Multichannel Audio Source Separation with Probabilistic Reverberation Priors

Simon Leglaive, Roland Badeau, Gael Richard

Multichannel Audio Source Separation with Probabilistic Reverberation Priors

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

***Abstract*—Incorporating prior knowledge about the sources and/or the mixture is a way to improve under-determined audio source separation performance. A great number of informed source separation techniques concentrate on taking priors on the sources into account, but fewer works have focused on constraining the mixing model. In this paper we address the problem of under-determined multichannel audio source separation in reverberant conditions. We target a semi-informed scenario where some room parameters are known. Two probabilistic priors on the frequency response of the mixing filters are proposed. Early reverberation is characterized by an autoregressive model while according to statistical room acoustics results, late reverberation is represented by an autoregressive moving average model. Both reverberation models are defined in the frequency domain. They aim to transcribe the temporal characteristics of the mixing filters into frequency-domain correlations. Our approach leads to a maximum a posteriori estimation of the mixing filters which is achieved thanks to an expectation-maximization algorithm. We experimentally show the superiority of this approach compared with a maximum likelihood estimation of the mixing filters.**

***Index Terms*—Multichannel audio source separation, probabilistic priors, mixing model, MAP estimation, EM algorithm.**

I. Introduction

AUDIO source separation is the task that aims to recover a set of source signals from the observation of one or several mixtures. Audio source separation can be used as a pre-processing step for classification or recognition tasks. Indeed, more relevant features can be computed on the separated sources or higher-level information can be extracted from parametric models used for source separation. In music information retrieval, audio source separation can for example help improving the performance of tasks such as instrument recognition [1], main melody extraction [2], [3] or singing voice detection [4]. Audio source separation is more challenging when the separated sources are meant to be listened to by humans, such as for interference removal in close-microphone live recordings, remixing or upmixing applications.

When a punctual source is emitting in an enclosed space, the signal recorded by a microphone will include multiple reflections of the source signal on the surfaces and objects in the room. The acoustic channel between a source and a microphone is characterized by a Room Impulse Response (RIR) and represents how the sound propagates in the room. In the context of audio source separation the mixture is referred to as convolutive due to the filters involved in the mixing process. Each RIR between a microphone and a source is called a mixing filter. Non-punctual source models have also been proposed using spatial covariance matrices in [5], [6].

Source separation is commonly achieved in a Time-Frequency (TF) domain. Firstly because using a short-term representation can be useful to describe the spectro-temporal evolution of the sources. Secondly because it is possible to exploit the sparsity of the sources in this domain. And thirdly because expressing the mixture in this domain is easier: it can be defined according to a frequency-dependent mixing matrix built from the frequency response of the mixing filters.

Recent and now widely used approaches for tackling under-determined audio source separation are based on *variance modeling* frameworks [5]. In these probabilistic approaches, the Short-Term Fourier Transform (STFT) coefficients of each source are modeled as latent random variables following a complex circularly symmetric distribution with a time and frequency-dependent scale parameter. Within this framework, Non-negative Matrix Factorization (NMF) techniques are popular to represent the spectro-temporal characteristics of the sources [7]–[9]. The interpretability of a non-negative decomposition makes it possible to incorporate deterministic constraints or probabilistic priors on the sources [10], [11]. One can for example consider specific spectral structures as a source/filter model [2]. User input can be taken into account in the form of humming [12]. Even the musical score can be incorporated within the decomposition [13], [14]. Deep neural networks have also been used recently in the context of these variance modeling frameworks [15].

Compared with the large number of constraints that have been proposed about the sources, only a few methods consider priors on the spatial mixing parameters (*i.e.* the mixing filters or the spatial covariance matrices). In [16] the authors consider a complex Wishart prior on the inverse of the spatial covariance matrices. Geometrically calculated or pre-measured steering vectors representing the direct path between the source and the microphones are used to parametrize the prior. In [17] the authors consider Inverse-Wishart and Gaussian priors over the spatial covariance matrices. The prior aims to represent that, on average over all possible source and microphone positions in the room, the spatial covariance matrix is equal to the sum of two terms: the outer-product of the steering vector and a term modeling the spatial correlations between the microphones due to late reverberation. The method requires the knowledge of the source positions and certain room characteristics.

In this article we propose two probabilistic priors over the

S. Leglaive, R. Badeau and G. Richard are with LTCI, CNRS, 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).

frequency response of the mixing filters. The first one corresponds to an autoregressive model of early reverberation. It is based on our previous works [18], [19] in which we neglected the influence of late reverberation. This is not the case in this paper as we consider a second prior for late reverberation. This prior aims to transcribe the time-domain exponential decay of the late reverberation power into frequency-domain correlations. Based on our late reverberation model in the frequency domain [20], we use an Autoregressive Moving-Average (ARMA) representation to characterize the late part of the mixing filters. This prior requires the knowledge of the reverberation time, the volume of the room and the total wall area. The priors are taken into account within an Expectation-Maximization (EM) algorithm so as to perform a Maximum A Posteriori (MAP) estimation of the mixing filters. We show that the source separation performance is improved compared to the standard Maximum Likelihood (ML) estimation.

Section II introduces the general source separation framework on which our work is built. It is mainly based on the previous methods [21], [22] where the mixing filters are estimated in the ML sense. In section III we introduce the reverberation models that will be used in section IV to define the priors on the mixing filters. The models are shown to be consistent with real data. Source separation experiments are detailed in section V where we compare the performance for ML and MAP estimations of the mixing filters. We finally draw conclusions in section VI.

II. Source separation framework

We consider a mixture of J source signals $s_j(t) \in \mathbb{R}$ on I channels denoted by $\mathbf{x}(t) = [x_1(t), ..., x_I(t)]^T \in \mathbb{R}^I$ where $(\cdot)^T$ is the transposition operator. The mixture can be decomposed according to the following additive model:

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{y}_j(t) + \mathbf{b}(t), \tag{1}$$

where $\mathbf{y}_j(t) = [y_{1j}(t), ..., y_{Ij}(t)]^T \in \mathbb{R}^I$ is the j-th source image, *i.e.* the vector of size I containing the image of the source signal $s_j(t)$ at each microphone i, and $\mathbf{b}(t) = [b_1(t), ..., b_I(t)]^T \in \mathbb{R}^I$ is a noise vector. Introducing this noise term in (1) is formally equivalent to considering an extra source image. However as we will see later, the noise model is usually different from the source image model; the noise rarely corresponds to a sound source propagating in an environment. It rather corresponds to a sensor noise or a modeling error. Moreover this noise term can be necessary for the separation algorithm, in order to prevent from potential numerical instabilities or slow convergence.

A. Mixing Model

Using the STFT, we can write the mixture for all $(f, n) \in \{0, ..., F-1\} \times \{0, ..., N-1\}$ as:

$$\mathbf{x}_{fn} = \sum_{j=1}^{J} \mathbf{y}_{j,fn} + \mathbf{b}_{fn}, \tag{2}$$

where $\mathbf{x}_{fn} = [x_{1,fn}, ..., x_{I,fn}]^T$, $\mathbf{y}_{j,fn} = [y_{1j,fn}, ..., y_{Ij,fn}]^T$ and $\mathbf{b}_{fn} = [b_{1,fn}, ..., b_{I,fn}]^T$ are the complex-valued vectors containing the STFT coefficients[1] of $x_i(t)$, $y_{ij}(t)$ and $b_i(t)$ respectively, at TF point (f, n). Each source image can be characterized by the associated source signal and a set of mixing filters. A widely used approximation in source separation consists in considering that the mixing filters are short compared with the length of the STFT analysis window. Under this hypothesis the convolutive mixing can be approximated by an instantaneous mixing in each frequency band [23]. A source image thus writes:

$$\mathbf{y}_{j,fn} = \mathbf{a}_{j,f} s_{j,fn}, \tag{3}$$

where $\mathbf{a}_{j,f} = [a_{1j,f}, ..., a_{Ij,f}]^T \in \mathbb{C}^I$ contains the frequency response of the mixing filters between source j and the I microphones and $s_{j,fn} \in \mathbb{C}$ represents the STFT of source j. This short mixing filters assumption is experimentally discussed in section V-C. From (2) and (3) the mixture can be written in matrix form:

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_{fn}, \tag{4}$$

where $\mathbf{A}_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I \times J}$ is referred to as the mixing matrix and $\mathbf{s}_{fn} = [s_{1,fn}, ..., s_{J,fn}]^T \in \mathbb{C}^J$ contains the coefficients of the J source STFTs. Compared with the mixing equation (2) which only involves the source images, the convolutive model (4) highlights the way the source signals are mixed, through the introduction of the mixing matrix. This formalism is thus well suited for representing the physics of the mixing process such as the propagation of sound in a room. However it reaches its limits when the sources cannot be considered as punctual or when the mixing process involves long reverberations for example. In these cases, other models less oriented on the representation of the mixing physical phenomena can be used. One possible approach consists in directly modeling the source images using spatial covariance matrices [5], [6]. As we focus in this paper on incorporating physically motivated priors on the mixing filters, we will only consider the convolutive model (4).

