On the Use of Latent Mixing Filters in Audio Source Separation

Laurent Girin[1] and Roland Badeau[2]

[1] GIPSA-lab, Grenoble Alpes Univ., INRIA Rhône-Alpes, France
laurent.girin@gipsa-lab.grenoble-inp.fr
[2] LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, France
roland.badeau@telecom-paristech.fr

Abstract. In this paper, we consider the underdetermined convolutive audio source separation (UCASS) problem. In the STFT domain, we consider both source signals and mixing filters as latent random variables, and we propose to estimate each source image, i.e. each individual source-filter product, by its posterior mean. Although, this is a quite straightforward application of the Bayesian estimation theory, to our knowledge, there exist no similar study in the UCASS context. In this paper, we discuss the interest of this estimator in this context and compare it with the conventional Wiener filter in a semi-oracle configuration.[3]

Keywords: Audio source separation, source image, latent mixing filters, MMSE estimator, MCMC sampling.

1 Introduction

To address the difficult problem of underdetermined audio source separation, probabilistic methods working in the Short-Term Fourier Transform (STFT) domain have been developed, e.g. [1–4]. These methods combine a physical mixture model, including source-to-microphone channel, with a source prior model. The mixture is often considered as convolutive, while using a (complex-valued) local Gaussian model (LGM) for the sources is now very popular. The convolutive mixture is generally approximated in the STFT domain as an instantaneous mixture at each frequency [2], even though this approximation can be questioned when the impulse response of the mixing filters is longer than the STFT window. A more general channel model has been proposed in [3], and combined with source LGM: the covariance matrix of the source image[4] is modeled as the product of the source power spectral density (PSD) with a spatial covariance matrix (SCM). A full-rank SCM is claimed to appropriately model diffuse sources and

[3] This work is partly supported by the French National Research Agency (ANR) as a part of the EDISON 3D project (ANR-13-CORD-0008-02), and by the European Research Council (ERC) Advanced Grant VHIA 340113.

[4] A source image is defined as the multichannel version of the source signal, as recorded at the microphones [5].

to overcome to some extent the limitation of the convolutive-to-multiplicative approximation [3], whereas the SCM model reduces to the convolutive model when the SCM is rank-1.

In all these papers, the coefficients of the channel model are considered as parameters of the overall probabilistic model. The source signals are considered as latent variables. The inference of sources and the estimation of (source and channel) parameters are made using an Expectation-Maximization (EM) algorithm or a similar two-step iterative procedure. Considering the channel coefficients as random variables or random processes, hence additional latent variables, has been recently proposed in a few audio source separation studies [6–9]. In [6, 8], a prior distribution is assigned to the channel coefficients. This enables to introduce prior information about the mixture and acoustic environment in a principled manner, e.g. dependencies of the channel with source location or knowledge on room acoustics. In [7, 9] a time-varying channel is considered as a random process and is estimated using a hidden Markov model with states corresponding to source direction of arrival (DoA) [7] or using a more general Kalman smoother [9]. The whole model solution is obtained following the variational EM methodology, which relies on the approximation of the joint posterior distribution of hidden variables (for instance source and channel) into a factorized form [10].

In a general manner, in all the above-mentioned studies, the extraction of the source signals from the mixture signal is made by some kind of Wiener filtering, in the E-step of the EM. Wiener filters are built from the current value of source parameters and from the current value of channel coefficients, be these latter considered as parameters or random variables. In turn, the new source estimates are used to update the channel coefficients (in the M-step or in some other part of the E-step). Therefore, *channel estimation and source signal estimation are two separate sequential processes*. Yet, in a fully Bayesian approach, where source and channel coefficients are considered as random variables, the posterior distribution of the source, and the associated source MMSE estimator, take a more general form: a stochastic integral that is generally not tractable [1, 11]. Therefore, the (standard or variational) EM methodology can be seen as a way to break this intractability into an iterative sequential process that is suboptimal at each iteration but that is globally efficient.

In the present study, we consider the convolutive case, and we consider the mixing matrix in the STFT domain as a latent variable affected with a prior distribution. Instead of the sequential channel estimation and Wiener filtering inherent to the EM, we propose to *directly estimate the source image, i.e. the product of a monochannel source and the corresponding mixing vector, by its posterior expectation*, i.e. the "fully Bayesian" MMSE estimator applied to the source image. Hence, in contrast to the EM, the mixing filters are considered here as a latent random variable *during the source inference step*. This may sound quite trivial at first sight, but the inference of the product of two random variables is not easy. In particular, we assume that the posterior probability of the filter-source product does *not* factorize, as opposed to what is done in approximate variational methods [7, 9]. In Section 2, we discuss this important

point in more details and explain how the source image estimator contrasts with the conventional (convolutive) Wiener estimator. Actually, we put in evidence theoretical links between latent mixing vectors and the SCM model of [3].

