represented as centered Gaussian random variables with a TF-dependent variance. This approach is referred to as the *variance modeling framework* in [10]. In addition, a piece of music is generally composed of musical events that repeat over time such as tonal musical notes or percussive sounds. In audio source separation it is very relevant to take this redundancy into account, precisely because of the under-determined nature of the problem. It allows us to reduce the number of parameters involved in the modeling of the sources. A common approach consists in assuming that the short-term (magnitude or power) spectrum of a source follows a compositional model [11]. Under Gaussianity and local stationarity assumptions [12], this approach aims to model the short-term power spectral density (PSD) of the source signals, represented by the TF-dependent variance. In this context non-negative matrix factorization (NMF) techniques are very popular. The non-negativity constraint of this kind of factorization generally leads to a meaningful decomposition [11]. The idea of using a non-negative decomposition for representing the short-term PSD of audio sources has first been proposed in [13]. A generative Gaussian model was then developed in [14]. In the recent literature many different probability distributions have been studied for audio signal modeling based on NMF [15]–[19].

B. Convolutive Mixing Process

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 channel 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*.

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).

More formally, let $s_j(t) \in \mathbb{R}$, $t = 0, \dots, L_s - 1$, $j = 1, \dots, J$, be the j -th source signal and $a_{ij}(t) \in \mathbb{R}$, $t = 0, \dots, L_a - 1$, $i = 1, \dots, I$, 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 $t = 0, \dots, T - 1$ with $T = L_s + L_a - 1$ as follows:

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

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

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

As previously mentioned, the source signals are usually represented in a TF domain. Consequently 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) . Such an approach has for example been followed in [20] 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 [21], [22]. 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 [23]. On the contrary the length of the STFT window is typically short (between 30 and 120 ms) in order to fit the local stationarity of the signal over the support of the 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 was developed in [24] to represent non-punctual sources. The authors introduced a *spatial covariance matrix* that encodes the correlations between the different channels.

In [25] 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 [26] when the filter bank corresponds to the STFT. In [27], 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 [28] and [29].

Finally, the methods in [4] and [30] 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 [31], 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 [32] we extended this approach by using a Student's t distribution for modeling the source MDCT coefficients¹. We considered a semi-blind setting assuming that the mixing filters were known and fixed. Finally, in [33] we generalized our approach in [31] 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 [34] [35, 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.

¹Note 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 $\mathcal{B}_j = \{0, \dots, F_j - 1\} \times \{0, \dots, N_j - 1\}$ as follows:

$$s_j(t) = \sum_{(f,n) \in \mathcal{B}_j} s_{j,fn} \psi_{j,fn}(t). \quad (4)$$

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

$$\begin{aligned} \psi_{j,fn}(t) = & \sqrt{\frac{4}{M_j}} w_j(t - nH_j) \\ & \cos\left(\frac{2\pi}{M_j} \left(t - nH_j + \frac{1}{2} + \frac{M_j}{4}\right) \left(f + \frac{1}{2}\right)\right), \end{aligned} \quad (5)$$

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 [33] that using the STFT does not improve the source separation performance. We can also mention that in [36] 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 $g_{ij,fn}(t) = [a_{ij} \star \psi_{j,fn}](t)$ as follows:

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

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

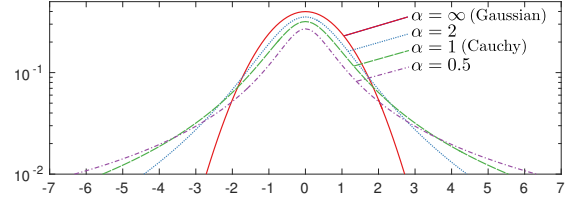


Fig. 1. Pdf of the $\mathcal{T}_\alpha(0, 1)$ distribution (in logarithmic scale) for different values of the shape parameter α .

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.

Let $\mathcal{T}_\alpha(0, \lambda)$ denote the centered Student's t distribution over a real-valued random variable (r.v.) where $\alpha > 0$ and $\lambda > 0$ are respectively the shape parameter (also called degrees of freedom) and scale parameter. Its probability density function (pdf) is given by:

$$T_\alpha(x; 0, \lambda) = \frac{1}{\sqrt{\alpha\pi\lambda^2}} \frac{\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{\alpha}{2})} \left(1 + \frac{x^2}{\alpha\lambda^2}\right)^{-\frac{\alpha+1}{2}}, \quad (7)$$

where $\Gamma(\cdot)$ denotes the Gamma function. It can be shown (see e.g. [37]) that the Student's t distribution results from compounding a centered Gaussian distribution $\mathcal{N}(0, v\lambda^2)$ with an inverse-gamma distribution $\mathcal{IG}(\alpha/2, \alpha/2)$ over v :

$$T_\alpha(x; 0, \lambda) = \int_0^{+\infty} \mathcal{N}(x; 0, v\lambda^2) \mathcal{IG}\left(v; \frac{\alpha}{2}, \frac{\alpha}{2}\right) dv. \quad (8)$$

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

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad (9)$$

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

$$\mathcal{IG}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right). \quad (10)$$

In the sequel we will use the two following properties of the inverse-gamma distribution: $\mathbb{E}[x^{-1}] = \alpha/\beta$ and $\mathbb{E}[\ln(x)] = \ln(\beta) - \text{di}\Gamma(\alpha)$ where $\text{di}\Gamma(\cdot)$ is the digamma function.