B. Local Gaussian Source Model

1) Source Distribution: The local Gaussian model is widely used in audio source separation [24], [21], [6], [10], [25]. Each coefficient $s_{j,fn}$ is modeled as a random variable following a centered proper complex Gaussian distribution:

$$s_{j,fn} \sim \mathcal{N}_c(0, v_{j,fn}). \tag{5}$$

$\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate proper complex Gaussian distribution with Probability Density Function (PDF):

$$N_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\det(\pi \boldsymbol{\Sigma})} \exp[-(\mathbf{x} - \boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})], \tag{6}$$

where $\boldsymbol{\Sigma}$ is the positive definite covariance matrix, $(\cdot)^H$ is the Hermitian transposition operator and $\det(\cdot)$ the matrix determinant. The term *proper* means that the pseudo-covariance matrix $\mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]$ is zero where $\mathbb{E}[\cdot]$ is

[1] Due to the Hermitian symmetry property that holds for real-valued signals, the STFT is a redundant TF transform. Therefore, the set of frequency bins $\{0, ..., F-1\}$ only includes the positive frequencies.

the mathematical expectation. Moreover if the mean vector $\boldsymbol{\mu}$ is also zero, the distribution is circularly symmetric. The sources are assumed mutually independent, we can thus write the distribution of the source vector introduced in (4) as:

$$\mathbf{s}_{fn} \sim \mathcal{N}_c(0, \boldsymbol{\Sigma}_{\mathbf{s},fn}), \qquad \boldsymbol{\Sigma}_{\mathbf{s},fn} = \mathrm{diag}\left([v_{j,fn}]_j\right), \tag{7}$$

where $\mathrm{diag}([c_m]_m)$ is the diagonal matrix constructed from the coefficients c_m for $m = 1, ..., M$.

2) NMF Source Power Parametrization: The variance $v_{j,fn}$ in the source model (5) represents the short-term Power Spectral Density (PSD) of source j; it is the PSD of each frame of the short-term analysis of the source signal. The popularity of the local Gaussian model comes in particular from the possibility of introducing a parametric modeling of the source PSD. The importance of such a parametrization lies in the underdetermined nature of the source separation problem. Indeed we have to estimate the FN parameters $\{v_{j,fn}\}_{fn}$ for each source. The idea is thus to represent the PSD of each source by a set of parameters whose cardinality is lower than FN. In this work we chose the widely used rank-reduction technique based on NMF. The idea of representing the short-term PSD of audio sources by a non-negative decomposition can be traced back to [26]. The Gaussian generative model based on NMF was then introduced in [7] and extended to multichannel audio source separation in [21]. It is based on the following factorization of the matrix $\mathbf{V}_j = [v_{j,fn}]_{fn} \in \mathbb{R}_+^{F \times N}$:

$$\mathbf{V}_j = \mathbf{W}_j \mathbf{H}_j, \tag{8}$$

where $\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$ and $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times N}$. K_j is the factorization rank and is generally chosen such that $K_j(F+N) \ll FN$. $\mathbf{W}_j$ is a matrix containing spectral templates while $\mathbf{H}_j$ contains the activation of these templates over the time frames. The non-negativity constraint of this decomposition generally leads to physically meaningful spectral templates. For example they can correspond to the spectrum of the music notes in the source signal. The rows of $\mathbf{H}_j$ then represent the activations of the notes over time frames.

3) Local Stationarity: One important hypothesis that will be used in the following consists in assuming the independence of the source STFT frames and the local stationarity of the source signals (*i.e.* over the support of each frame) [27]. For some category of signals, particularly Gaussian and generally harmonisable α-stable [28], this *local stationarity* assumption implies the independence of the TF points of the STFT such that:

$$p(\{\mathbf{s}_{fn}\}_{f,n}) = \prod_{f,n} p(\mathbf{s}_{fn}). \tag{9}$$

This widely used assumption is introduced for convenience because it strongly simplifies the models. However it is clearly not consistent with an STFT representation of the signals, in particular due to the framing and overlapping procedure.

C. Noise Model

As in [21], we assume a time-stationary and isotropic noise following a centered proper complex Gaussian distribution:

$$\mathbf{b}_{fn} \sim \mathcal{N}_c(0, \boldsymbol{\Sigma}_{\mathbf{b},f}), \qquad \boldsymbol{\Sigma}_{\mathbf{b},f} = \sigma_{b,f}^2 \mathbf{I}_I, \tag{10}$$

where $\sigma_{b,f}^2 > 0$ and $\mathbf{I}_m$ is the identity matrix of size m.

D. Parameter Estimation

Let $\mathbf{X} = \{\mathbf{x}_{fn}\}_{f,n}$ be the set of observed data and $\eta = \left\{\{\mathbf{W}_j\}_j, \{\mathbf{H}_j\}_j, \{\mathbf{A}_f\}_f, \{\sigma_{b,f}^2\}_f\right\}$ the set of parameters. In the most general case the parameters are estimated in a MAP sense:

$$\begin{aligned} \eta^\star &= \arg\max_{\eta} \ \ln p(\eta|\mathbf{X}) \\ &= \arg\max_{\eta} \ \ln p(\mathbf{X}|\eta) + \ln p(\eta), \end{aligned} \tag{11}$$

where $\ln p(\mathbf{X}|\eta)$ is the log-likelihood and $\ln p(\eta)$ the log-prior over the parameters. This estimation can be done by maximizing a lower bound of the criterion (11), typically with an EM algorithm [29]. Let $\mathbf{S} = \{\mathbf{s}_{fn}\}_{f,n}$ denote the set of latent or hidden variables. At the E-step of the EM algorithm we compute the following lower bound, from the current estimation η' of the parameters:

$$Q_{\mathrm{MAP}}(\eta|\eta') = Q_{\mathrm{ML}}(\eta|\eta') + \ln p(\eta), \tag{12}$$

where $Q_{\mathrm{ML}}(\eta|\eta')$ is defined as the conditional expectation of the complete-data log-likelihood. From (4), (5) and (10) it writes:

$$\begin{aligned} Q_{\mathrm{ML}}(\eta|\eta') =& \mathbb{E}_{\mathbf{S}|\mathbf{X},\eta'}[\ln p(\mathbf{X}, \mathbf{S}|\eta)] \\ \stackrel{c}{=}& -N \sum_{f=0}^{F-1} \Big[\ln\det(\boldsymbol{\Sigma}_{\mathbf{b},f}) \\ &+ \mathrm{Trace}\Big(\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \hat{\mathbf{R}}_{\mathbf{xx},f} - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \mathbf{A}_f \hat{\mathbf{R}}_{\mathbf{xs},f}^H \\ &- \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \hat{\mathbf{R}}_{\mathbf{xs},f} \mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \mathbf{A}_f \hat{\mathbf{R}}_{\mathbf{ss},f} \mathbf{A}_f^H \Big)\Big] \\ &- \sum_{j=1}^{J} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \Big[\ln(v_{j,fn}) + \frac{\hat{p}_{j,fn}}{v_{j,fn}} \Big], \end{aligned} \tag{13}$$

where $\mathbb{E}_{\mathbf{S}|\mathbf{X},\eta'}[\cdot]$ is the conditional expectation of $\mathbf{S}$ given $\mathbf{X}$, $\stackrel{c}{=}$ denotes equality up to a constant (independent of η in (13)) and the statistics denoted by letters with a hat are defined at the E-step in Algorithm 1. If no prior is considered on the parameters the second term in the right-hand side of (11) and (12) disappears. The estimation is thus done in an ML sense. We see that for both MAP and ML estimations the E-step can be reduced to the computation of the conditional expectation in (13). In the M-step we then maximize (13) (for ML estimation) or (12) (for MAP estimation) in order to obtain the new estimation of the parameters. These two steps are iterated until convergence. An alternative consists in only increasing and not maximizing the Q-function at the M-step. In that case the algorithm is referred to as a Generalized EM (GEM). In this work we will only consider priors on the mixing filters, *i.e.* on the set $\{\mathbf{A}_f\}_f$, so that only the update of the mixing matrix will be modified in the M-step compared with the ML case. We summarize in Algorithm 1 one iteration of the GEM algorithm. For more details about the derivation see [21], [22]. For ML estimation the update of the mixing filters at line 11 of Algorithm 1 is given by:

$$\mathbf{A}_f = \hat{\mathbf{R}}_{\mathbf{xs},f} \hat{\mathbf{R}}_{\mathbf{ss},f}^{-1}. \tag{14}$$

Finally, up to an additive constant independent of $v_{j,fn}$, we can recognize in the last line of (13) the Itakura-Saito

Algorithm 1: One iteration of the GEM algorithm

E-step:

1: $\mathbf{\Sigma}_{\mathbf{s},fn} = \text{diag}([v_{j,fn}]_j)$ with $v_{j,fn} = [\mathbf{W}_j\mathbf{H}_j]_{fn}$
2: $\mathbf{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},f}$
3: $\mathbf{G}_{\mathbf{s},fn} = \mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}$
4: $\hat{\mathbf{s}}_{fn} = \mathbf{G}_{\mathbf{s},fn}\mathbf{x}_{fn}$
5: $\mathbf{\Sigma}_{\mathbf{s},fn}^{post} = (\mathbf{I}_J - \mathbf{G}_{\mathbf{s},fn}\mathbf{A}_f)\mathbf{\Sigma}_{\mathbf{s},fn}$
6: $\hat{\mathbf{R}}_{\mathbf{ss},fn} = \hat{\mathbf{s}}_{fn}\hat{\mathbf{s}}_{fn}^H + \mathbf{\Sigma}_{\mathbf{s},fn}^{post}$
7: $\hat{p}_{j,fn} = [\hat{\mathbf{R}}_{\mathbf{ss},fn}]_{j,j}$
8: $\hat{\mathbf{R}}_{\mathbf{ss},f} = \frac{1}{N}\sum_n \hat{\mathbf{R}}_{\mathbf{ss},fn}$
9: $\hat{\mathbf{R}}_{\mathbf{xx},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{x}_{fn}^H$
10: $\hat{\mathbf{R}}_{\mathbf{xs},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\hat{\mathbf{s}}_{fn}^H$

M-step:

11: Update $\mathbf{A}_f$ according to (14) (ML estimation) or Algorithm 2 (MAP estimation)
12: $\mathbf{\Sigma}_{\mathbf{b},f} = \text{Trace}(\hat{\mathbf{R}}_{\mathbf{xx},f} - \mathbf{A}_f\hat{\mathbf{R}}_{\mathbf{xs},f}^H - \hat{\mathbf{R}}_{\mathbf{xs},f}\mathbf{A}_f^H + \mathbf{A}_f\hat{\mathbf{R}}_{\mathbf{ss},f}\mathbf{A}_f^H)\mathbf{I}_I/I$
13: $\mathbf{W}_j, \mathbf{H}_j = \text{IS-NMF}(\hat{\mathbf{P}}_j)$ with $\hat{\mathbf{P}}_j = [\hat{p}_{j,fn}]_{fn}$

(IS) divergence [7] between $v_{j,fn} = [\mathbf{W}_j\mathbf{H}_j]_{fn}$ and the posterior mean of the source power spectrograms $\hat{p}_{j,fn} = \mathbb{E}_{\mathbf{S}|\mathbf{X},\eta'}[|s_{j,fn}|^2]$. Therefore the update of the source parameters at line 13 of Algorithm 1 is done by computing an NMF on $\hat{\mathbf{P}}_j = [\hat{p}_{j,fn}]_{fn} \in \mathbb{R}_+^{F\times N}$ using the IS divergence. It can be done with the standard multiplicative update rules (see [7]).