Unfortunately, just like in the general case, the source image MMSE estimator takes the form of an intractable stochastic integral. Nevertheless, we further derive an advanced formulation of that stochastic integral in the case where the mixing filters follow a complex Gaussian distribution. The resulting expression depends on the source distribution, and though we use the LGM source model in the present study, the formulation is valid for any other distribution. We then turn to numerical approximation techniques to compute values of the source image estimator. We conduct experiments using a very basic sampling technique, for instance the Metropolis algorithm [12, chap. 3]. We validate this approach in a "semi-oracle" configuration, where the source and channel parameters are estimated "offline" from the individual source images. In the present study, we only implement and discuss the inference step (in this semi-oracle configuration). The design of a complete blind separation process based on the proposed inference scheme and most likely of iterative nature, is out of the scope of the present paper. This paper must be considered as a prospective paper that discusses the use of a direct source image inference scheme in the UCASS framework, and positions this approach w.r.t. Wiener filtering.

Note that although the principle of the direct source image estimator is simple in essence, we could not find any paper exploring this idea in the present UCASS framework and reporting associated experiments. Probably the need to resort to computationally heavy sampling schemes can explain it. For example, a sampling process was applied to source separation in [11], but this study only dealt with instantaneous mixtures with mixing parameters assumed to be known. The present study however echoes [13], in which a *joint system and signal Kalman filter* was proposed and applied to single-channel speech enhancement and speech dereverberation. Interestingly, this joint scheme was opposed to a *dual* scheme, with sequential system and signal estimation, which can be seen, according to the authors of [13], "as a sequential variant of the EM procedure." The distribution of the source-system product within the joint Kalman filter was sampled using the Unscented Transform. In short, [13] considered a unique speech signal and a dynamic system, and the present paper considers a source separation problem, with stationary filters that can be easily extended to non-stationary filters.

2 Latent mixing filters and estimation of source image

2.1 Principle

As in many source separation methods, the mixture signal is modeled as a convolutive noisy mixture of the source signals. Relying on the so-called narrow-band assumption, i.e. the impulse responses of the mixing filters are shorter than the time-frequency (TF) analysis window, the $I \times 1$ mixture signal is expressed in

the short-time Fourier transform (STFT) domain as:

$$\mathbf{x}_{f\ell} = \mathbf{A}_{f\ell}\mathbf{s}_{f\ell} + \mathbf{b}_{f\ell} = \sum_{j=1}^{J} \mathbf{a}_{j,f\ell} s_{j,f\ell} + \mathbf{b}_{f\ell}, \tag{1}$$

where $f \in [1, F]$ is the frequency bin index, $\ell \in [1, L]$ is the frame index, $\mathbf{s}_{f\ell} = [s_{1,f\ell}, \ldots, s_{J,f\ell}]^\top \in \mathbb{C}^J$ (where symbol $.^\top$ denotes the transpose operator) is the vector of source coefficients, considered as a latent variable, $\mathbf{A}_{f\ell} = [\mathbf{a}_{1,f\ell}, \ldots, \mathbf{a}_{J,f\ell}] \in \mathbb{C}^{I\times J}$ is the mixing matrix ($\mathbf{a}_{j,f\ell} \in \mathbb{C}^I$ is the mixing vector for source j), and $\mathbf{b}_{f\ell} = [b_{1,f\ell}, \ldots, b_{I,f\ell}]^\top \in \mathbb{C}^I$ is a residual noise.

In the present study, we consider the mixing filter matrix $\mathbf{A}_{f\ell}$ as a latent variable, as opposed to a parameter as done in most audio source separation studies.[5] Moreover, in contrast to the classical use of a Wiener filter, we propose to estimate a source image signal $\mathbf{y}_{j,f\ell} = \mathbf{a}_{j,f\ell} s_{j,f\ell}$ directly by its posterior expectation, i.e. the MMSE estimator:

$$\hat{\mathbf{y}}_{j,f\ell} = \mathbb{E}_{q(\mathcal{H})}[\mathbf{a}_{j,f\ell} s_{j,f\ell}] = \mathbb{E}_{q(\mathbf{a}_{j,f\ell}, s_{j,f\ell})}[\mathbf{a}_{j,f\ell} s_{j,f\ell}], \tag{2}$$