Equation (8) shows that the Student's t distribution corresponds to a scale mixture of Gaussians (SMoG) [38]. Consequently, it is a super-Gaussian distribution [39] which exhibits thicker tails than the Gaussian as can be seen in Figure 1. The inverse-gamma distribution $\mathcal{IG}(\alpha/2, \alpha/2)$ over the r.v. v becomes degenerate at $v = 1$ as α tends to infinity (i.e. the pdf tends to a Dirac impulse centered at one). Therefore, in this limit the Student's t distribution reduces to the Gaussian one. 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 (8) can be interpreted in the following way: realizations of x are globally expected to be distributed according to a centered 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 [14], 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:

$$\lambda_{j,fn}^2 = [\mathbf{W}_j \mathbf{H}_j]_{fn}, \quad (11)$$

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 $K_j(F_j + N_j) \ll F_j N_j$. \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 [40]. However it can be used similarly for a source representation based on the real-valued MDCT as already proposed in [41] 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 [14] 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 [16], the symmetric alpha-stable [17] and the Student's t [18] 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 [18].

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:

$$s_{j,fn} \sim \mathcal{T}_{\alpha_v}(0, \lambda_{j,fn}). \quad (12)$$

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

$$\begin{cases} s_{j,fn} | v_{j,fn} & \sim \mathcal{N}(0, v_{j,fn} \lambda_{j,fn}^2); \\ v_{j,fn} & \sim \mathcal{IG}\left(\frac{\alpha_v}{2}, \frac{\alpha_v}{2}\right). \end{cases} \quad (13)$$

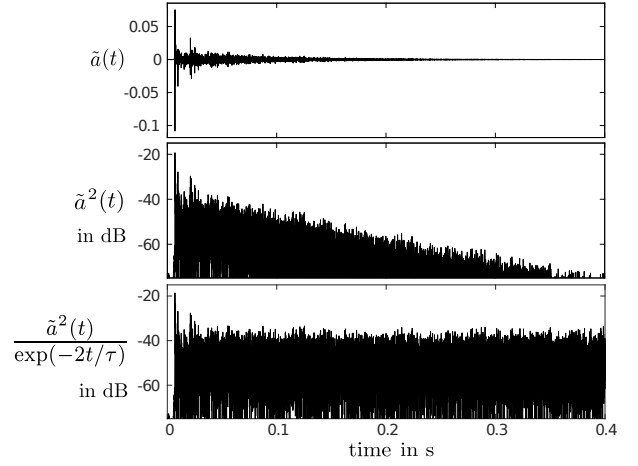


Fig. 2. Amplitude (top), power (middle) and normalized power (bottom) of the first 400 ms of an RIR from the MIRD database [42]. The RIR was recorded in a room with a reverberation time of 610 ms and with a source-to-microphone distance of 2 m.

We further assume that the squared scale parameters are structured by means of an NMF model as in (11). 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 acoustical 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 [42]. This RIR was recorded in a room with a reverberation time of 610 ms. The source-to-microphone distance was approximately 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. It means that the energy density is uniformly distributed across the room and over all directions [43].

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 $\mathbb{E}[a^2(t)]$ [44], [45]. 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 [46] that concert halls impulse responses sounded very similarly to white noise with an exponentially decaying envelope. In [47], 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:

$$a(t) \sim \mathcal{N}(0, r^2(t)) \quad (14)$$

with

$$r(t) = \sigma_r \exp(-t/\tau), \quad \tau = \frac{T_{60} f_s}{3 \ln(10)}. \quad (15)$$

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 the properties of the Fourier transform and Schroeder's model of room frequency response (RFR) [45], [48]. In this model, the RFR is represented as a random process whose real and imaginary parts are independent centered and stationary Gaussian processes with the same variance.

3) *Monte Carlo Experiment*: In [49] 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 (14) 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 (15) 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.

From the Gaussian model (14) we should expect the normalized RIR coefficients $a(t)/\exp(-t/\tau)$ to be independent and identically distributed (i.i.d) according to $\mathcal{N}(0, \sigma_r^2)$. We verify this hypothesis by means of a Monte-Carlo experiment using a set of RIRs from the MIRD database [42]. They 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 from which we compute the empirical pdf of the normalized RIR coefficients as displayed in Figure 3 (black solid line). We also fit from these data a Gaussian distribution using a maximum likelihood approach. The estimated pdf is represented with a blue dotted line in the same figure (the estimated σ_r is 0.006). 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 (14) as already noticed from the bottom plot of Figure 2.

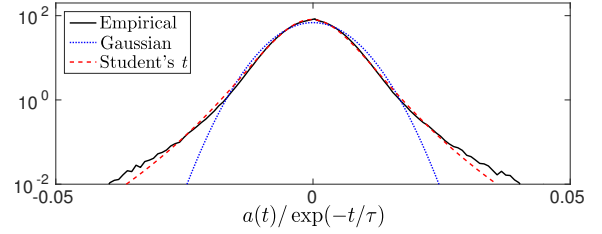


Fig. 3. Empirical pdf (black solid line) and pdf of the Gaussian (blue dotted line) and Student's t (red dashed line) distributions fitted on 624 normalized RIRs from the MIRD database [42]. The RIRs were recorded in a room with a reverberation time of 610 ms.