E. Source Reconstruction

The sources are estimated in the Minimum Mean Square Error (MMSE) sense:

$$\hat{\mathbf{s}}_{fn} = \mathbb{E}_{\mathbf{s}_{fn}|\mathbf{x}_{fn}}[\mathbf{s}_{fn}]. \tag{15}$$

We know that $\mathbf{s}_{fn}$ and $\mathbf{b}_{fn}$ are two independent proper complex Gaussian random vectors. From the linearity of the normal law and (4) we can write the posterior distribution of the sources as $\mathbf{s}_{fn}|\mathbf{x}_{fn} \sim \mathcal{N}_c(\hat{\mathbf{s}}_{fn}, \mathbf{\Sigma}_{\mathbf{s},fn}^{post})$ where $\hat{\mathbf{s}}_{fn}$ and $\mathbf{\Sigma}_{\mathbf{s},fn}^{post}$ are defined at lines 4 and 5 of Algorithm 1 respectively.

The sources and mixing filters are however estimated up to a frequency-dependent scale factor. This ambiguity can be fixed with a normalization strategy as in [21]. Nevertheless we prefer to provide as the output of the source separation system the reconstructed source images:

$$\hat{\mathbf{y}}_{j,fn} = \hat{\mathbf{a}}_{j,f}\hat{s}_{j,fn}, \tag{16}$$

where $\hat{s}_{j,fn} = [\hat{\mathbf{s}}_{fn}]_j$ and $\hat{\mathbf{a}}_{j,f}$ is the j-th column of the estimated mixing matrix $\hat{\mathbf{A}}_f$. This strategy implicitly solves the scale ambiguity. The time-domain signals are then reconstructed by inverse STFT.

III. Reverberation Modeling

The aim of this work is to propose a new way of estimating the mixing matrix $\mathbf{A}_f$. Based on physically motivated reverberation models, we define probabilistic priors over the frequency response of the mixing filters $a_{ij,f}$. We then derive a MAP estimation of the mixing matrix. Compared with the ML estimation given by (14), we take the dependencies between frequency points into account for estimating the mixing matrix. As explained in this section, these frequency correlations are induced by the specific temporal dynamics of the mixing filters, which actually are room responses. For the sake of clarity we first introduce the reverberation models by considering a single room response and by using specific notations not related to the previously introduced source separation problem. We then use these models in section IV to define priors on the frequency responses $a_{ij,f}$ of the mixing filters in the context of source separation.

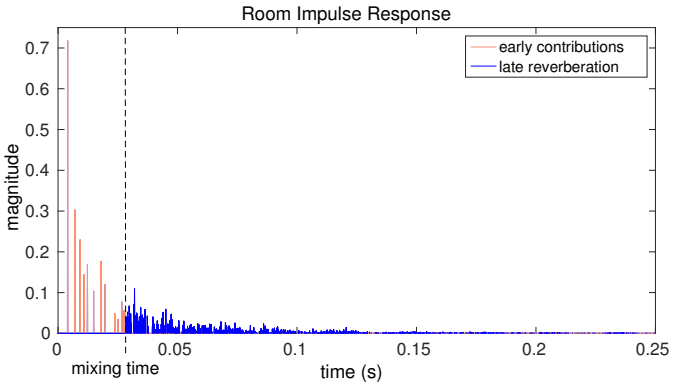


Fig. 1: Simulated room impulse response using the Roomsimove toolbox [31]. The room is a $10 \times 6.6 \times 3$ m shoebox. The reverberation time is 250 ms. The source to microphone distance is 1.38 m. The mixing time is defined in (17).

Let $h(t) = h_e(t) + h_l(t)$, $t = 0, ..., T-1$, denote the whole RIR and $h_e(t)$, $h_l(t)$ the early and late parts respectively, having disjoint temporal supports (see Fig. 1). The room frequency response (RFR) is similarly defined for $f = 0, ..., T-1$ by $H(f) = H_e(f) + H_l(f)$ where $H_{(\cdot)}(f)$ is the T-point Discrete Fourier Transform (DFT) of $h_{(\cdot)}(t)$. As the RIR is real-valued, the RFR satisfies the Hermitian symmetry, *i.e.* $H_{(\cdot)}(T-f) = H_{(\cdot)}(f)^*$ where $(\cdot)^*$ denotes the complex conjugate. We assume that T is even and we will refer to as *positive frequencies* the set of indices $\{0, ..., T/2\}$.

The time instant when late reverberation starts is referred to as the *mixing time*. It is usually defined according to the volume V of the room in m^3 [30]:

$$t_0 = \lfloor C_0\sqrt{V}f_s \rfloor \text{ samples}, \tag{17}$$

where $C_0 = 2 \times 10^{-3}$ is a normalization constant, f_s is the sampling rate in Hz and $\lfloor\cdot\rfloor$ the floor function. The time instant separating early and late reverberations in Fig. 1 is set according to this definition of the mixing time.

A. Early Reverberation

1) Model of Early Contributions: As represented in Fig. 1 the early part of the room response can be represented by a sum of *early contributions*. Let us consider R early contributions, each one is associated with an attenuation term ρ_k and a delay τ_k, $k = 0, ..., R-1$, such that $H_e(f) \approx G(f)$ with:

$$G(f) = \sum_{k=0}^{R-1} \rho_k\delta_k^f \quad \text{where} \quad \delta_k = e^{-j2\pi\tau_k/T}. \tag{18}$$

From (18) it follows that $\{G(f)\}_{f=R,...,T-1}$ satisfies a recursive equation of the form (see, *e.g.*, [32]):

$$\sum_{r=0}^{R} \varphi_r^e G(f-r) = 0, \tag{19}$$

such that $\{\varphi_r^e\}_{r=0}^R$ and $\{\delta_k\}_{k=0}^{R-1}$ are respectively the coefficients and roots of the same polynomial of order R. Without loss of generality we assume that φ_0^e is equal to one. We consider that $H_e(f)$ follows (19) up to a deviation $\kappa(f)$:

$$\sum_{r=0}^{R} \varphi_r^e H_e(f-r) = \kappa(f). \tag{20}$$

$\kappa(f)$ is modeled as a centered proper complex white Gaussian noise with variance σ_κ^2. From (20), $\{H_e(f)\}_f$ can be seen as following an autoregressive (AR) model of order R, denoted AR(R). We can finally write the joint PDF of the sequence of $H_e(f)$ for $f = 0, ..., T/2$:

$$\begin{aligned} p(\{H_e(f)\}_f) &= p(H_e(0), ..., H_e(R-1)) \\ &\prod_{f=R}^{T/2} N_c\Big(-\sum_{r=1}^{R} \varphi_r^e H_e(f-r), \sigma_\kappa^2\Big). \end{aligned} \tag{21}$$

2) Anechoic Case ($R = 1$): Theoretically the order R of the AR model (20) should be equal to the number of early contributions we consider. For example, for the RIR represented in Fig. 1 we should choose $R = 13$. However by considering that the direct path is dominating the early echoes, it is sufficient to choose $R = 1$. In this case, according to (18)-(20) we can write:

$$H_e(f) = \delta_0 H_e(f-1) + \kappa(f), \tag{22}$$

where $\delta_0 = e^{-j2\pi\tau_0/T}$ and $\tau_0 = \lfloor (r_0/c) f_s \rfloor$ with r_0 the distance between the source and the microphone in meters and c the speed of sound in m.s^{-1}. This AR(1) model also expresses that, as the temporal support of the early part of the RIR is concentrated on the time instants close to zero, the associated frequency response tends to be smooth. From equation (22) we can write:

$$|H_e(f)| \approx |H_e(f-1)|, \tag{23}$$

$$\arg(H_e(f)) \approx \arg(H_e(f-1)) - 2\pi\tau_0/T. \tag{24}$$

We represent in Fig. 2 the phase and magnitude relations between $H_e(f)$ and $H_e(f-1)$ according to (23) and (24). $H_e(f)$ corresponds to the early part of the RIR represented in Fig. 1 (red part), it contains 13 early contributions. Nonetheless we observe that considering an order $R = 1$ is a reasonable approximation.

B. Late Reverberation

Late reverberation corresponds to a stage of propagation where we usually consider the sound as diffuse. It means that the sound energy is uniformly distributed in the room and over all directions [33]. It corresponds to the part of the RIR where many reflections occur. Contrary to what we did for early reverberation, there are so many reflections that we cannot characterize each one individually. We conversely have to use statistical methods to describe late reverberation.