where $\mathbb{E}_q$ denotes the mathematical expectation w.r.t. the probability density function (PDF) q, $q(.)$ denotes the posterior probability of a variable, i.e. $q(.) = p(.|\mathbf{x})$, and $\mathcal{H}$ denotes the complete set of hidden variables, i.e. $\mathcal{H} = \{\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}\}_{f,\ell=1}^{F,L}$ $= \{\mathbf{a}_{j,f\ell}, s_{j,f\ell}\}_{f,\ell,j=1}^{F,L,J}$. Note that we assume for simplicity that all distributions factorize over f and ℓ. We also naturally assume that sources and filters are independent in the prior sense, i.e. $p(\mathbf{a}_{j,f\ell}, s_{j,f\ell}) = p(\mathbf{a}_{j,f\ell})p(s_{j,f\ell})$. However, and very importantly, we do *not* want here $q(\mathbf{a}_{j,f\ell}, s_{j,f\ell})$ to factorize over $\mathbf{a}_{j,f\ell}$ and $s_{j,f\ell}$, as opposed to what was done in the variational approximation approach, e.g. [7, 9]. This is for two reasons: i) In a general manner, a joint process is optimal compared to a combination of subprocesses. For instance, we want to take benefit from a possible posterior correlation between source and mixing filter. ii) We want the proposed inference process to account for a diffuse source, seen as the "sum" of (possibly many) punctual sources with identical PSD and filtered with slightly different filters. Here the expectation in (2) takes the role of such summation. In contrast, factorizing $q(\mathbf{a}_{j,f\ell}, s_{j,f\ell})$ over $\mathbf{a}_{j,f\ell}$ and $s_{j,f\ell}$ would lead to $\hat{\mathbf{y}}_{j,f\ell} = \mathbb{E}_{q(\mathbf{a}_{j,f\ell})}[\mathbf{a}_{j,f\ell}]\mathbb{E}_{q(s_{j,f\ell})}[s_{j,f\ell}] = \hat{\mathbf{a}}_{j,f\ell}\hat{s}_{j,f\ell}$, i.e. a "unique" filtered source estimate, loosing the ability to represent diffuse sources. Note that the EM/Wiener approach within the convolutive mixture model is also problematic

[5] Considering the filters as latent variables enables us to make them depend on the time frame ℓ at no additional cost, compared to frame-independent latent filters $\mathbf{A}_f$, given that both models have the same set of parameters. This also comes at a much lower cost than the parametric case. However this does not necessarily mean that we have "trajectories" of filters, as for the moving sources or moving sensors in [7, 9]. This simply allows the realization of the filters to be different for each frame, e.g. modeling slight movements of sources around their mean position. In the following, $\mathbf{a}_{j,f\ell}$ is assumed wide-sense stationary (WSS) along ℓ, hence its mean and covariance matrix do not depend on ℓ.

in this regard: a single mixing vector estimate $\hat{\mathbf{a}}_{j,f\ell}$ is used to build a single Wiener filter, whose ability to filter out diffuse sources is questionable.

Before entering into the technical derivation of (2), we now want to mention that considering the mixing filters as (WSS) latent variables also has a very interesting interpretation in terms of spatial properties of the sources from the prior distribution point of view. Indeed, let us define $\boldsymbol{\mu}_{\mathbf{a},j,f} = \mathbb{E}_{p(\mathbf{a}_{j,f\ell})}[\mathbf{a}_{j,f\ell}]$ the (prior) mean vector of $\mathbf{a}_{j,f\ell}$, and $\boldsymbol{\Sigma}_{\mathbf{a},j,f} = \mathbb{E}_{p(\mathbf{a}_{j,f\ell})}[(\mathbf{a}_{j,f\ell} - \boldsymbol{\mu}_{\mathbf{a},j,f})(\mathbf{a}_{j,f\ell} - \boldsymbol{\mu}_{\mathbf{a},j,f})^{\mathrm{H}}]$ its (prior) covariance matrix ($.^{\mathrm{H}}$ denotes the conjugate transpose operator). Then, assuming prior uncorrelation between source and filter, the prior covariance matrix of a source image is given by:

$$\mathbf{R}_{\mathbf{y},j,f\ell} = \mathbb{E}_{p(\mathcal{H})}[\mathbf{y}_{j,f\ell}\mathbf{y}_{j,f\ell}^{\mathrm{H}}] = \mathbb{E}_{p(s_{j,f\ell})}[|s_{j,f\ell}|^2]\,\mathbb{E}_{p(\mathbf{a}_{j,f\ell})}[\mathbf{a}_{j,f\ell}\mathbf{a}_{j,f\ell}^{\mathrm{H}}], \tag{3}$$

hence

$$\mathbf{R}_{\mathbf{y},j,f\ell} = v_{j,f\ell}\mathbf{R}_{\mathbf{a},j,f}, \tag{4}$$

where $v_{j,f\ell} = \mathbb{E}_{p(s_{j,f\ell})}[|s_{j,f\ell}|^2]$ is the PSD of source j at TF-bin (f,ℓ), and

$$\mathbf{R}_{\mathbf{a},j,f} = \mathbb{E}_{p(\mathbf{a}_{j,f\ell})}[\mathbf{a}_{j,f\ell}\mathbf{a}_{j,f\ell}^{\mathrm{H}}] = \boldsymbol{\mu}_{\mathbf{a},j,f}\boldsymbol{\mu}_{\mathbf{a},j,f}^{\mathrm{H}} + \boldsymbol{\Sigma}_{\mathbf{a},j,f} \tag{5}$$