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 illustrate this point we also fitted a Student's t distribution from the set of normalized RIR coefficients. The estimated pdf is represented in Figure 3 with a red dashed line. The maximum likelihood estimation led to $\sigma_r = 0.005$ and a shape parameter that is equal to 7.2. 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 experiment we assume that the mixing filter coefficients $a_{ij}(t)$ independently follow a Student's t distribution:

$$a_{ij}(t) \sim \mathcal{T}_{\alpha_u}(0, r(t)), \quad (16)$$

where $r(t)$ models the exponential decay as defined in (15). This model admits Polack's Gaussian one in (14) as a special case when $\alpha_u \rightarrow \infty$. It can be rewritten in a hierarchical manner as follows:

$$\begin{cases} a_{ij}(t) | u_{ij}(t) & \sim \mathcal{N}(0, u_{ij}(t) r^2(t)); \\ u_{ij}(t) & \sim \mathcal{IG}\left(\frac{\alpha_u}{2}, \frac{\alpha_u}{2}\right). \end{cases} \quad (17)$$

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:

$$x_i(t) | \mathbf{s}, \mathbf{a}; \sigma_i^2 \sim \mathcal{N}\left(\sum_{j=1}^J y_{ij}(t), \sigma_i^2\right), \quad (18)$$

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

III. VARIATIONAL INFERENCE

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

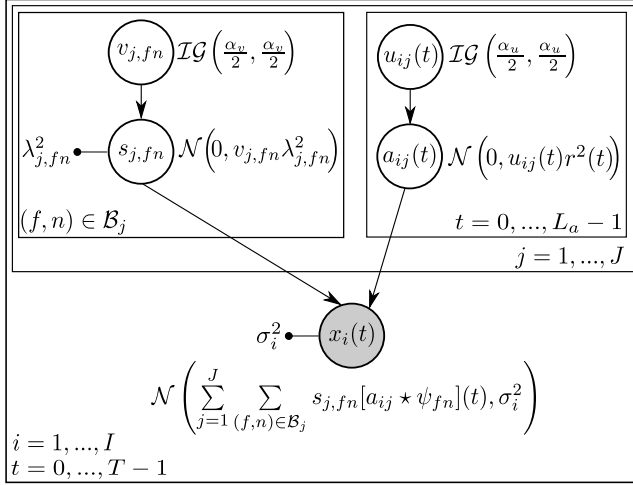


Fig. 4. Bayesian network corresponding to the proposed model. Latent random variables are represented with empty circles, observations with shaded circles, and the deterministic model parameters with dots. Rectangles denote the plate notation. The sub-graph contained in each rectangle is repeated according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

Performing inference aims to compute the posterior distribution of the latent variables $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ 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 [37].

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^* in the variational family \mathcal{F} that minimizes the Kullback-Leibler (KL) divergence from the true posterior:

$$q^*(\mathbf{z}) = \arg \min_{q(\mathbf{z}) \in \mathcal{F}} D_{\text{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})), \quad (19)$$

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

$$D_{\text{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) = \ln p(\mathbf{x};\boldsymbol{\theta}) - \mathcal{L}(q;\boldsymbol{\theta}), \quad (20)$$

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

$$\mathcal{L}(q;\boldsymbol{\theta}) = \langle \ln(p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})/q(\mathbf{z})) \rangle_q. \quad (21)$$

$\ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$ is referred to as the complete data log-likelihood. From (20) 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 [37] which consists in iterating two steps until convergence: the E-step where we compute $q^* = \arg \max_{q \in \mathcal{F}} \mathcal{L}(q;\boldsymbol{\theta}^*)$ and the M-step where we compute $\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} \mathcal{L}(q^*;\boldsymbol{\theta})$.

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:

$$q(\mathbf{z}) = \prod_{j=1}^J \left[\prod_{(f,n) \in \mathcal{B}_j} q_{j,fn}^s(s_{j,fn}) q_{j,fn}^v(v_{j,fn}) \right] \left[\prod_{i=1}^I \prod_{t=0}^{L_a-1} q_{ijt}^a(a_{ij}(t)) q_{ijt}^u(u_{ij}(t)) \right]. \quad (22)$$

For simplicity of notation we will drop the superscripts and indices when referring to one of the factors in (22).

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:

$$\hat{s}_j(t) = \sum_{(f,n) \in \mathcal{B}_j} \hat{s}_{j,fn} \psi_{j,fn}(t). \quad (23)$$

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 $\hat{g}_{ij,fn}(t) = [\hat{a}_{ij} \star \psi_{j,fn}](t)$. Finally, the estimate of the j -th source image as seen by microphone i is given by:

$$\begin{aligned} \hat{y}_{ij}(t) &= [\hat{a}_{ij} \star \hat{s}_j](t) \\ &= \sum_{(f,n) \in \mathcal{B}_j} \hat{s}_{j,fn} \hat{g}_{ij,fn}(t). \end{aligned} \quad (24)$$

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* [50]. 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 $q^*(z)$, $z \in \mathbf{z}$, will have the same form as the prior distribution of z conditionally on its parents [50]. For our model we consequently have:

$$q^*(v_{j,fn}) = IG(v_{j,fn}; \nu_v, \beta_{j,fn}); \quad (25)$$

$$q^*(s_{j,fn}) = N(s_{j,fn}; \hat{s}_{j,fn}, \gamma_{j,fn}); \quad (26)$$

$$q^*(u_{ij}(t)) = IG(u_{ij}(t); \nu_u, d_{ij}(t)); \quad (27)$$

$$q^*(a_{ij}(t)) = N(a_{ij}(t); \hat{a}_{ij}(t), \rho_{ij}(t)). \quad (28)$$

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 (25) to (28). In order not to vainly overload this section with equations, we choose to detail the expression of the VFE in the Appendix.