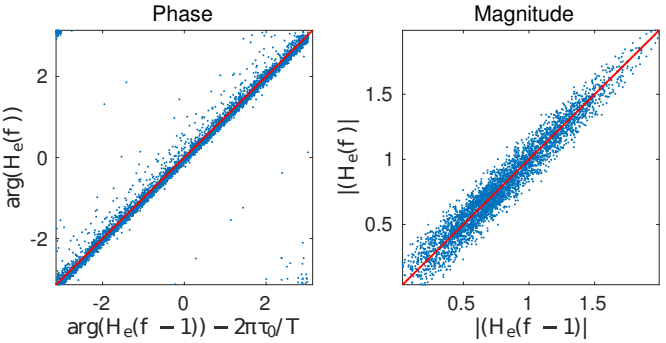


Fig. 2: Phase (left figure) and magnitude (right figure) relations between $H_e(f)$ and $H_e(f-1)$ illustrating (23) and (24). Blue dots: data from the early part of the RIR represented in Fig. 1, red lines: line $x = y$.

1) Statistical Characterization: It is well known that for a diffuse sound field the reverberation power decays exponentially [34]. This temporal dynamics induces specific frequency correlations between the coefficients of an RFR. Schroeder in [35] defined the frequency correlation functions of frequency responses in rooms. However he neglected the influence of the direct path and early echoes, resulting in a mismatch between theoretical and experimental results, as observed in [36]. We showed in [20] that considering the validity of the exponential decay only for late reverberation yields a more accurate match between theory and experiments.

We first characterize the temporal dynamics of the late RIR by defining the *Power Temporal Profile* (PTP) for $t = 0, ..., T-1$:

$$\bar{h}_l(t) = \mathbb{E}[|h_l(t)|^2] = P_0^2 e^{-2t/\tau} \mathbb{1}_{t\in\{t_0,...,T-1\}}(t), \tag{25}$$

where $\mathbb{1}_{t\in\mathcal{T}}(t)$ is the indicator function which equals 1 if $t \in \mathcal{T}$, 0 otherwise, P_0^2 is a constant related to the total power of late reverberation and τ is linked to the reverberation time[2] T_{60} through:

$$\tau = \frac{T_{60} f_s}{3\ln(10)}. \tag{26}$$

It is important to mention that different realizations of the room responses can be interpreted as different observations at several source and microphone positions in the room. According to the theory of statistical room acoustics, $\{H_l(f)\}_f$ is a proper centered and Wide Sense Stationary (WSS) complex Gaussian random process. We define the autocovariance function (ACVF) $\gamma(m)$ and the PSD $\phi(t)$ of this process by:

$$\gamma(m) = \mathbb{E}[H_l(f) H_l(f-m)^*]; \tag{27}$$

$$\phi(t) = \frac{1}{T}\mathbb{E}[|\mathcal{F}_T\{H_l(f)\}|^2], \tag{28}$$

where $\mathcal{F}_T\{\cdot\}$ is the T-point DFT. We have to mention that as we work in discrete time and frequency, all signals are T-periodic. Strictly speaking the RFR is thus a T-periodic WSS random process. Moreover $\phi(t)$ has to be understood as a discretized PSD function. One can refer to [37] for a review

[2]The reverberation time in seconds is defined as the time it takes for the sound energy to decrease by 60 dB after extinction of the source.

of some properties of periodic random processes. It can be shown that the PSD is related to the PTP by [20]:

$$\phi(t) = T\bar{h}_l(T-t). \tag{29}$$

From the Wiener-Khinchin theorem and (29) we obtain the theoretical ACVF:

$$\gamma(m) = \sigma^2_{rev} \frac{1-e^{2/\tau}}{1-e^{2(T-t_0+1)/\tau}} \frac{1-e^{(j2\pi m/T+2/\tau)(T-t_0+1)}}{1-e^{j2\pi m/T+2/\tau}}, \tag{30}$$

where σ^2_{rev} is defined by:

$$\sigma^2_{rev} = \mathbb{E}[|H_l(f)|^2] = C_{rev}\frac{1-\alpha}{\pi\alpha\mathcal{S}}, \tag{31}$$

with α the average absorption coefficient (without dimension) and $\mathcal{S}$ the total wall area in m^2. $C_{rev} = 75$ is an empirical constant that has been shown to be consistent with real data in [20]. The average absorption coefficient can be computed using Norris-Eyring's formula [38, p. 24]:

$$\alpha = 1 - e^{-24\ln(10)V/(c\mathcal{S}T_{60})}. \tag{32}$$

2) ARMA Model: As proposed in [20] we consider an invertible and causal ARMA(P,Q) model for representing the late part of the RFR $\{H_l(f)\}_f$:

$$\Phi(L)H_l(f) = \Theta(L)\epsilon(f), \tag{33}$$

where $\Phi(L) = \sum_{p=0}^{P}\varphi_p^l L^p$, $\Theta(L) = \sum_{q=0}^{Q}\theta_q L^q$ with $\varphi_0^l = \theta_0 = 1$ and L is the lag operator, *i.e.* $LH_l(f) = H_l(f-1)$. $\epsilon(f)$ is a centered proper complex white Gaussian noise of variance σ^2_ϵ for $f \in [0,...,T-1]$ and is extended by T-periodicity elsewhere. The important point here is that the ARMA parameters can be estimated from the sole knowledge of the theoretical ACVF given by (30), without the need of any data. Finally from this ARMA model we can write the following relationship involving the joint PDF of the sequence of $H_l(f)$ for $f = 0,...,T/2$:

$$p(\{H_l(f)\}_f) \propto \prod_{f=0}^{T/2} N_c\left(\frac{\Phi(L)}{\Theta(L)}H_l(f); 0, \sigma^2_\epsilon\right). \tag{34}$$

It is not an equality because we have omitted a term related to the Jacobian of the inverse ARMA filter transformation which does not depend on $\{H_l(f)\}_f$.

3) Validation: To validate the late reverberation model we verify on data the consistency of the theoretical ACVF and its ARMA parametrization. We consider a $10\times6.6\times3$ m shoebox with reverberation time $T_{60} = 250$ ms. We simulate 196 RIRs from the image source method using the Roomsimove toolbox [31]. According to a uniform grid of points in the room, the relative source/microphone position varies between the 196 RIRs and each RIR corresponds to one realization of the same random process. From this set of realizations we conduct a Monte-Carlo simulation to obtain the empirical ACVF for the late part of the RFR. We also compute the theoretical ACVF (30) from the same room parameters as the ones used to simulate the RIRs. Finally we compute an ARMA$(7,2)$ parametrization from the theoretical ACVF (see [20] and [39, ch. 2] for the estimation procedure). We represent in Fig. 3 the three ACVFs: computed from the data, from expression (30), and the ARMA parametrization. We observe a good match between theory and experiments. The validity of the model is also verified on measured RIRs in [20].

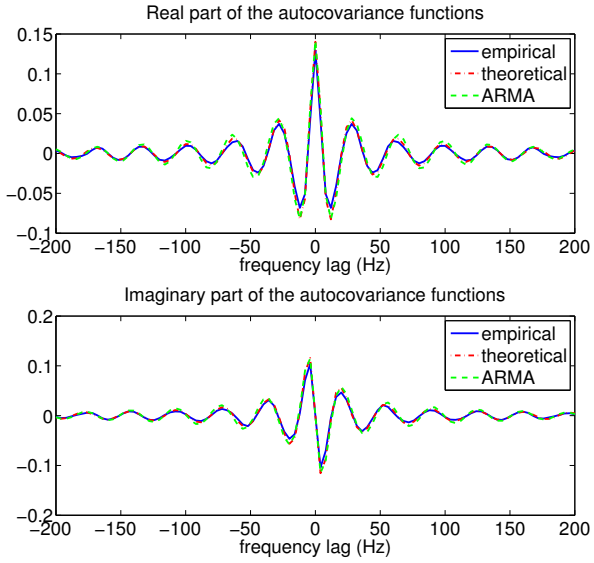


Fig. 3: Real (top figure) and imaginary (bottom figure) parts of the RFR ACVFs for late reverberation. Blue solid line: empirical ACVF, red dash-dot line: theoretical ACVF from (30), green dashed line: ARMA$(7,2)$ parametrization.

IV. Priors on the Mixing Filters

We now incorporate the reverberation models introduced in section III into the source separation framework presented in section II. We consider that the mixing matrix $\mathbf{A}_f$ can be decomposed as:

$$\mathbf{A}_f = \mathbf{A}_{e,f} + \mathbf{A}_{l,f}, \tag{35}$$

where $\mathbf{A}_{e,f} = [a^e_{ij,f}]_{ij} \in \mathbb{C}^{I\times J}$ and $\mathbf{A}_{l,f} = [a^l_{ij,f}]_{ij} \in \mathbb{C}^{I\times J}$ correspond respectively to the early and late parts of the mixing filters.

A. Early Reverberation Prior

1) Prior: We consider that, independently for all i,j, the filter $a^e_{ij,f}$ follows the AR(1) model (22). According to (21) for $R = 1$ and by considering no prior on $a^e_{ij,0}$ for all i,j we can write:

$$\ln p(\{\mathbf{A}_{e,f}\}_f) = -IJ(F-1)\ln(\pi\sigma^2_\kappa) - \frac{1}{\sigma^2_\kappa}\sum_{f=1}^{F-1}\left|\left|\mathbf{A}_{e,f} - \mathbf{\Delta}\circ\mathbf{A}_{e,f-1}\right|\right|^2_F, \tag{36}$$

where $\mathbf{\Delta} = [\delta_{ij}]_{ij} \in \mathbb{C}^{I\times J}$, $||\cdot||^2_F$ is the Frobenius norm and $\circ$ is the element-wise matrix product. $\delta_{ij} = e^{-j2\pi\tau_{ij}/L_a}$ with L_a the length of the mixing filters and $\tau_{ij} = \lfloor(r_{ij}/c)f_s\rfloor$ where r_{ij} is the distance between the j-th source and the i-th microphone in meters.