is the 2nd-order moment of the corresponding mixing filter. In conventional studies using the (time-invariant) convolutive model with $\mathbf{a}_{j,f}$ considered as a parameter, (4) holds with $\mathbf{R}_{\mathbf{a},j,f}$ being defined as $\mathbf{R}_{\mathbf{a},j,f} = \mathbf{a}_{j,f}\mathbf{a}_{j,f}^{\mathrm{H}}$ and thus limited to be rank-1. In the parametric context $\mathbf{R}_{\mathbf{a},j,f}$ is referred to as the spatial covariance matrix (SCM) of source j, and an extension to a full-rank SCM has been proposed in [3]. This full-rank matrix is assumed to model well a diffuse source, though interpreting this model in terms of the process generating the source image is not easy. An interpretation was given in [4, 6] in the form of a finite summation of punctual sources filtered by different filters, all considered as parameters during the source inference step. *In contrast, considering the mixing filter as a latent variable as proposed in the present study enables to directly define* $\mathbf{R}_{\mathbf{a},j,f}$ *as a full-rank matrix with* (5), *while keeping the mixture comfortably described by a simple convolutive model (i.e. one source-filter product per image source signal).* Obviously, the proposed filter model reduces to the parametric convolutive case when $\boldsymbol{\Sigma}_{\mathbf{a},j,f}$ tends to zero. Hence, we believe that the fully probabilistic model presented in the present paper generalizes —or at least provides an elegant interpretation of—, the "parametric" definition of the SCM. It actually provides an elegant probabilistic interpretation of both the generation of diffuse source signals, as a "probabilistic convolution" (a probabilistic source-filter product in the TF domain), and their estimation, as a continuous summation of source-filter products.

2.2 General expression of the source image MMSE estimator

Let us now provide some technical derivations, starting with a general formulation of the source image MMSE estimator. Eq. (2) writes:

$$\hat{\mathbf{y}}_{j,f\ell} = \int\int \mathbf{a}_{j,f\ell}s_{j,f\ell}p(\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}|\mathbf{x}_{f\ell})d\mathbf{A}_{f\ell}d\mathbf{s}_{f\ell}. \tag{6}$$

Since we have $p(\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}|\mathbf{x}_{f\ell}) = \frac{p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell},\mathbf{s}_{f\ell})p(\mathbf{A}_{f\ell})p(\mathbf{s}_{f\ell})}{p(\mathbf{x}_{f\ell})}$, (6) rewrites:

$$\hat{\mathbf{y}}_{j,f\ell} = \frac{1}{p(\mathbf{x}_{f\ell})} \int s_{j,f\ell} \Big(\int \mathbf{a}_{j,f\ell} p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}) p(\mathbf{A}_{f\ell}) d\mathbf{A}_{f\ell} \Big) p(\mathbf{s}_{f\ell}) d\mathbf{s}_{f\ell}. \quad (7)$$

Note that this expression is completely independent of the form of all densities. It only relies on definition (6) and the Bayes product rule. Obviously, (7) can be extended in the Bayesian sense by including priors on the parameters of the different distributions. In the present work, we stick to the above form.

2.3 The Gaussian case

In this section, we go a bit further and derive a "simplified" or "advanced" form of the source image MMSE estimator in the case where the mixing filters are assumed to follow a complex Gaussian distribution. For this aim, let us first specify and reshape $p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell})p(\mathbf{A}_{f\ell})$. As in several other studies, $\mathbf{b}_{f\ell}$ is assumed to be a zero-mean circular stationary complex Gaussian noise, i.e. $p(\mathbf{b}_{f\ell}) = \mathcal{N}_c(\mathbf{b}_{f\ell}; \mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b},f})$, where $\boldsymbol{\Sigma}_{\mathbf{b},f}$ is the noise covariance matrix to be estimated.[6] In addition, $\mathbf{b}_{f\ell}$ may be assumed to be isotropic, i.e. $\boldsymbol{\Sigma}_{\mathbf{b},f} = v_{\mathbf{b},f}\mathbf{I}_I$ with $v_{\mathbf{b},f} \in \mathbb{R}^+$ and $\mathbf{I}_I$ denoting the identity matrix of size I. We thus have $p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}) = \mathcal{N}_c(\mathbf{x}_{f\ell}; \mathbf{A}_{f\ell}\mathbf{s}_{f\ell}, \boldsymbol{\Sigma}_{\mathbf{b},f})$. Now it is natural to assume that the mixing filters $\mathbf{A}_{f\ell}$ follow a complex Gaussian prior distribution, since the latter is the conjugate prior of the Gaussian distribution for the mean parameter. For the sake of technical derivation, $\mathbf{A}_{f\ell}$ is first vectorized by vertically concatenating its J columns $\mathbf{a}_{j,f\ell}$ into a single column vector $\mathbf{a}_{:,f\ell}$, i.e. $\mathbf{a}_{:,f\ell} = \mathrm{vec}(\mathbf{A}_{f\ell}) = [\mathbf{a}_{1,f\ell}^\top, \ldots, \mathbf{a}_{J,f\ell}^\top]^\top \in \mathbb{C}^{IJ}$. Then we assume:

$$p(\mathbf{A}_{f\ell}) = p(\mathbf{a}_{:,f\ell}) = \mathcal{N}_c(\mathbf{a}_{:,f\ell}; \boldsymbol{\mu}_{\mathbf{a},f}, \boldsymbol{\Sigma}_{\mathbf{a},f}), \quad (8)$$

where the mean vector $\boldsymbol{\mu}_{\mathbf{a},f} \in \mathbb{C}^{IJ}$ and the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{a},f} \in \mathbb{C}^{IJ\times IJ}$ are parameters to be estimated. $\boldsymbol{\mu}_{\mathbf{a},f}$ is the concatenation of the individual mean mixing vectors $\boldsymbol{\mu}_{\mathbf{a},j,f}, j \in [1, J]$, defined for each source. $\boldsymbol{\Sigma}_{\mathbf{a},f}$ is block diagonal, assuming prior decorrelation of filters corresponding to different sources.

Let us then rewrite $\mathbf{A}_{f\ell}\mathbf{s}_{f\ell} = \sum_{j=1}^{J} \mathbf{a}_{j,f\ell}s_{j,f\ell} = (\mathbf{s}_{f\ell}^\top \otimes \mathbf{I}_I)\mathbf{a}_{:,f\ell} = \mathbf{U}_{f\ell}\mathbf{a}_{:,f\ell}$, with $\mathbf{U}_{f\ell} = \mathbf{s}_{f\ell}^\top \otimes \mathbf{I}_I$ ($\otimes$ denotes the Kronecker matrix product). Then, we can write:

$$p(\mathbf{x}_{f\ell}|\mathbf{a}_{:,f\ell}, \mathbf{s}_{f\ell})p(\mathbf{a}_{:,f\ell}) = p(\mathbf{a}_{:,f\ell}|\mathbf{x}_{f\ell}, \mathbf{s}_{f\ell})p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}), \quad (9)$$

since both sides are equal to $p(\mathbf{x}_{f\ell}, \mathbf{a}_{:,f\ell}|\mathbf{s}_{f\ell})$. Because $p(\mathbf{a}_{:,f\ell})$ is the conjugate prior of $p(\mathbf{x}_{f\ell}|\mathbf{a}_{:,f\ell}, \mathbf{s}_{f\ell})$, $p(\mathbf{a}_{:,f\ell}|\mathbf{x}_{f\ell}, \mathbf{s}_{f\ell})$ is a complex-Gaussian distribution that can be written $p(\mathbf{a}_{:,f\ell}|\mathbf{x}_{f\ell}, \mathbf{s}_{f\ell}) = \mathcal{N}_c(\mathbf{a}_{:,f\ell}; \boldsymbol{\mu}_{\mathbf{d},f\ell}, \boldsymbol{\Sigma}_{\mathbf{d},f\ell})$. Then, since $\mathbf{a}_{:,f\ell}$ is Gaussian and $\mathbf{b}_{f\ell}$ is Gaussian, it follows that $p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell})$ is a Gaussian distribution

[6] The proper complex Gaussian distribution is defined as $\mathcal{N}_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\pi\boldsymbol{\Sigma}|^{-1} \exp\big(-[\mathbf{x}-\boldsymbol{\mu}]^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}[\mathbf{x}-\boldsymbol{\mu}]\big)$, where $|.|$ denotes the matrix determinant [14].

that can be written $p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}) = \mathcal{N}_c(\mathbf{x}_{f\ell}; \boldsymbol{\mu}_{\mathbf{e},f\ell}, \boldsymbol{\Sigma}_{\mathbf{e},f\ell})$. Identifying the quadratic terms in $\mathbf{a}_{:,f\ell}$ in (9), we get:

$$\boldsymbol{\Sigma}_{\mathbf{d},f\ell}^{-1} = \mathbf{U}_{f\ell}^{\mathrm{H}} \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \mathbf{U}_{f\ell} + \boldsymbol{\Sigma}_{\mathbf{a},f}^{-1}. \tag{10}$$