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) [37], [51]. It consists in maximizing the VFE by cycling over each individual variational parameter. We can indeed show that the optimal distribution $q^*(z)$, $z \in \mathbf{z}$, should satisfy [37]:

$$\ln q^*(z) = \langle \ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \rangle_{q(\mathbf{z}|z)} + \text{constant}, \quad (29)$$

where $\mathbf{z}|z$ denotes the set of all latent variables but z . This technique aims to identify the distribution of the individual factors $q^*(z)$ along with the expression of its parameters by developing the right hand side of (29). 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 (29) only involves the random variables belonging to the Markov blanket $MB(z)$ of z , i.e. its parents, children and co-parents [50]. This inference method therefore leads to a set of coupled solutions, as the variational parameters for $q^*(z)$ 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 (26) and (28), we have to solve two quadratic and convex optimization problems to individually update the set of variational parameters $\{\hat{s}_{j,fn}\}_{j,fn}$ and $\{\hat{a}_{ij}(t)\}_{i,j,t}$ (see the form of the VFE defined in the Appendix). The preconditioned conjugate gradient (PCG) method is a natural choice for solving such an optimization problem [52]. The PCG method has already been used for variational inference in [53]. The preconditioner was chosen as the Fisher information matrix so the preconditioned gradient was corresponding to the natural gradient.

Finally, all the other variational parameters in our model will be updated relying on (29), or equivalently by differentiating the VFE when appropriate².

1) *E-V Step*: We can develop (29) for the latent variable $v_{j,fn}$ as follows:

$$\begin{aligned} \ln q^*(v_{j,fn}) \stackrel{c}{=} & -\ln(v_{j,fn}) \left(\frac{\alpha_v + 1}{2} + 1 \right) \\ & - \frac{1}{v_{j,fn}} \left(\frac{\alpha_v}{2} + \frac{\langle s_{j,fn}^2 \rangle_{q(s_{j,fn})}}{2\lambda_{j,fn}^2} \right), \end{aligned} \quad (30)$$

where $\stackrel{c}{=}$ denotes equality up to an additive constant. From this equation we indeed recognize that the optimal distribution $q^*(v_{j,fn})$ is Inverse-Gamma as already indicated in (25). Moreover we can identify its parameters:

$$\nu_v = \frac{\alpha_v + 1}{2}; \quad (31)$$

$$\beta_{j,fn} = \frac{\alpha_v}{2} + \frac{\hat{s}_{j,fn}^2 + \gamma_{j,fn}}{2\lambda_{j,fn}^2}. \quad (32)$$

²Due to the gamma and digamma functions involved in the VFE, the only simple way of identifying ν_v and ν_u is to rely on (29). On the contrary, the other variational parameters can be identified by differentiating this criterion.

2) *E-U Step*: Similarly, using (29) we have:

$$\begin{aligned} \ln q^*(u_{ij}(t)) \stackrel{c}{=} & -\ln(u_{ij}(t)) \left(\frac{\alpha_u + 1}{2} + 1 \right) \\ & - \frac{1}{u_{ij}(t)} \left(\frac{\alpha_u}{2} + \frac{\langle a_{ij}^2(t) \rangle_{q(a_{ij}(t))}}{2r^2(t)} \right), \end{aligned} \quad (33)$$

where the parameters of this distribution as introduced in (27) are given by:

$$\nu_u = \frac{\alpha_u + 1}{2}; \quad (34)$$

$$d_{ij}(t) = \frac{\alpha_u}{2} + \frac{\hat{a}_{ij}^2(t) + \rho_{ij}(t)}{2r^2(t)}. \quad (35)$$

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

$$\gamma_{j,fn} = \left[\frac{\nu_v}{\beta_{j,fn}\lambda_{j,fn}^2} + \sum_{i=1}^I \frac{1}{\sigma_i^2} \left(\sum_{\tau=0}^{L_a-1} \rho_{ij}(\tau) + \|\hat{\mathbf{g}}_{ij,fn}\|_2^2 \right) \right]^{-1}, \quad (36)$$

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

Let us now focus on the update of the mean parameters $\{\hat{s}_{j,fn}\}_{j,fn}$. 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 $\mathbf{x}_i = [x_i(0), \dots, x_i(T-1)]^\top$ 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:

$$\Delta\hat{\mathbf{s}} = \Lambda_{\hat{\mathbf{s}}}\hat{\mathbf{s}} - \sum_{i=1}^I \frac{1}{\sigma_i^2} \hat{\mathbf{G}}_i^\top \mathbf{x}_i, \quad (37)$$

where for any arbitrary mapping³ between the triplet (j, f, n) and $b \in \{1, \dots, B\}$ with $B = \sum_{j=1}^J F_j N_j$, the matrix $\Lambda_{\hat{\mathbf{s}}} \in \mathbb{R}^{JFN \times JFN}$ is defined as:

$$\begin{aligned} \Lambda_{\hat{\mathbf{s}}} = \text{diag} \left(\left\{ \frac{\nu_v}{\beta_{j,fn}\lambda_{j,fn}^2} + \sum_{i=1}^I \frac{1}{\sigma_i^2} \sum_{\tau=0}^{L_a-1} \rho_{ij}(\tau) \right\} \right) \\ + \sum_{i=1}^I \frac{1}{\sigma_i^2} \hat{\mathbf{G}}_i^\top \hat{\mathbf{G}}_i. \end{aligned} \quad (38)$$