2) Hyperparameters: This prior is parametrized by the set of AR coefficients $\{\delta_{ij}\}_{i,j}$ and the variance σ_κ^2. The AR coefficients can be estimated within the GEM algorithm or fixed if we know the distance r_{ij} between each source j and microphone i. In a Bayesian context the variance σ_κ^2 is fixed and expresses how confident we are about the fact that $a^e_{ij,f}$ is close to $\delta_{ij}a^e_{ij,f-1}$.

B. Late Reverberation Prior

1) Prior: We consider that each filter $a^l_{ij,f}$ independently follows for all i,j the ARMA model (33). We can also write that the filter vector $\mathbf{a}^l_{j,f} = [a^l_{1j,f},...,a^l_{Ij,f}]^T \in \mathbb{C}^I$ follows:

$$\Phi(L)\mathbf{a}^l_{j,f} = \Theta(L)\boldsymbol{\epsilon}_f, \tag{37}$$

with $\boldsymbol{\epsilon}_f \sim \mathcal{N}_c(0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},f} = \sigma_\epsilon^2\mathbf{I}_I)$. As $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},f}$ is proportional to the identity matrix we do not consider any spatial correlation between the mixing filters. Finally, according to (34) we can write the following late reverberation prior:

$$\ln p(\{\mathbf{A}_{l,f}\}_f) \overset{c}{=} \sum_{f=0}^{F-1} -J\ln\det(\pi\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},f}) - \mathrm{Trace}\left[\left(\frac{\Phi(L)}{\Theta(L)}\mathbf{A}_{l,f}\right)^H \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\epsilon},f}\left(\frac{\Phi(L)}{\Theta(L)}\mathbf{A}_{l,f}\right)\right]. \tag{38}$$

2) Hyperparameters: This prior is parametrized by the ARMA coefficients $\{\varphi^l_p\}_p$, $\{\theta_q\}_q$ and the variance σ^2_ϵ. We will use an ARMA$(7,2)$ model. The ARMA coefficients are learned from the theoretical ACVF in (30). This function only depends on the reverberation time, the room volume and the total wall area through the exponential decay factor τ (26), the mixing time t_0 (17), the variance σ^2_{rev} (31) and the absorption coefficient α (32). These parameters are assumed to be known. As for the early reverberation prior, the variance σ^2_ϵ is fixed and expresses how confident we are about the prior.

C. MAP Estimation

As explained in section II-D, only the M-step of the GEM algorithm (summarized in Algorithm 1) is modified in order to take the priors on the mixing filters into account. The MAP estimation of the set of mixing filters $\mathbf{A}_e = \{\mathbf{A}_{e,f}\}_f$ and $\mathbf{A}_l = \{\mathbf{A}_{l,f}\}_f$ at the M-step is done by minimizing:

$$\mathcal{L}(\mathbf{A}_e, \mathbf{A}_l) = -Q_{\mathrm{ML}}(\eta|\eta') - \ln p(\{\mathbf{A}_{e,f}\}_f) - \ln p(\{\mathbf{A}_{l,f}\}_f), \tag{39}$$

where $Q_{\mathrm{ML}}(\eta|\eta')$ is defined in (13) with $\mathbf{A}_f = \mathbf{A}_{e,f} + \mathbf{A}_{l,f}$, $\ln p(\{\mathbf{A}^e_f\}_f)$ is defined in (36) and $\ln p(\{\mathbf{A}^l_f\}_f)$ in (38). The whole procedure is summarized in Algorithm 2 and is explained below. The final GEM algorithm for MAP estimation corresponds to Algorithm 1 where line 11 is replaced by Algorithm 2.

1) M-step for $\mathbf{A}_e$: Canceling the gradient of $\mathcal{L}(\mathbf{A}_e, \mathbf{A}_l)$ with respect to $\mathbf{A}_{e,f}$ leads to the updates summarized in Algorithm 2 Part 1. $\otimes$ denotes the Kroenecker product and $\mathrm{vec}(\cdot)$ concatenates the columns of a matrix into a single column vector. We see that at each frequency the update of $\mathbf{A}_{e,f}$ depends on $\mathbf{A}_{e,f-1}$ and/or $\mathbf{A}_{e,f+1}$. This coordinate descent should thus be repeated over the whole set of frequencies until a stopping criterion is reached (*e.g.* convergence or maximum number of iterations).

2) Update of the AR coefficients: The AR coefficients $\{\delta_{ij}\}_{i,j}$ that parametrize the early reverberation prior (36) can be either fixed if we know the relative distance between each source and microphone, or estimated at the M-step of the GEM algorithm, from the current estimation of the early mixing filters $\{a^e_{ij,f}\}$. We choose the second option because firstly, knowing the source to microphone distance can be difficult in practice, and secondly we did not observe a significant improvement over the source separation results by setting the AR coefficients to their true values. We thus estimate in the M-step the AR coefficients that minimize the cost function (39) under the constraint $|\delta_{ij}| = 1$ imposed by the model. This is equivalent to maximizing the prior (36) under the same constraint. It is important to note that this prior is here seen as a log-likelihood because we assume that the filters $\{a^e_{ij,f}\}$ are observed from their current estimation. The update of the AR coefficients is obtained using the method of Lagrange multipliers. It is given in Part 2 of Algorithm 2.

3) M-step for $\mathbf{A}_l$: Unfortunately we cannot easily cancel out the gradient of $\mathcal{L}(\mathbf{A}_e, \mathbf{A}_l)$ with respect to $\mathbf{A}_{l,f}$ due to the MA part of the ARMA model. We thus have to consider another descent method to minimize $\mathcal{L}(\mathbf{A}_e, \mathbf{A}_l)$ with respect to $\mathbf{A}_l$. We studied various approaches including a gradient descent with optimal step size, the Barzilai and Borwein approach [40] and a conjugate gradient method without or with preconditioning [41]. We obtained the best results in terms of convergence speed with the third approach (with preconditioning). We detail the method here.

Our objective is to minimize $\mathcal{L}(\mathbf{A}_e, \mathbf{A}_l)$ with respect to the vector of parameters of size IJF:

$$\mathbf{a}_l = \left[\mathrm{vec}\left(\mathbf{A}_{l,0}\right)^T, \mathrm{vec}\left(\mathbf{A}_{l,1}\right)^T, ..., \mathrm{vec}\left(\mathbf{A}_{l,F-1}\right)^T\right]^T. \tag{40}$$

We decompose the gradient vector $\mathbf{g}$ (column vector of size IJF) as the concatenation of the F gradients:

$$\mathbf{g} = [\mathbf{g}_0^T, \mathbf{g}_1^T, ..., \mathbf{g}_{F-1}^T]^T, \tag{41}$$

with $\mathbf{g}_f = \mathrm{vec}\left(\frac{1}{2}\nabla_{\mathbf{A}_{l,f}}\mathcal{L}(\mathbf{A}_e, \mathbf{A}_l)\right)$. We can show from (39) that:

$$\begin{aligned}\mathbf{g}_f =& N\,\mathrm{vec}\left(\boldsymbol{\Sigma}^{-1}_{\mathbf{b},f}(\mathbf{A}_{e,f}\hat{\mathbf{R}}_{\mathbf{ss},f} - \hat{\mathbf{R}}_{\mathbf{xs},f})\right) \\ &+ N\left(\hat{\mathbf{R}}^T_{\mathbf{ss},f} \otimes \boldsymbol{\Sigma}^{-1}_{\mathbf{b},f}\right)\mathrm{vec}\left(\mathbf{A}_{l,f}\right) \\ &+ \frac{\Phi^*(L^{-1})}{\Theta^*(L^{-1})}\left((\mathbf{I}_J \otimes \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\epsilon},f})\frac{\Phi(L)}{\Theta(L)}\mathrm{vec}\left(\mathbf{A}_{l,f}\right)\right),\end{aligned} \tag{42}$$

where $\Phi^*(L^{-1}) = \sum_{p=0}^P(\varphi^l_p)^*L^{-p}$ and $\Theta^*(L^{-1}) = \sum_{q=0}^Q\theta^*_qL^{-q}$. We want to solve $\mathbf{g} = 0$ which is here equivalent to solving a positive definite linear system. For that purpose we use the Preconditioned Conjugate Gradient (PCG) method [41]. From (42) and the fact that $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},f}$ is diagonal we can define the following preconditioning matrix:

$$\mathbf{D} = \begin{pmatrix}\mathbf{D}_0 & 0 & \cdots & 0 \\ 0 & \mathbf{D}_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{D}_{F-1}\end{pmatrix}, \tag{43}$$