Then, identifying the linear terms in $\mathbf{a}_{:,f\ell}$ in (9), we get:

$$\boldsymbol{\mu}_{\mathbf{d},f\ell} = \boldsymbol{\Sigma}_{\mathbf{d},f\ell}(\mathbf{U}_{f\ell}^{\mathrm{H}} \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \mathbf{x}_{f\ell} + \boldsymbol{\Sigma}_{\mathbf{a},f}^{-1} \boldsymbol{\mu}_{\mathbf{a},f}). \tag{11}$$

Then, identifying the quadratic terms in $\mathbf{x}_{f\ell}$ in (9) and applying the matrix inversion lemma [15, pp. 18-19], we get:

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{e},f\ell}^{-1} &= \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \mathbf{U}_{f\ell} \boldsymbol{\Sigma}_{\mathbf{d},f\ell} \mathbf{U}_{f\ell}^{\mathrm{H}} \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \\ \Leftrightarrow \boldsymbol{\Sigma}_{\mathbf{e},f\ell} &= \boldsymbol{\Sigma}_{\mathbf{b},f} + \mathbf{U}_{f\ell} \boldsymbol{\Sigma}_{\mathbf{a},f} \mathbf{U}_{f\ell}^{\mathrm{H}}. \end{aligned} \tag{12}$$

Finally, identifying the remaining linear terms in $\mathbf{x}_{f\ell}$, we get:

$$\boldsymbol{\mu}_{\mathbf{e},f\ell} = \boldsymbol{\Sigma}_{\mathbf{e},f\ell}(\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \mathbf{U}_{f\ell} \boldsymbol{\Sigma}_{\mathbf{d},f\ell} \boldsymbol{\Sigma}_{\mathbf{a},f}^{-1} \boldsymbol{\mu}_{\mathbf{a},f}). \tag{13}$$

Now we can inject (9) into (7), and we get:

$$\hat{\mathbf{y}}_{j,f\ell} = \frac{1}{p(\mathbf{x}_{f\ell})} \int s_{j,f\ell} \boldsymbol{\mu}_{\mathbf{d},j,f\ell} \mathcal{N}_c(\mathbf{x}_{f\ell}; \boldsymbol{\mu}_{\mathbf{e},f\ell}, \boldsymbol{\Sigma}_{\mathbf{e},f\ell}) p(\mathbf{s}_{f\ell}) d\mathbf{s}_{f\ell}, \tag{14}$$

with $\boldsymbol{\mu}_{\mathbf{d},j,f\ell}$ being the sub-vector of $\boldsymbol{\mu}_{\mathbf{d},f\ell}$ that corresponds to source j. If we concatenate the source images as $\mathbf{y}_{f\ell} = [\mathbf{y}_{1,f\ell}^{\top}, \ldots, \mathbf{y}_{J,f\ell}^{\top}]^{\top} \in \mathbb{C}^{IJ}$, we can rewrite (14) for all sources in compact form:

$$\hat{\mathbf{y}}_{f\ell} = \frac{1}{p(\mathbf{x}_{f\ell})} \int (\mathbf{s}_{f\ell} \otimes \mathbf{I}_I) \boldsymbol{\mu}_{\mathbf{d},f\ell} \mathcal{N}_c(\mathbf{x}_{f\ell}; \boldsymbol{\mu}_{\mathbf{e},f\ell}, \boldsymbol{\Sigma}_{\mathbf{e},f\ell}) p(\mathbf{s}_{f\ell}) d\mathbf{s}_{f\ell}. \tag{15}$$

Note that (14) and (15) are valid for any source distribution. In the following, we use the LGM with diagonal covariance matrix: $p(\mathbf{s}_{f\ell}) = \mathcal{N}_c\big(\mathbf{s}_{f\ell}; \mathbf{0}, \mathbf{v}_{f\ell} = \mathrm{diag}_J(v_{j,f\ell})\big)$. Even for such a classical source distribution, the integral in (14) or (15) has no closed-form expression since $\boldsymbol{\mu}_{\mathbf{d},f\ell}$ is a non-linear function of $\mathbf{s}_{f\ell}$ implying the inversion of a quadratic form (which is also present in $\boldsymbol{\mu}_{\mathbf{e},f\ell}$). Also, (14) or (15) requires the calculation of the observation marginal density $p(\mathbf{x}_{f\ell})$, which is a classical obstacle in inference problems. Therefore we have to turn towards sampling techniques.

2.4 Inference of source image using Metropolis algorithm

For the computation of values of the source image estimator (15), in the present study, we propose to use the Metropolis algorithm. Because this algorithm is very classical and quite basic, and because of room limitation, we will not present it into details. The reader is referred to [12] for a general overview of sampling techniques, and to [12, chap. 3] for a tutorial on the Metropolis algorithm.