$\text{diag}(\{c(b)\}_b)$ denotes the diagonal matrix constructed from the sequence of coefficients $c(b)$. It is clear from (37) and (38) 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 [52]. The preconditioning matrix is equal to the diagonal part of $\Lambda_{\hat{\mathbf{s}}}$. We can show that $[\Lambda_{\hat{\mathbf{s}}}]_{b,b} = \gamma_{j,fn}^{-1}$. 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

³This 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 $\hat{\mathbf{G}}_i \in \mathbb{R}^{T \times JFN}$.

mention that $\gamma_{j,fn}^{-1}$ is equal to the Fisher information defined by $\mathcal{I}(\hat{s}_{j,fn}) = \left\langle -\partial^2 (\ln q^*(s_{j,fn})) / (\partial \hat{s}_{j,fn}^2) \right\rangle_{q^*(s_{j,fn})}$. The Fisher information is indeed used for characterizing the geometry of an optimization problem related to probability distributions [53] 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:

$$\rho_{ij}(t) = \left[\frac{\nu_u}{d_{ij}(t)r^2(t)} + \frac{1}{\sigma_i^2} \left(\|\hat{\mathbf{s}}_j\|_2^2 + \sum_{(f,n) \in \mathcal{B}_j} \gamma_{j,fn} \right) \right]^{-1}, \quad (39)$$

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

Let us define $\hat{\mathbf{a}}_{ij} = [\hat{a}_{ij}(0), \dots, \hat{a}_{ij}(L_a - 1)]^\top$ 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 [52] matrix constructed from the sequence $\{c(k)\}_{k=0}^{L_a-1}$;
- $\hat{r}_j^{ss}(k) = \sum_{t=0}^{L_s-1-k} \hat{s}_j(t)\hat{s}_j(t+k)$;
- $\hat{r}_{j,fn}^{\psi\psi}(k) = \sum_{t=0}^{L_s-1-k} \psi_{j,fn}(t)\psi_{j,fn}(t+k)$;
- $\epsilon_{ij}(t) = x_i(t) - \sum_{j' \neq j} \hat{y}_{ij'}(t)$;
- $\hat{r}_{ij}^{s\epsilon}(k) = \sum_{t=0}^{L_s-1-k} \hat{s}_j(t)\epsilon_{ij}(t+k)$.

We can show that the gradient is given as follows:

$$\Delta \hat{\mathbf{a}}_{ij} = \mathbf{\Lambda}_{\hat{\mathbf{a}}_{ij}} \hat{\mathbf{a}}_{ij} - \frac{1}{\sigma_i^2} \hat{\mathbf{r}}_{ij}^{s\epsilon}, \quad (40)$$

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

$$\mathbf{\Lambda}_{\hat{\mathbf{a}}_{ij}} = \text{diag} \left(\left\{ \frac{\nu_u}{d_{ij}(t)r^2(t)} \right\}_t \right) + \frac{1}{\sigma_i^2} T_{L_a} \left\{ \hat{r}_j^{ss}(k) + \sum_{(f,n) \in \mathcal{B}_j} \gamma_{j,fn} \hat{r}_{j,fn}^{\psi\psi}(k) \right\}. \quad (41)$$

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 $\hat{r}_{j,fn}^{\psi\psi}(0) = 1$, 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 [20]). Alternatively, we can show that canceling the derivative of the VFE with respect to this parameter leads to the following update:

$$\sigma_i^2 = \frac{1}{T} \bar{e}_i, \quad (42)$$

where \bar{e}_i is defined in the Appendix in (50).

2) *NMF Parameters*: We recall that $\lambda_{j,fn}^2 = [\mathbf{W}_j \mathbf{H}_j]_{fn}$. We can show that maximizing $\mathcal{L}(q^*, \boldsymbol{\theta})$ 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:

$$\mathcal{C}(\mathbf{W}_j, \mathbf{H}_j) = \sum_{(f,n) \in \mathcal{B}_j} d_{IS}(\hat{p}_{j,fn}, [\mathbf{W}_j \mathbf{H}_j]_{fn}), \quad (43)$$

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

$$\hat{p}_{j,fn} = \frac{\hat{s}_{j,fn}^2 + \gamma_{j,fn}}{\beta_{j,fn}/\nu_v}. \quad (44)$$

This minimization problem can be solved by using the standard multiplicative update (MU) rules given in [14]. It has been shown in [54] that this procedure leads to a monotonic decrease of the IS divergence.

It is also interesting to note that by using (31) and (32) we can rewrite $\hat{p}_{j,fn}$ as follows:

$$\hat{p}_{j,fn} = \left(\frac{\alpha_v (\hat{s}_{j,fn}^2 + \gamma_{j,fn})^{-1} + [\mathbf{W}_j \mathbf{H}_j]_{fn}^{-1}}{\alpha_v + 1} \right)^{-1}. \quad (45)$$

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 NMF parametrization. Injecting (45) in (43) we obtain an optimization problem very similar to the one presented in the recent Student's t NMF framework [18]. 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 [14].

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) *Gaussian Source Model and Deterministic Time-Domain Mixing Filters*: The first baseline method corresponds to our previous work [31] that has already been described in Section I-B. In the following, the results using this method will be denoted by "Gaussian - unconstrained filters" as it is based on a Gaussian source model and deterministic unconstrained mixing filters in the time domain.