Algorithm 2: MAP update of the mixing filters

Part 1: Update $\{\mathbf{A}_{e,f}\}_f$ by coordinate descent

while stopping criterion not reached **do**

$$\mathrm{vec}(\mathbf{A}_{e,0}) = \Big[N\hat{\mathbf{R}}^T_{\mathbf{ss},0} \otimes \mathbf{I}_I + \tfrac{1}{\sigma_\kappa^2}(\mathbf{I}_J \otimes \mathbf{\Sigma}_{\mathbf{b},0})\Big]^{-1} \times \mathrm{vec}\Big[N\hat{\mathbf{R}}_{\mathbf{xs},0} - N\mathbf{A}_{l,0}\hat{\mathbf{R}}_{\mathbf{ss},0} + \tfrac{1}{\sigma_\kappa^2}\mathbf{\Sigma}_{\mathbf{b},0}\big(\mathbf{\Delta}^* \circ \mathbf{A}_{e,1}\big)\Big]$$

for all $f \in \{2, ..., F-2\}$ **do**

$$\mathrm{vec}(\mathbf{A}_{e,f}) = \Big[N\hat{\mathbf{R}}^T_{\mathbf{ss},f} \otimes \mathbf{I}_I + \tfrac{2}{\sigma_\kappa^2}(\mathbf{I}_J \otimes \mathbf{\Sigma}_{\mathbf{b},f})\Big]^{-1} \times \mathrm{vec}\Big[N\hat{\mathbf{R}}_{\mathbf{xs},f} - N\mathbf{A}_{l,f}\hat{\mathbf{R}}_{\mathbf{ss},f} + \tfrac{1}{\sigma_\kappa^2}\mathbf{\Sigma}_{\mathbf{b},f}\big(\mathbf{\Delta} \circ \mathbf{A}_{e,f-1} + \mathbf{\Delta}^* \circ \mathbf{A}_{e,f+1}\big)\Big]$$

end for

$$\mathrm{vec}(\mathbf{A}_{e,F-1}) = \Big[N\hat{\mathbf{R}}^T_{\mathbf{ss},F-1} \otimes \mathbf{I}_I + \tfrac{1}{\sigma_\kappa^2}(\mathbf{I}_J \otimes \mathbf{\Sigma}_{\mathbf{b},F-1})\Big]^{-1} \times \mathrm{vec}\Big[N\hat{\mathbf{R}}_{\mathbf{xs},F-1} - N\mathbf{A}_{l,F-1}\hat{\mathbf{R}}_{\mathbf{ss},F-1} + \tfrac{1}{\sigma_\kappa^2}\mathbf{\Sigma}_{\mathbf{b},F-1}\big(\mathbf{\Delta} \circ \mathbf{A}_{e,F-2}\big)\Big]$$

end while

Part 2: Update $\{\delta_{ij}\}_{i,j}$

$$\delta_{ij} = \Big(\sum_{f=1}^{F-1} a^e_{ij,f}(a^e_{ij,f-1})^*\Big) \Big/ \Big|\sum_{f=1}^{F-1} a^e_{ij,f}(a^e_{ij,f-1})^*\Big|$$

Part 3: Update $\{\mathbf{A}_{l,f}\}_f$ with the PCG method

Initialize $\mathbf{g}$ from (41) and (42)

Initialize $\boldsymbol{\omega} = [\boldsymbol{\omega}_0^T, \boldsymbol{\omega}_1^T, ..., \boldsymbol{\omega}_{F-1}^T]^T$ with $\boldsymbol{\omega}_f = \mathbf{D}_f^{-1}\mathbf{g}_f$ and $\mathbf{D}_f$ given by (44)

while stopping criterion not reached **do**

$\boldsymbol{\gamma} = [\boldsymbol{\gamma}_0^T, \boldsymbol{\gamma}_1^T, ..., \boldsymbol{\gamma}_{F-1}^T]^T$ with $\boldsymbol{\gamma}_f = N(\hat{\mathbf{R}}^T_{\mathbf{ss},f} \otimes \mathbf{\Sigma}^{-1}_{\mathbf{b},f})\boldsymbol{\omega}_f + \frac{\Phi^*(L^{-1})}{\Theta^*(L^{-1})}(\mathbf{I}_J \otimes \mathbf{\Sigma}^{-1}_{\boldsymbol{\epsilon},f})\frac{\Phi(L)}{\Theta(L)}\boldsymbol{\omega}_f$

$\mu = (\boldsymbol{\omega}^H\mathbf{g})/(\boldsymbol{\omega}^H\boldsymbol{\gamma})$

$\mathbf{a}_l \leftarrow \mathbf{a}_l - \mu\boldsymbol{\omega}$

Compute $\mathbf{g}$ from (41) and (42)

$\mathbf{g}_p = \mathbf{D}^{-1}\mathbf{g}$ where $\mathbf{D}$ is given by (43) and (44)

$\alpha = -(\boldsymbol{\gamma}^H\mathbf{g}_p)/(\boldsymbol{\omega}^H\boldsymbol{\gamma})$

$\boldsymbol{\omega} \leftarrow \mathbf{g}_p + \alpha\boldsymbol{\omega}$

end while

Part 4: Update $\{\mathbf{A}_f\}_f$

$\mathbf{A}_f = \mathbf{A}_{e,f} + \mathbf{A}_{l,f}$

where the diagonal matrix $\mathbf{D}_f$ is given by:

$$\mathbf{D}_f = N\,\mathrm{diag}(\hat{\mathbf{R}}^T_{\mathbf{ss},f} \otimes \mathbf{\Sigma}^{-1}_{\mathbf{b},f}) + \frac{1}{\sigma_\epsilon^2}\mathbf{I}_{IJ}\sum_{s=0}^{N_\psi}|\psi_s|^2. \qquad (44)$$

The parameters $\{\psi_s\}_{s=0}^{N_\psi}$ are the coefficients of the polynomial $\Psi(L)$ which approximates the transfer function of the inverse ARMA model $\Phi(L)/\Theta(L)$. For the chosen ARMA$(7, 2)$ model, we obtain a very accurate approximation by setting $N_\psi = 2048$. We summarize the PCG method for the update of $\{\mathbf{A}_{l,f}\}_f$ in Part 3 of Algorithm 2.

V. Experiments

We present in this section the dataset and the evaluation criteria used for the experiments. We discuss the short mixing filters assumption used for writing the source image model (3). We also describe our initialization strategy before running the GEM algorithm. We explain how the algorithm parameters are chosen and we finally detail the source separation results. Audio examples are available from our demo web page[3].

mix	duration	src. prop.	s1	s2	s3	s4	s5
1	28 s	instr.	piano	brushes	bass	-	-
		DoA	-45°	0°	45°	-	-
2	14 s	instr.	drums	voice	piano	bass	-
		DoA	0°	-45°	45°	-20°	-
3	24 s	instr.	drums	guitar	bass	-	-
		DoA	0°	-45°	45°	-	-
4	28 s	instr.	drums	voice	guitar	guitar	-
		DoA	0°	-45°	45°	-20°	-
5	18 s	instr.	bass	guitar	drums	-	-
		DoA	45°	-45°	0°	-	-
6	15 s	instr.	drums	voice	guitar	bass	-
		DoA	0°	-45°	45°	-20°	-
7	25 s	instr.	drums	voice	guitar	guitar	bass
		DoA	0°	-45°	-20°	45°	20°
8	12 s	instr.	drums	voice	bass	-	-
		DoA	0°	-45°	45°	-	-

TABLE I: Description of the dataset: duration (in seconds) and source properties, *i.e.* instrument and DoA (in degrees), for each mixture.

A. Dataset

Our experiments were conducted from the audio tracks without effects provided by the Musical Audio Signal Separation (MASS) dataset [42]. We created 8 stereo mixtures by simulating mixing filters with the Roomsimove toolbox [31]. The room was a $4.45 \times 3.55 \times 2.5$ m shoebox with a reverberation time of 128 ms. The microphone spacing was set to 1 m, as the distance between the source and the center of the microphone pair. The direction of arrival (DoA) and the type of the sources are described in Table I for each mixture, along with the duration in seconds. As the MASS dataset provides stereo sources, each one is first converted to mono, downsampled to 16 kHz and filtered with the associated RIRs to create a source image. We finally sum all the source images to create a mixture.

B. Evaluation Criteria

We evaluate the source separation performance in terms of reconstructed source images. We use standard energy ratios: the Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ratio (SAR) and source Image-to-Spatial distortion Ratio (ISR). These criteria expressed in decibels (dB) are defined in [43]. We used the BSS Eval Toolbox available at [44] to compute these measures.

C. Short Mixing Filters Assumption

The aim of this section is to discuss the short mixing filters assumption used to express the convolutive mixture in the STFT domain. We assume that the mixing filters length is equal to the reverberation time. On the one hand, for rewriting the time-domain convolution as a multiplication in the STFT

[3]http://perso.telecom-paristech.fr/leglaive/demoMASSwPRP.html

measures	setting	average scores
SDR	$\mathbf{A}_f$ is blindly estimated	0.0
	$\mathbf{A}_{e,f}$ is oracle and $\mathbf{A}_{l,f}$ is zero	0.9
	$\mathbf{A}_f$ is oracle	**7.9**
ISR	$\mathbf{A}_f$ is blindly estimated	5.3
	$\mathbf{A}_{e,f}$ is oracle and $\mathbf{A}_{l,f}$ is zero	7.6
	$\mathbf{A}_f$ is oracle	**15.6**
SIR	$\mathbf{A}_f$ is blindly estimated	2.1
	$\mathbf{A}_{e,f}$ is oracle and $\mathbf{A}_{l,f}$ is zero	3.9
	$\mathbf{A}_f$ is oracle	**13.5**
SAR	$\mathbf{A}_f$ is blindly estimated	6.4
	$\mathbf{A}_{e,f}$ is oracle and $\mathbf{A}_{l,f}$ is zero	10.8
	$\mathbf{A}_f$ is oracle	**12.0**

TABLE II: Preliminary source separation results (in dB) averaged over the 29 sources of the dataset.

domain as in (3), we theoretically have to use an STFT analysis window longer than the reverberation time [23]. On the other hand, we cannot choose a too long analysis window because of the local stationarity assumption on the source signals. Moreover, the true mixing filters are still tractable in the limit case where the analysis window length equals the reverberation time. For example, considering the dataset presented above, we can choose a 128 ms long window and think of the mixing matrix $\mathbf{A}_f$ in (4) as containing the frequency response of the mixing filters. Even if this approach is theoretically arguable, it can be shown to be effective in terms of oracle source separation results[4]. This may explain why this mixing model is still widely used in audio source separation. Indeed, even if it is not accurate for representing the true mixing process (time-domain convolution), its simplicity is useful for blind audio source separation and it allows us to achieve satisfactory oracle performance.

