Multichannel audio source separation: variational inference of time-frequency sources from time-domain observations

Simon Leglaive, Roland Badeau, Gaël Richard

MULTICHANNEL AUDIO SOURCE SEPARATION: VARIATIONAL INFERENCE OF TIME-FREQUENCY SOURCES FROM TIME-DOMAIN OBSERVATIONS

Simon Leglaive, Roland Badeau, Gaël Richard

LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France.

ABSTRACT

A great number of methods for multichannel audio source separation are based on probabilistic approaches in which the sources are modeled as latent random variables in a Time-Frequency (TF) domain. For reverberant mixtures, it is common to approximate the time-domain convolutive mixing process as being instantaneous in the short-term Fourier transform domain, under a short mixing filters assumption. The TF latent sources are then inferred from the TF mixture observations. In this paper we propose to infer the TF latent sources from the time-domain observations. This approach allows us to exactly model the convolutive mixing process. The inference procedure relies on a variational expectation-maximization algorithm. In significant reverberation conditions, our approach leads to a signal-to-distortion ratio improvement of 5.5 dB compared with the usual TF approximation of the convolutive mixing process.

***Index Terms*—** Multichannel audio source separation, time-domain convolutive model, time-frequency source model, non-negative matrix factorization, variational EM algorithm.

1. INTRODUCTION

Audio source separation is the task that aims to recover a set of audio source signals from one or several mixtures. When the sources are recorded in an enclosed space, the reflections of sound on the surfaces and objects of the room induce reverberation in the recordings. Modeling such reverberant mixtures basically involves two stages: modeling the source signals and the way they are mixed together.

Source separation is commonly achieved in a Time-Frequency (TF) domain because it provides a meaningful and often sparse representation of the source signals. In this TF domain, many approaches are based on probabilistic modeling and statistical inference [1, 2]. Within such a framework, Non-negative Matrix Factorization (NMF) techniques are popular to represent the spectro-temporal characteristics of the sources [3, 4, 5, 6].

The propagation between a punctual source and a microphone in a room can be characterized by a room impulse response referred to as a *mixing filter* in the context of source separation. A signal recorded by a microphone is thus generally represented as the convolution of a source signal with a mixing filter. Modeling convolutive mixtures in a TF domain is quite challenging. Therefore, most of the methods rely on the assumption that the mixing filters are short compared with the TF analysis window [7]. Under this hypothesis the convolutive mixing process can be approximated as instantaneous in each frequency band (see, *e.g.*, [8], [9]). This assumption, although widely used, remains one of the main limitations of reverberant audio source separation methods. It prevents us from obtaining good separation results in recordings with significant reverberation.

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).

Some methods working with TF source models have nonetheless investigated other mixture models. In [10] the authors introduced spatial covariance matrices to model non-punctual sources. They experimentally showed that the flexibility of this model helps improving the performance of reverberant audio source separation. The method presented in [11] considers a time-domain modeling of the mixture while keeping a sparse constraint on the sources in the TF domain. In [12] the time-domain convolution is approximated by subband filtering in the Short-Term Fourier Transform (STFT) domain. Finally it has been shown in [13] that convolution in the time domain can be accurately represented in the TF domain by a two-dimensional filter.

In this paper the sources are modeled in the Modified Discrete Cosine Transform (MDCT) domain as centered Gaussian random variables, whose variances are structured by means of an NMF model. The convolutive mixing process is exactly modeled by staying in the time domain. We then use the time-domain observations to infer the TF latent source variables[1]. Our inference procedure relies on a variational expectation-maximization algorithm.

The models are introduced in section 2. The variational inference is presented in section 3. Oracle experiments are conducted in section 4 and we finally draw conclusions in section 5.

2. MODELS

We denote $s_j(t) \in \mathbb{R}$, $t = 0, ..., L_s - 1$, $j = 1, ..., J$, the j-th source signal and $a_{ij}(t) \in \mathbb{R}$, $t = 0, ..., L_a$, $i = 1, ..., I$, the mixing filter between source j and microphone i. Let define $T = L_s + L_a - 1$. The signal $x_i(t)$ recorded by the i-th microphone is represented for $t = 0, ..., T - 1$ as:

$$x_i(t) = \sum_{j=1}^{J} y_{ij}(t) + b_i(t), \tag{1}$$

where $y_{ij}(t) = [a_{ij} \star s_j](t)$ is referred to as a source image, with $\star$ the discrete convolution operator, and $b_i(t)$ is an additive noise.