2) *Spatial Covariance Matrix Model*: The second baseline method was proposed in [55]. It generalizes in a same framework the punctual and non-punctual convolutive mixing models used in [20], [56] and [24], [30] respectively. This approach is based on the following Gaussian multichannel source image model:

$$\mathbf{y}_{j,fn} \sim \mathcal{N}(0, \lambda_{j,fn}^2 \mathbf{R}_{j,f}), \quad \text{with } \mathbf{R}_{j,f} = \mathbf{A}_{j,f} \mathbf{A}_{j,f}^H, \quad (46)$$

where $\mathbf{y}_{j,fn} = [y_{1j,f}, \dots, y_{Ij,f}]^T \in \mathbb{C}^I$, $\lambda_{j,fn}^2 \in \mathbb{R}_+$ is the short-term PSD of the source parametrized with an NMF model as in (11) and $\mathbf{R}_{j,f} \in \mathbb{C}^{I \times I}$ is the spatial covariance matrix (SCM), further structured using a matrix $\mathbf{A}_{j,f} \in \mathbb{C}^{I \times R_j}$ of rank $0 < R_j \leq I$. 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). Source separation is performed by means of an expectation-maximization (EM) algorithm. We use here the implementation provided by the authors and available at [57]. In the following the results using this framework when the SCM is of rank 1 and 2 will be denoted by "SCM rank-1" and "SCM rank-2" respectively.

The (V)EM 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 [58]. We created 8 stereo mixtures sampled at 16 kHz using RIRs from the MIRD database [42]. 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. We expect this approach along with the proposed priors to efficiently separate highly reverberant mixtures. 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 (15). All other parameters are blindly estimated. This semi-blind scenario will be used for the proposed method and the baselines.

3) *Performance Measures*: We evaluate the quality of the separation in terms of reconstructed stereo source images. We use standard energy ratios defined in [59] 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 [60]. We also consider the Overall Perceptual Score (OPS) which is a perceptually motivated objective measure [61], [62]. It is expressed in percentage and computed using the PEASS toolbox [63].

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 (31) and (34) 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 (25)-(28) 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 [31] 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 initialize $\mathbf{A}_{j,f}$ in (46) as a rank-1 matrix (column vector of size I) whose i -th entry corresponds to the discrete Fourier transform (DFT) of $\hat{a}_{ij}(t)$. The number of points for computing the DFT is equal to the length of the STFT window.

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 [18] 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 (43) under the same constraint with:

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

For the baseline methods where the source model is 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 rules⁴ given in [14]. 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.

⁴In 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.

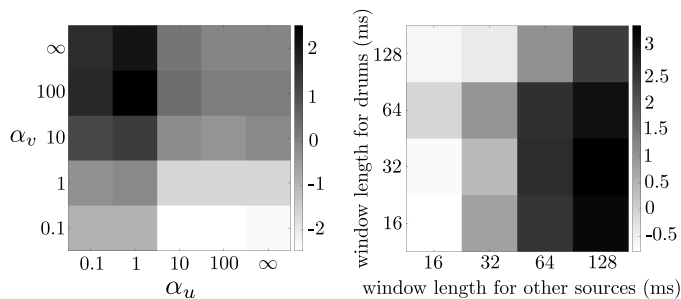


Fig. 5. Average SDR according to the shape parameters α_v and α_u (left plot) and according to the window length for the drums and other sources (right plot). The reverberation time of the mixtures is equal to 360 ms.

D. Model Hyperparameters and Algorithm Parameters

The specific choices for model hyperparameters and algorithm parameters are described in this subsection. The experiments have been performed using the mixtures associated with a reverberation time of 360 ms.

1) *Noise Variance*: As explained in Section III-D1, instead of using (42) we decrease the noise variance σ_i^2 over the iterations from 10^{-2} to 10^{-6} by using the same schedule as in [20].

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 (15) 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 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

TABLE I
AVERAGE SOURCE SEPARATION RESULTS.

Reverberation time: 160 ms					
	SDR	ISR	SIR	SAR	OPS
SCM rank-1 [55]	2.5	6.4	3.6	9.8	18.8
SCM rank-2 [55]	2.1	5.9	3.0	9.5	17.7
Gaussian - unconstrained filters [31]	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
SCM rank-1 [55]	1.9	5.7	2.4	9.4	16.6
SCM rank-2 [55]	1.5	5.3	1.9	9.4	15.7
Gaussian - unconstrained filters [31]	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
SCM rank-1 [55]	1.8	5.4	2.4	9.4	14.6
SCM rank-2 [55]	1.5	5.1	1.7	9.1	14.0
Gaussian - unconstrained filters [31]	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

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 [64]. However this is left for future work as this paper mostly focuses on proposing a new mixing model for reverberant mixtures.

E. Comparison with the Baselines

We compare below 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. Audio examples illustrating those results are available at [65]. 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.

As can be seen from Table I, the proposed method is very robust to highly reverberant conditions. On the contrary, the performance of the baseline methods decreases when the reverberation time increases. This shows that the SCM model (of rank 1 or 2) is not sufficiently accurate for modeling highly reverberant mixtures, confirming the results given in [28] and [29]. We also observe that the rank-1 SCM model outperforms the rank-2 model. This may be due to the fact that the latter involves more parameters than the former. However we can mention that it is difficult to hear any difference between the two approaches from the audio examples available online.

Our previous approach [31] 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 the methods based on the SCM model. 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 illustrated by audio examples available at [65].

Finally we also notice that even though the model hyperparameters (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.