3 Experiments

In this section, we report experiments conducted with three different stereo ($I = 2$) mixtures of $J = 3$ speech signals.[7] In Mix 1 and Mix 2, the source signals were monochannel 16 kHz signals randomly taken from the TIMIT database [16]. The source images $\mathbf{y}_j(t)$ were individually generated using the room impulse response (RIR) simulator of AudioLabs Erlangen.[8] The setting was the following: room size 7 m $\times$ 5 m $\times$ 2.5 m, sensor array placed at (3.5 m , 1.5 m , 1.5 m), distance between microphones $d = 0.15$ m, reverberation time $T_{60} = 150$ ms, source-to-sensor distance 1.2 m. In Mix 1, sources s_1, s_2 and s_3 are initially located at azimuths $-45°$, $0°$, $45°$, respectively, and they all move by $20°$ around the microphone array, within the signal duration of 2 s. In Mix 2, they start at azimuths $-75°$, $-25°$, $25°$ and they all move by $50°$. Finally, for Mix 3, three speakers were (separately) recorded in an office ($T_{60} \approx 0.6$ s). They were initially located at azimuths $-45°$, $0°$, $45°$, at 1.5 m from a two-microphone array (omnidirectional), and moved by about $45°$ in 2 s.

The STFT window was a 1024-point sine window with 50% overlap. The parameters $\boldsymbol{\mu}_{\mathbf{a},f}$ and $\boldsymbol{\Sigma}_{\mathbf{a},f}$ were set to "semi-oracle" values calculated from the individual source images. More precisely, for each $j \in [1, J]$, $\boldsymbol{\mu}_{\mathbf{a},j,f}$ and $\boldsymbol{\Sigma}_{\mathbf{a},j,f}$ were calculated from $\mathbf{y}_{j,f\ell}$, the STFT of $\mathbf{y}_j(t)$, following the spirit of the full-rank SCM initialization in [3]: $\mathbf{y}_{j,f\ell}$ was first normalized in phase, i.e. we calculated $\tilde{\mathbf{y}}_{j,f\ell} = \mathbf{y}_{j,f\ell} e^{-i\arg(y_{1,f\ell})}$; then $\boldsymbol{\mu}_{\mathbf{a},j,f}$ and $\boldsymbol{\Sigma}_{\mathbf{a},j,f}$ were calculated as the empirical mean and empirical covariance matrix of $\tilde{\mathbf{y}}_{j,f\ell}$, $\ell \in [1, L]$; finally, $v_{j,f\ell}$ was calculated for each frame by $v_{j,f\ell} = \frac{1}{I}\mathrm{trace}(\mathbf{R}_{\mathbf{a},j,f}^{-1}\mathbf{y}_{j,f\ell}\mathbf{y}_{j,f\ell}^H)$. The noise variance $v_{\mathbf{b},f}$ was set to 10^{-6} times the average PSD of the mixture signal. The semi-oracle setting of the parameters is of course an artificial close-to-optimal configuration that ensures very good separation performance (as verified in Table 1).

The computation of the Metropolis source image estimator was made using the semi-oracle values of the parameters and the mixture signal, with the PDF in the integral of (15) used as the target distribution and a complex-Gaussian distribution used as the candidate distribution. 15,000 samples were drawn at each TF bin (1,000 for burn-in). The separation of each 2s-mixture required about 4 hours on a 4-core 2.3GHz Intel Core i7 using the Matlab Parallel Toolbox. For comparison, the rank-1 Wiener estimator (R1W; as used in [2]) and the full-rank Wiener estimator (FRW; as used in [3]), using the same semi-oracle values of the parameters, were calculated as:

$$\hat{\mathbf{y}}_{j,f\ell} = v_{j,f\ell}\boldsymbol{\mu}_{\mathbf{a},j,f}\boldsymbol{\mu}_{\mathbf{a},j,f}^H \left(\sum_{k=1}^{J} v_{k,f\ell}\boldsymbol{\mu}_{\mathbf{a},k,f}\boldsymbol{\mu}_{\mathbf{a},k,f}^H + v_{\mathbf{b},f}\mathbf{I}_I\right)^{-1} \mathbf{x}_{f\ell}, \qquad (16)$$

$$\hat{\mathbf{y}}_{j,f\ell} = v_{j,f\ell}\mathbf{R}_{\mathbf{a},j,f} \left(\sum_{k=1}^{J} v_{k,f\ell}\mathbf{R}_{\mathbf{a},k,f} + v_{\mathbf{b},f}\mathbf{I}_I\right)^{-1} \mathbf{x}_{f\ell}. \qquad (17)$$

[7] Matlab code and data are available at: `www.gipsa-lab.grenoble-inp.fr/~laurent.girin/demo/lva2017.zip`.

[8] `www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator`.