To validate this argument, we conducted a preliminary source separation experiment with the baseline method summarized in Algorithm 1 (for the ML estimation case). The STFT analysis window length is equal to the reverberation time (128 ms). We consider three settings: (1) $\mathbf{A}_f$ is blindly estimated in the M-step according to equation (14); (2) the early part $\mathbf{A}_{e,f}$ is fixed to oracle values (using the true mixing filters truncated at the mixing time) and the late part $\mathbf{A}_{l,f}$ is fixed to zero; (3) the whole mixing matrix $\mathbf{A}_f = \mathbf{A}_{e,f} + \mathbf{A}_{l,f}$ is fixed to oracle values. All other parameters are blindly estimated. The results averaged on the 29 sources of the dataset are shown in Table II. We first observe that even if the length of the STFT analysis window equals the length of the mixing filters, we achieve good separation results when $\mathbf{A}_f$ is oracle. We also observe that when we neglect the late reverberation by setting $\mathbf{A}_{l,f}$ to zero, the separation performance drastically decreases compared with the case where the whole mixing matrix is oracle. This shows the importance of achieving a good estimation for both the early and late parts of the mixing filters.

The conclusion of this preliminary experiment is the following one: if we were able to estimate the true mixing filters, which correspond to room responses, we would achieve good source separation performance, even if the mixing filters length equals the STFT analysis window length. The aim of this work is precisely to incorporate constraints on the estimation of the mixing filters to go towards this objective.

[4]Moreover, one could wonder if the residual noise in the mixture model (4) does not become overwhelmingly large due to the short mixing filters assumption which is not satisfied. Experimental results which are not detailed here show that most of the mixture power is represented by the estimated source images and the residual noise is negligible.

D. Blind Initialization

The GEM algorithm is very sensitive to the parameter initialization. In order to obtain satisfactory separation results, we have to provide a "good initialization". As in [6], [19] we first tried the initialization procedure based on a hierarchical clustering algorithm to estimate the mixing filters [45] and a permutation solving method [46]. From the estimated mixing filters the sources can be recovered via projection of the mixture over the source directions and binary masking in the TF plan. The NMF parameters can then be estimated from the separated sources. However with this approach we obtained less satisfactory initialization results than with a more empirical method inspired from the one described in [21, section IV.H]. The initialization procedure we used is restricted to stereo mixtures and is detailed in the Appendix.

E. Algorithm Parameters

1) STFT parameters: We use half-overlapping sine windows. The STFTs are computed using the MATLAB code provided in the context of the SiSEC challenge at [47]. According to the discussion in section V-C, we fix the length of the analysis window to 128 ms or 2048 points at a sampling rate of 16 kHz.

2) NMF order: We arbitrarily choose an NMF of order 10 for all the sources.

3) Stopping criterion for the descent algorithms: It was sufficient in terms of convergence to repeat the coordinate descent for the update of $\{\mathbf{A}_{e,f}\}_f$ for 2 iterations. The stopping criterion for the PCG method used in the update of $\{\mathbf{A}_{l,f}\}_f$ is set to 20 iterations.

4) Prior hyperparameters: The AR coefficients for the early reverberation prior are estimated in the M-step of the GEM algorithm. The $\mathrm{ARMA}(7,2)$ coefficients for the late reverberation prior are learned from the theoretical ACVF (30) which is defined according to some known room parameters. The variances of both priors are critical hyperparameters because they have a strong influence on the separation results. Indeed, we see in Fig. 4 that the average SDR for the mixture 4 of the dataset varies from 0.2 to 4.2 dB according to the value of the variances. These variances control the strength of the priors, the smallest they are, the strongest the priors will be and conversely. In particular, if their values are too high the priors have no effect on the results compared with an ML estimation of the mixing filters; their estimation will indeed be dominated by the contribution from the log-likelihood.

F. Source Separation Results

For comparing ML and MAP estimations we run 500 iterations of the GEM algorithms from the same initialization. As can be seen in Fig. 4, the source separation results for

	SDR			ISR			SIR			SAR		
mix	ML	MAP_BEST	MAP_CV	ML	MAP_BEST	MAP_CV	ML	MAP_BEST	MAP_CV	ML	MAP_BEST	MAP_CV
1	-4.4	**-4.3**	-5.9	3.9	**4.1**	3.5	**0.4**	**0.4**	-3.4	**10.4**	9.8	7.3
2	-1.3	**0.0**	-1.0	4.1	4.0	**4.5**	-0.7	**1.2**	0.5	7.4	**8.5**	7.5
3	2.6	**3.8**	3.7	6.1	8.1	**8.4**	5.6	6.5	**6.7**	6.6	**9.4**	6.9
4	0.6	**4.2**	4.0	5.8	**8.4**	8.2	2.3	**6.8**	6.7	7.9	**9.6**	9.5
5	1.1	**2.0**	1.2	6.5	**7.6**	6.3	**2.0**	1.6	1.0	6.2	**6.3**	5.1
6	1.3	**2.5**	**2.5**	4.4	**5.9**	**5.9**	2.9	**5.7**	**5.7**	4.6	**6.1**	**6.1**
7	-0.7	**0.1**	0.0	5.0	**5.1**	**5.1**	0.5	**0.8**	0.7	3.4	**4.6**	4.3
8	**1.2**	**1.2**	-0.1	**7.5**	**7.5**	6.9	**5.2**	**5.2**	3.6	6.7	6.7	**6.8**
mean	0.0	**1.2**	0.7	5.3	**6.2**	6.0	2.1	**3.4**	2.7	6.4	**7.5**	6.6

TABLE III: Source separation results in dB: without prior (ML), with prior and best variances on the grid-search (MAP_BEST), with prior and variances chosen by cross-validation (MAP_CV). The mean is computed over the 29 sources of the dataset.

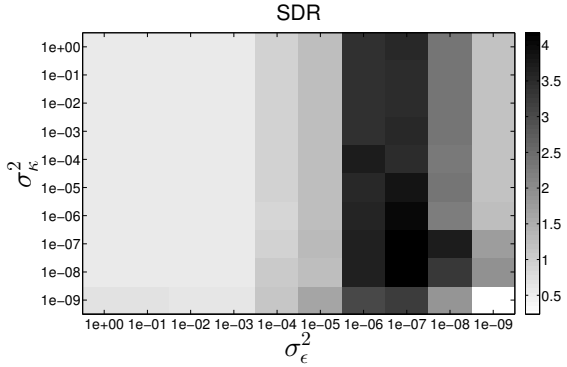


Fig. 4: Average SDR over a grid-search on the prior variances for mixture 4 of the dataset.

MAP estimation are highly dependent on the prior variances. Moreover the results we observe in Fig. 4 are considerably different if we take another mixture from our dataset. This justifies the use of a grid-search on the variances to assess the performance of MAP estimation.

1) Results with best variances on the grid: For all the variances in the grid $(\sigma^2_\kappa, \sigma^2_\epsilon) \in \{10^0, 10^{-1}, ..., 10^{-9}\} \times \{10^0, 10^{-1}, ..., 10^{-9}\}$ we first compute the source separation performance that leads to the best average SDR individually for each mixture. ML and MAP results with this strategy are shown on average for each mixture in Table III, in columns ML and MAP_BEST respectively. The mean results in the last line of Table III are computed by averaging over all the sources of the dataset. We see that over all the mixtures, the SDR improvement due to MAP estimation ranges from 0.1 to 3.6 dB. On average, MAP estimation leads to an improvement of 0.9 dB of ISR, 1.3 dB of SIR, 1.1 dB of SAR and 1.2 dB of SDR. We also represent in Fig. 5a a scatter plot of the SDR for the 29 sources of our dataset. On y and x axes we represent respectively the MAP and ML results. Red circles (resp. blue stars) above (resp. below) the line $x = y$ represent better separation results with MAP (resp. ML) estimation. We observe that globally the improvement due to MAP estimation is significant as it leads to an increase of the SDR on almost all the sources. When it is not the case, the decrease is small as blue stars are very close to the line $x = y$. Those results show that for each mixture there exist values for the variances that lead to an improvement of the source separation performance.

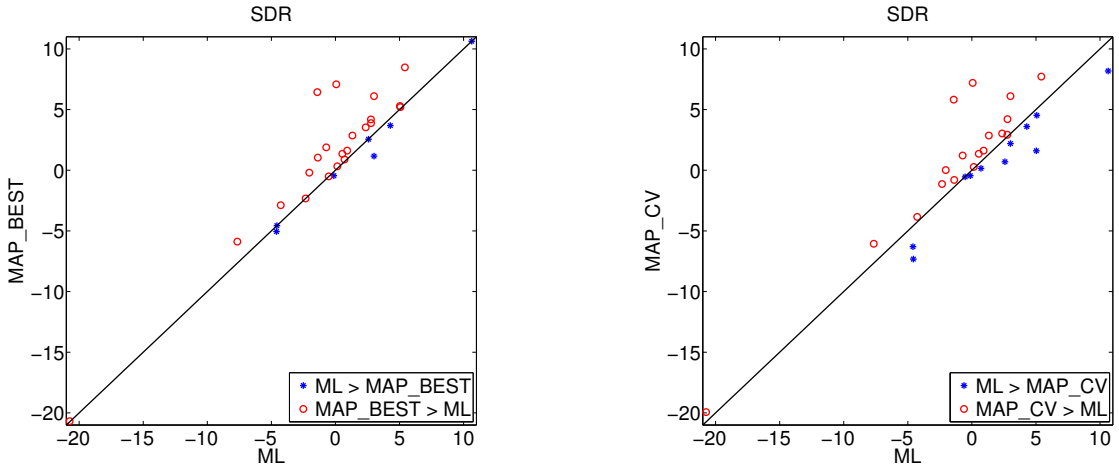
2) Results with variances chosen by cross-validation: We now present the results where the variances are chosen by cross-validation on the same grid as before. For each fold we consider 7 mixtures out of 8 as a training set and we keep the last one for testing. The variances are chosen as the ones that lead to the best average SDR over all the sources in the training set. Interestingly, for every fold the chosen variances are $(\sigma^2_\kappa, \sigma^2_\epsilon) = (10^{-6}, 10^{-7})$. The MAP results for this cross-validation are shown in column MAP_CV of Table III. Compared with ML estimation we observe a decrease of the SDR on mixtures 1 and 8 by 1.5 and 1.3 dB respectively. However for all the other mixtures, the SDR is increased by taking the priors into account. The improvement ranges from 0.1 to 3.4 dB. On average MAP estimation with cross-validation leads to an improvement of 0.7 dB of ISR, 0.6 dB of SIR, 0.2 dB of SAR and 0.7 dB of SDR. We observe from the scatter plot in Fig. 5b that with MAP estimation the quality of the separated sources never degrades much, but on some sources the improvement is important. Indeed, red circles tend to be further from line $x = y$ than blue stars.