For the sake of reproducibility, a Matlab implementation of the proposed algorithm is available at [65].

V. CONCLUSION

This paper introduced a new Bayesian framework for multichannel audio source separation based on the Student's t distribution. 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 [66]. This idea will be investigated in future works, in the context of the proposed source separation framework.

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 [67]. One could also use more sophisticated NMF-based source models such as the ones in [55] (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 (12). Such an approach has been followed for the Gaussian SCM model in [68], [69].

APPENDIX

The variational free energy can be decomposed from its definition in (21) as follows:

$$\begin{aligned} \mathcal{L}(q^*, \theta) &= \left\langle \ln \left(\frac{p(\mathbf{x}, \mathbf{s}, \mathbf{v}, \mathbf{a}, \mathbf{u}; \theta)}{q^*(\mathbf{s})q^*(\mathbf{v})q^*(\mathbf{a})q^*(\mathbf{u})} \right) \right\rangle_{q^*(\mathbf{s})q^*(\mathbf{v})q^*(\mathbf{a})q^*(\mathbf{u})} \\ &= \langle \ln p(\mathbf{x}|\mathbf{s}, \mathbf{a}; \sigma) \rangle_{q^*(\mathbf{s})q^*(\mathbf{a})} \\ &\quad + \langle \ln p(\mathbf{s}|\mathbf{v}; \lambda) - \ln q^*(\mathbf{s}) \rangle_{q^*(\mathbf{s})q^*(\mathbf{v})} \\ &\quad + \langle \ln p(\mathbf{v}) - \ln q^*(\mathbf{v}) \rangle_{q^*(\mathbf{v})} \\ &\quad + \langle \ln p(\mathbf{a}|\mathbf{u}) - \ln q^*(\mathbf{a}) \rangle_{q^*(\mathbf{a})q^*(\mathbf{u})} \\ &\quad + \langle \ln p(\mathbf{u}) - \ln q^*(\mathbf{u}) \rangle_{q^*(\mathbf{u})}. \end{aligned} \quad (48)$$

Note that $\langle -\ln q(\cdot) \rangle$ is the entropy of the distribution q . Let us detail each term in the right hand side of (48) from the model presented in Section II and the variational distributions given by equations (25) to (28). 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:

$$\langle \ln p(\mathbf{x}|\mathbf{s}, \mathbf{a}; \sigma) \rangle = -\frac{IT}{2} \ln(2\pi) - \frac{T}{2} \sum_{i=1}^I \ln(\sigma_i^2) - \frac{1}{2} \sum_{i=1}^I \frac{1}{\sigma_i^2} \bar{e}_i, \quad (49)$$

where \bar{e}_i is given by:

$$\begin{aligned} \bar{e}_i &= \left\| \mathbf{x}_i - \sum_{j=1}^J \hat{\mathbf{y}}_{ij} \right\|_2^2 + \sum_{j=1}^J \left[\|\hat{\mathbf{s}}_j\|_2^2 \sum_{\tau=0}^{L_a-1} \rho_{ij}(\tau) \right] \\ &\quad + \sum_{j=1}^J \sum_{(f,n) \in \mathcal{B}_j} \gamma_{j,fn} \left[\|\hat{\mathbf{g}}_{j,fn}\|_2^2 + \sum_{\tau=0}^{L_a-1} \rho_{ij}(\tau) \right], \end{aligned} \quad (50)$$

with $\hat{\mathbf{g}}_{ij,fn} = [\hat{g}_{ij,fn}(0), \dots, \hat{g}_{ij,fn}(T-1)]^\top$, $\hat{\mathbf{s}}_j = [\hat{s}_j(0), \dots, \hat{s}_j(L_s-1)]^\top$ and $\hat{\mathbf{y}}_{ij} = [\hat{y}_{ij}(0), \dots, \hat{y}_{ij}(T-1)]^\top$. Note that $\hat{s}_j(t)$ and $\hat{y}_{ij}(t)$ are related to the variational parameters $\{\hat{s}_{j,fn}\}_{j,fn}$ and $\{\hat{a}_{ij}(t)\}_{i,j,t}$ by (23) and (24) respectively.

2) *S-term*:

$$\begin{aligned} \langle \ln p(\mathbf{s}|\mathbf{v}; \boldsymbol{\lambda}) - \ln q^*(\mathbf{s}) \rangle = & -\frac{1}{2} \sum_{j=1}^J \sum_{(f,n) \in \mathcal{B}_j} \left[\ln \left(\frac{\beta_{j,fn}}{\gamma_{j,fn}} \right) \right. \\ & \left. - \text{di}\Gamma(\nu_v) + \ln(\lambda_{j,fn}^2) + \frac{\nu_v}{\beta_{j,fn}} \frac{\hat{s}_{j,fn}^2 + \gamma_{j,fn}}{\lambda_{j,fn}^2} - 1 \right]. \end{aligned} \quad (51)$$

3) *V-term*:

$$\begin{aligned} \langle \ln p(\mathbf{v}) - \ln q^*(\mathbf{v}) \rangle = & -\sum_{j=1}^J \sum_{(f,n) \in \mathcal{B}_j} \left[\frac{\alpha_v \nu_v}{2} \frac{1}{\beta_{j,fn}} \right. \\ & + \frac{\alpha_v}{2} \ln(\beta_{j,fn}) + \ln \Gamma \left(\frac{\alpha_v}{2} \right) + \frac{\alpha_v}{2} \ln \left(\frac{2}{\alpha_v} \right) \\ & \left. + \text{di}\Gamma(\nu_v) \left(\nu_v - \frac{\alpha_v}{2} \right) - \nu_v - \ln \Gamma(\nu_v) \right]. \end{aligned} \quad (52)$$