		Mix 1			Mix 2			Mix 3		
Method	Meas.	s_1	s_2	s_3	s_1	s_2	s_3	s_1	s_2	s_3
Rank-1 Wiener	SDR	14.28	12.49	8.83	12.52	10.01	7.47	-3.39	-1.28	-2.55
	SAR	16.28	15.12	8.78	14.95	13.12	8.70	2.21	1.62	1.92
	SIR	17.18	14.25	7.77	15.52	12.71	7.05	0.56	-0.96	0.93
	ISR	16.30	18.70	14.56	14.59	14.88	14.14	1.35	3.73	3.96
Full-Rank Wiener	SDR	19.88	15.54	13.98	18.88	14.56	**13.67**	7.68	8.96	8.22
	SAR	22.43	17.25	**16.44**	20.89	**16.79**	12.05	**10.74**	**6.71**	**7.13**
	SIR	24.94	19.31	18.14	23.47	**19.46**	12.44	**11.56**	**6.14**	**7.27**
	ISR	24.04	20.29	19.74	23.33	18.59	19.38	11.82	12.64	12.26
Proposed	SDR	**19.99**	**15.82**	14.02	18.86	14.64	13.56	7.62	8.94	8.29
	SAR	**22.62**	**18.27**	16.24	**21.30**	16.46	**12.41**	9.70	6.14	6.85
	SIR	**26.24**	**21.93**	**18.38**	**25.16**	19.21	**13.05**	10.88	5.54	7.10
	ISR	**24.48**	**20.57**	**20.12**	23.32	**18.88**	**19.77**	**12.10**	**12.89**	**12.84**

Table 1. Separation performance (in dB). Best scores across methods are in bold (when the difference is larger than 0.1 dB).

Four standard audio source separation objective measures were calculated between the estimated and ground truth source images, namely: signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) signal-to-artifact ratio (SAR) and image-to-spatial distortion ratio (ISR) [17]. The results are presented in Table 1. We can see that both the FRW and the proposed estimator provide separation measures that are notably larger than the R1W. This confirms that both are able to efficiently exploit the spatial information on the mixture encoded in the SCM (remember that for each source, the SCM is equivalent to the second-order moment of the mixing filter, see (5)). For Mix 1 (sources moving relatively slowly), the proposed estimator performs globally better than the FRW. For Mix 2 (sources moving more rapidly), the results of the proposed estimator and FRW are more similar. Finally, the results for the real recordings tend to slightly favor FRW, even if the difference in SDR is especially small.[9]

4 Conclusion

Altogether, these results show the potential of the proposed method to overcome the state-of-the-art. As opposed to the Wiener filter build from the full-rank spatial covariance matrix of [3], the proposed source image estimator has the freedom to use the latter to independently estimate (an infinite set of) filter values at every frame and use it for image source estimation. In contrast, the Wiener filter of [3] directly uses the same spatial information at every frame. Yet, the results for real recordings are mitigated. The proposed estimator may be more sensible than the full-rank Wiener filter to the convolutive-to-multiplicative approximation for long mixing filters, for reasons that must be investigated. We will also work on improving the sampling scheme, and integrating the proposed estimator in a fully blind (iterative) separation process.

[9] So far, no statistical test could be performed on a large set of mixtures to test the significativity of the results because of the huge computational cost of the Metropolis.

References

1. E. Vincent, M. Jafari, S. Abdallah, M. Plumbley, and M. Davies, "Probabilistic modeling paradigms for audio source separation," *Machine Audition: Principles, Algorithms and Systems*, pp. 162–185, 2010.
2. A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.
3. N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.
4. A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.
5. N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet, "Linear mixing models for active listening of music productions in realistic studio conditions," in *Convention of the Audio Engineering Society (AES)*, Budapest, Hungary, 2012.
6. N. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for Gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 149, 2013.
7. T. Higuchi, N. Takamune, N. Tomohiko, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based on DOA-HMM," in *Proc. of the Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2014.
8. S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation with probabilistic reverberant modeling," in *IEEE Workshop Applicat. Signal Process. to Audio and Acoust. (WASPAA)*, 2015.
9. D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of moving sound sources," in *IEEE Workshop Applicat. Signal Process. to Audio and Acoust. (WASPAA)*, 2015.
10. V. Smidl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Berlin: Springer-Verlag, 2006.
11. A. Cemgil, C. Févotte, and S. Godsill, "Variational and stochastic inference for Bayesian source separation," *Digital Signal Processing*, vol. 2007, no. 17, pp. 891–913, 2007.
12. F. Liang, C. Liu, and R. Carroll, *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. Wiley, Aug. 2010.
13. S. Gannot and M. Moonen, "On the application of the unscented Kalman filter to speech processing," in *IEEE Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, 2003, p. 811.
14. F. Neeser and J. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Info. Theory*, vol. 39, no. 4, pp. 1293–1302, 1993.
15. R. Horn and C. Johnson, *Matrix analysis*. Cambridge: Cambridge University Press, 1985.
16. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," 1993, Linguistic Data Consortium, Philadelphia.
17. E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Int. Conf. on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.