VI. Conclusion

In this paper we introduced two probabilistic priors for constraining the estimation of the mixing filters in multichannel audio source separation. The priors aim to capture the temporal characteristics of the mixing filters by modeling the frequency correlations of their frequency responses. Experiments showed that these new priors lead to better source separation results than a standard approach with unconstrained mixing filters. However the results are highly dependent on the variances of the priors. In future works we could use a conjugate prior on the variances so that their values are constrained to be close to the ones we obtained with cross-validation, but not exactly equal.

This work targeted a semi-informed scenario where the reverberation time, the volume and the total wall area of the room were assumed to be known to define the late reverberation prior. In future works we could try to blindly estimate the reverberation time (see, *e.g*, [48]). Moreover it could be interesting to conduct a sensitivity analysis of the method with respect to errors on these room parameters.

Future works will also include the extension of our model to non-punctual sources using spatial covariance matrices. These matrices can be factorized as the outer product of *sub-sources* mixing matrices [10] on which we could consider the same priors as the ones presented in this paper.



(a) MAP_BEST results.

(b) MAP_CV results.

Fig. 5: Scatter plot of the SDR (in dB) for the 29 sources of the dataset. MAP and ML results are on y and x axis respectively. The source with a SDR around -20 dB corresponds to the brushes in the mixture 1 of the dataset. This source is not accurately modeled by an NMF.

Finally we could consider not only modeling the frequency correlations of the mixing filters but also the spatial correlations. As used in [6], [17] for stereo audio source separation, the spatial correlation functions are theoretically defined according to the distance between the microphones. For that purpose we could consider a full noise covariance matrix in equation (37) or use a vector ARMA model to represent late reverberation in the frequency domain.

APPENDIX
BLIND INITIALIZATION FOR STEREO MIXTURES

The initialization procedure we used is as follows:

1) Stack left and right mixture STFTs so as to create a $2F \times N$ complex-valued matrix.
2) Compute a K_{init}-component IS-NMF and split the resulting matrix containing the frequency atoms in two matrices associated to left and right channels. K_{init} is chosen between 3 and 5 depending on the mixture.
3) Reconstruct K_{init} left ($i = 1$) and right ($i = 2$) components $c_{ik,fn}$ by Wiener filtering (see, *e.g.*, [7]). By considering an anechoic model we can write $c_{ik,fn} = \rho_{ik} e^{-j2\pi f \tau_{ik}/L_a} |c_{k,fn}| e^{j \arg(c_{k,fn})}$.
4) Compute $\tilde{c}_{ik,fn} = c_{ik,fn}/e^{j \arg(c_{1k,fn})}$ such that we can write $\tilde{c}_{1k,fn} = \rho_{1k}|c_{k,fn}|$ and $\tilde{c}_{2k,fn} = \rho_{2k}|c_{k,fn}|e^{-j2\pi f(\tau_{2k}-\tau_{1k})/L_a}$.
5) Compute $\bar{c}_{ik,f} = \frac{1}{N}\sum_n \tilde{c}_{ik,fn}$.
6) Compute and unwrap $\xi_{k,f} = \arg(\frac{\bar{c}_{2k,f}}{\bar{c}_{1k,f}})$ such that we can write $\xi_{k,f} = -2\pi f(\tau_{2k} - \tau_{1k})/L_a$.
7) Assuming punctual and spatially disjoint sources, $\xi_{k,f}$ forms J clusters as multiple components are spatially associated with the same source. Use the K-means algorithm to cluster the $\xi_{k,f}$.
8) By denoting $\mathcal{K}_j$ the set of indices k associated to source j, initialize the mixing filters as $a_{ij,f} = \frac{1}{\#\mathcal{K}_j}\sum_{k\in\mathcal{K}_j} \bar{c}_{ik,f}$.
9) Compute a pre-separation by projection of the mixture over the source direction and TF masking. We used the MATLAB code provided in the context of the SiSEC challenge at [47].
10) Initialize the NMF parameters from the separated sources using the Kullback-Leibler divergence.
11) Initialize $\sigma^2_{b,f} = 10^{-2}\sum_{i,n}|x_{i,fn}|^2/(IN)$.

MAP estimation requires the initialization of $a^e_{ij,f}$ and $a^l_{ij,f}$ individually. For that purpose we simply split the initial mixing filters in the time domain according to the mixing time in (17).

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation." in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2009, pp. 327–332.

[2] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1180–1191, 2011.

[3] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, USA, 2010, pp. 425–428.

[4] S. Leglaive, R. Hennequin, and R. Badeau, "Singing voice detection with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, 2015, pp. 121–125.

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

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

[7] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[8] 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.

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

[10] 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, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[11] 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.

[12] P. Smaragdis and G. J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2009, pp. 69–72.

[13] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel, "Evaluation of a score-informed source separation system." in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2010, pp. 219–224.

[14] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, 2011, pp. 45–48.

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

[16] T. Otsuka, K. Ishiguro, T. Yoshioka, H. Sawada, and H. G. Okuno, "Multichannel sound source dereverberation and separation for arbitrary number of sources based on Bayesian nonparametrics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2218–2232, 2014.

[17] N. Q. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP J. Adv. Signal. Process.*, vol. 2013, p. 149, 2013.

[18] S. Leglaive, R. Badeau, and G. Richard, "A priori probabiliste anéchoïque pour la séparation sous-déterminée de sources sonores en milieu réverbérant," in *Colloque GRETSI*, Lyon, France, 2015, Paper no. 127.

[19] ——, "Multichannel audio source separation with probabilistic reverberation modeling," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2015, pp. 1–5.

[20] ——, "Autoregressive moving average modeling of late reverberation in the frequency domain," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Budapest, Hungary, 2016, pp. 1–5.

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

[22] 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.

[23] 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.

[24] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency gaussian source models," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2005, pp. 78–81.

[25] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1408–1423, 2016.

[26] 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.

[27] 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.

[28] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, 2015, pp. 266–270.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B (Methodological)*, pp. 1–38, 1977.

[30] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," in *Audio Engin. Soc. Conv. 128*, 2010, Paper no. 8089.

[31] E. Vincent and D. R. Campbell, "Roomsimove," http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip, 2008.

[32] R. Kumaresan, "On the zeros of the linear prediction-error filter for deterministic signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 31, no. 1, pp. 217–220, 1983.

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

[34] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Audio Engin. Soc. Conv. 103*, 1997, Paper no. 4629.

[35] M. R. Schroeder, "Frequency-correlation functions of frequency responses in rooms," *J. Acoust. Soc. Am.*, vol. 34, no. 12, pp. 1819–1823, 1962.

[36] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 791–803, 2003.

[37] G. Gu, X. Gao, J. He, and M. Naraghi-Pour, "Parametric modeling of wideband and ultrawideband channels in frequency domain," *IEEE Trans. Veh. Technol.*, vol. 56, no. 4, pp. 1600–1612, 2007.

[38] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Springer Science & Business Media, 2010.

[39] S. Kay, "Spectral estimation," in *Advanced Topics in Signal Processing*, J. S. Lim and A. V. Oppenheim, Eds. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988, pp. 58–122.

[40] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA J. Numer. Anal.*, vol. 8, no. 1, pp. 141–148, 1988.

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

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

[43] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Evaluation Campaign (2007-2010): Achievements and Remaining Challenges," *Signal Process.*, vol. 92, pp. 1928–1936, 2012.

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

[45] 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 J. Adv. Signal Process.*, vol. 2007, 2007.

[46] 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, Lang. Process.*, vol. 15, no. 5, pp. 1592–1604, 2007.

[47] "Signal Separation Evaluation Campaign (SiSEC 2008)," http://sisec2008.wiki.irisa.fr, 2008.

[48] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Am.*, vol. 114, no. 5, pp. 2877–2892, 2003.

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. Since October 2014, he has been working toward the Ph.D. degree in the Department of Signal and Image Processing, Télécom ParisTech. His research interests include statistical models for audio signals, audio source separation, and machine learning applied to 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, Cachan, France, in 2001, and the Ph.D. degree in the field of signal processing from the ENST in 2005. He received the Habilitation degree from the Université Pierre et Marie Curie (UPMC), Paris VI, Paris, France, in 2010. In 2001, he joined the Department of Signal and Image Processing of Télécom ParisTech, CNRS LTCI, as an Assistant Professor, where he became Associate Professor in 2005. He is currently a Chief Engineer of the French Corps of Mines (foremost of the great technical corps of the French state) and an Associate Editor of the *EURASIP Journal on Audio, Speech, and Music Processing*. He is a coauthor of more than 20 journal papers, more than 80 international conference papers, and 2 patents. His research interests include statistical modeling of nonstationary signals (including adaptive high-resolution spectral analysis and Bayesian extensions to NMF), with applications to audio and music (source separation, multipitch estimation, automatic music transcription, audio coding, audio inpainting). He received the ParisTech Ph.D. Award in 2006 from UPMC, Paris VI, France.

Gaël Richard (SM'06) received the State Engineering degree from Télécom ParisTech, France (formerly ENST) 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 the Department of Signal and Image Processing, Télécom ParisTech, where he is now a Full Professor in audio signal processing and Head of the Signal and Image processing department. He is a coauthor of over 200 papers and inventor in 8 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 senior member of the IEEE.