4) *A-term*:

$$\begin{aligned} \langle \ln p(\mathbf{a}|\mathbf{u}) - \ln q^*(\mathbf{a}) \rangle = & -\frac{1}{2} \sum_{i=1}^I \sum_{j=1}^J \sum_{t=0}^{L_a-1} \left[\ln \left(\frac{d_{ij}(t)}{\rho_{ij}(t)} \right) \right. \\ & \left. - \text{di}\Gamma(\nu_u) + \ln(r^2(t)) + \frac{\nu_u}{d_{ij}(t)} \frac{\hat{a}_{ij}^2(t) + \rho_{ij}(t)}{r^2(t)} - 1 \right]. \end{aligned} \quad (53)$$

5) *U-term*:

$$\begin{aligned} \langle \ln p(\mathbf{u}) - \ln q^*(\mathbf{u}) \rangle = & -\sum_{i=1}^I \sum_{j=1}^J \sum_{t=0}^{L_a-1} \left[\frac{\alpha_u \nu_u}{2} \frac{1}{d_{ij}(t)} \right. \\ & + \frac{\alpha_u}{2} \ln(d_{ij}(t)) + \ln \Gamma \left(\frac{\alpha_u}{2} \right) + \frac{\alpha_u}{2} \ln \left(\frac{2}{\alpha_u} \right) \\ & \left. + \text{di}\Gamma(\nu_u) \left(\nu_u - \frac{\alpha_u}{2} \right) - \nu_u - \ln \Gamma(\nu_u) \right]. \end{aligned} \quad (54)$$

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] S. Rickard, "The duet blind source separation algorithm," in *Blind Speech Separation*, S. Makino, H. Sawada, and T.-W. Lee, Eds. Springer, 2007, pp. 217–241.
- [8] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1592–1604, 2007.
- [9] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 382–394, 2010.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2015, pp. 1–5.
- [17] U. Şimşekli, A. Liutkus, and A. T. Cemgil, "Alpha-stable matrix factorization," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2289–2293, 2015.
- [18] 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.
- [19] P. Magron, R. Badeau, and A. Liutkus, "Lévy NMF for robust nonnegative source separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, United States, 2017.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] D. A. Bies and C. H. Hansen, *Engineering noise control: theory and practice*. CRC press, 2009.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.

- [29] —, “An EM algorithm for audio source separation based on the convolutive transfer function,” in *Proc. IEEE Workshop Applcat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, United States, Oct. 2017.
- [30] S. Arberet and P. Vanderghenst, “Reverberant audio source separation via sparse and low-rank modeling,” *IEEE Signal Process. Lett.*, vol. 21, no. 4, pp. 404–408, 2014.
- [31] 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.
- [32] —, “Semi-blind Student’s t source separation for multichannel audio convolutive mixtures,” in *Proc. European Signal Processing Conference (EUSIPCO)*, Kos, Greece, 2017.
- [33] —, “Separating time-frequency sources from time-domain convolutive mixtures using non-negative matrix factorization,” in *Proc. IEEE Workshop Applcat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, United States, 2017.
- [34] 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.
- [35] R. Giri, “Bayesian sparse signal recovery using scale mixtures with applications to speech,” Ph.D. dissertation, UC San Diego, 2016.
- [36] 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.
- [37] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] T. Schultz, “Diffusion in reverberation rooms,” *Journal of Sound and Vibration*, vol. 16, no. 1, pp. 17–28, 1971.
- [44] 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.
- [45] —, “Frequency-correlation functions of frequency responses in rooms,” *Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1819–1823, 1962.
- [46] J. A. Moorer, “About this reverberation business,” *Computer music journal*, vol. 3, no. 2, pp. 13–28, 1979.
- [47] J.-D. Polack, “La transmission de l’énergie sonore dans les salles,” Ph.D. dissertation, Université du Maine, 1988.
- [48] 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.
- [49] 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.
- [50] J. Winn and C. M. Bishop, “Variational message passing,” *Journal of Machine Learning Research*, vol. 6, no. Apr., pp. 661–694, 2005.
- [51] 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.
- [52] G. H. Golub and C. F. Van Loan, *Matrix computations*. Johns Hopkins University Press, 1996.
- [53] 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.
- [54] C. Févotte and J. Idier, “Algorithms for nonnegative matrix factorization with the β -divergence,” *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] M. Vinyes, “MTG MASS dataset,” <http://mtg.upf.edu/download/datasets/mass>, 2008.
- [59] 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.
- [60] E. Vincent, “BSS Eval Toolbox Version 3.0 for Matlab,” http://bass-db.gforge.inria.fr/bss_eval/, 2007.
- [61] 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.
- [62] 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.
- [63] V. Emiya and E. Vincent, “PEASS Toolbox Version 2.0 for Matlab,” <http://bass-db.gforge.inria.fr/peass/>, 2011.
- [64] 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.
- [65] <https://perso.telecom-paristech.fr/leglaive/demoSSMM.html>.
- [66] D. Wipf and H. Zhang, “Revisiting Bayesian blind deconvolution,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3595–3634, 2014.
- [67] 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.
- [68] 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.
- [69] —, “Multichannel music separation with deep neural networks,” in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Budapest, Hungary, 2016.