Each signal $s_j(t)$ is represented by a set of TF synthesis coefficients $\{s_{j,fn} \in \mathbb{R}\}_{f,n}$, $(f, n) \in \{0, ..., F - 1\} \times \{0, ..., N - 1\}$:

$$s_j(t) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} s_{j,fn} \psi_{fn}(t). \tag{2}$$

In this work we use the MDCT for representing the source signals. $\psi_{fn}(t) \in \mathbb{R}$, $t = 0, ..., L_s - 1$, is thus an MDCT synthesis atom:

$$\psi_{fn}(t) = \sqrt{\frac{2}{F}} w(t - nH) \cos\left(\frac{2\pi}{L_w}\left(t - nH + \frac{1}{2} + \frac{L_w}{4}\right)\left(f + \frac{1}{2}\right)\right), \tag{3}$$

[1] In [14] time-domain observations were also used for inferring TF sources. However the convolutive mixture model was approximated in the STFT domain which is not the case in our approach.

where $w(t)$ is a sine-window defined by $w(t) = \sin(\pi(t+0.5)/L_w)$ if $0 \leq t \leq L_w - 1$, 0 otherwise, and $H = L_w/2$ is the hop size. Note that $F = L_w/2$. We choose the MDCT for mainly two reasons: firstly we do not need to approximate the time-domain convolutive mixing process in the TF domain, which is usually done for practical reasons in the STFT domain (see, *e.g.*, [7]); secondly the STFT is a redundant complex-valued transform so we need to deal with twice more real valued coefficients than with an MDCT. Moreover the MDCT exhibits the preservation of whiteness property [15]. It is thus more appropriate than the STFT for assuming the independence of the source TF points, as commonly done in audio source separation. Note that expression (2) allows one to use different TF resolutions for representing different sources. This can be useful for separating tonal and transient components [16, 17]. This extension is however let for future work as in this paper we choose the same TF dictionary for all the sources.

In a similar way as in [16], the synthesis coefficients are modeled as centered real-valued Gaussian random variables, whose variances are structured by means of an NMF model:

$$s_{j,fn} \sim \mathcal{N}(0, v_{j,fn} = [\mathbf{W}_j \mathbf{H}_j]_{fn}), \tag{4}$$

with $\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$, $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times N}$ and $\mathcal{N}(\mu, \sigma^2)$ is the real-valued Gaussian distribution with Probability Density Function (PDF):

$$N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right). \tag{5}$$

In the single-channel coding-based informed source separation method [18], the authors also use an MDCT along with a Gaussian source model and a non-negative factorization of the source power.

From equation (2) a source image writes:

$$y_{ij}(t) = [a_{ij} \star s_j](t) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} s_{j,fn} g_{ij,fn}(t), \tag{6}$$

where $g_{ij,fn}(t) = [a_{ij} \star \psi_{fn}](t)$. We finally assume a white Gaussian additive noise:

$$b_i(t) \sim \mathcal{N}(0, \sigma_i^2). \tag{7}$$

3. VARIATIONAL INFERENCE

Let $\mathbf{x}$ denote the set of observed time-domain variables $\{x_i(t)\}_{i,t}$, $\mathbf{s}$ the set of TF-domain latent variables $\{s_{j,fn}\}_{j,f,n}$ and $\boldsymbol{\theta}$ the set of parameters $\{\sigma_i^2\}_i$, $\{\mathbf{W}_j\}_j$, $\{\mathbf{H}_j\}_j$ and $\{a_{ij}(t)\}_{i,j,t}$. Our objective is to estimate the latent variables in a Minimum Mean Square Error (MMSE) sense:

$$\hat{\mathbf{s}} = \mathbb{E}_{\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^\star}[\mathbf{s}], \tag{8}$$

where the model parameters are estimated in a Maximum Likelihood (ML) sense:

$$\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \; p(\mathbf{x}; \boldsymbol{\theta}). \tag{9}$$

The solution of (8)-(9) can be found by means of an Expectation-Maximization (EM) algorithm [19]. However, according to the models defined in section 2, the posterior distribution $p(\mathbf{s}|\mathbf{x}; \boldsymbol{\theta})$ is Gaussian but parametrized by a full covariance matrix of too high dimensions to be implemented. We are thus resorting to a Variational EM algorithm (VEM) [20] in order to factorize this posterior distribution. Let $\mathcal{F}$ be a set of PDFs over the latent variables $\mathbf{s}$. For any PDF $q \in \mathcal{F}$ and any function $f(\mathbf{s})$, we note $\langle f(\mathbf{s}) \rangle_q = \int f(\mathbf{s}) q(\mathbf{s}) d\mathbf{s}$. Then for any $q \in \mathcal{F}$ and parameter set $\boldsymbol{\theta}$, the log-likelihood can be decomposed as:

$$\ln p(\mathbf{x}; \boldsymbol{\theta}) = \mathcal{L}(q; \boldsymbol{\theta}) + D_{\text{KL}}(q || p(\mathbf{s}|\mathbf{x}; \boldsymbol{\theta})), \tag{10}$$

where $\mathcal{L}(q; \boldsymbol{\theta}) = \left\langle \ln\left(\frac{p(\mathbf{x},\mathbf{s};\boldsymbol{\theta})}{q(\mathbf{s})}\right) \right\rangle_q$ and $D_{\text{KL}}(q||p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})) = \left\langle \ln\left(\frac{q(\mathbf{s})}{p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})}\right) \right\rangle_q$ is the Kullback-Leibler (KL) divergence between q and the posterior distribution $p(\mathbf{s}|\mathbf{x}; \boldsymbol{\theta})$. $\mathcal{L}(q; \boldsymbol{\theta})$ is called the variational free energy and can be further decomposed as $\mathcal{L}(q; \boldsymbol{\theta}) = E(q; \boldsymbol{\theta}) + H(q)$ where

$$E(q; \boldsymbol{\theta}) = \langle \ln p(\mathbf{x}, \mathbf{s}; \boldsymbol{\theta}) \rangle_q, \tag{11}$$

and $H(q) = -\langle \ln q(\mathbf{s}) \rangle_q$ is the entropy of distribution q. Since the KL divergence is always non-negative, the variational free energy is a lower bound of the log-likelihood. The variational EM algorithm consists in iterating two steps until convergence: the E-step where we compute $q^\star = \arg\max_{q \in \mathcal{F}} \mathcal{L}(q; \boldsymbol{\theta}_{\text{old}})$ and the M-step where we compute $\boldsymbol{\theta}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q^\star; \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} E(q^\star; \boldsymbol{\theta})$.

In this work we will use the mean field approximation where we assume that the density q can be factorized as:

$$q(\mathbf{s}) = \prod_{j=1}^{J} \prod_{f=0}^{F-1} \prod_{n=0}^{N-1} q_{jfn}(s_{j,fn}). \tag{12}$$

3.1. Source estimate under the variational approximation

$q(\mathbf{s})$ as defined in (12) aims to approximate the posterior distribution $p(\mathbf{s}|\mathbf{x}; \boldsymbol{\theta})$. Under this approximation the j-th TF source estimate is:

$$m_{j,fn} = \langle s_{j,fn} \rangle_{q^\star}. \tag{13}$$

The j-th time-domain source estimate is then obtained for $t = 0, ..., L_s - 1$ by inverse MDCT:

$$\hat{s}_j(t) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{j,fn} \psi_{fn}(t). \tag{14}$$

From the current mixing filters, we also define the estimate of the j-th source image seen by the i-th microphone for $t = 0, ..., T-1$:

$$\hat{y}_{ij}(t) = [a_{ij} \star \hat{s}_j](t) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{j,fn} g_{ij,fn}(t). \tag{15}$$

3.2. Variational free energy

From equations (1) to (7), the complete data log-likelihood $\ln p(\mathbf{x}, \mathbf{s}; \boldsymbol{\theta}) = \ln p(\mathbf{x}|\mathbf{s}; \boldsymbol{\theta}) + \ln p(\mathbf{s}; \boldsymbol{\theta})$ writes:

$$\begin{aligned} \ln p(\mathbf{x}, \mathbf{s}; \boldsymbol{\theta}) = & -\frac{1}{2}(IT + JFN)\ln(2\pi) \\ & - \frac{1}{2}\sum_{i=1}^{I}\sum_{t=0}^{T-1}\left[\ln(\sigma_i^2) + \frac{1}{\sigma_i^2}\left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t)\right)^2\right] \\ & - \frac{1}{2}\sum_{j=1}^{J}\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[\ln(v_{j,fn}) + \frac{s_{j,fn}^2}{v_{j,fn}}\right]. \end{aligned} \tag{16}$$

We are interested in computing the variational free energy $\mathcal{L}(q^\star; \boldsymbol{\theta}) = E(q^\star; \boldsymbol{\theta}) + H(q^\star)$. From (11) and (16) we have:

$$\begin{aligned} E(q^\star; \theta) = & -\frac{1}{2}(IT + JFN)\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{I}\sum_{t=0}^{T-1}\left[\ln(\sigma_i^2)\right. \\ & \left. + \frac{e_i(t)}{\sigma_i^2}\right] - \frac{1}{2}\sum_{j=1}^{J}\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[\ln(v_{j,fn}) + \frac{m_{j,fn}^2 + \gamma_{j,fn}}{v_{j,fn}}\right], \end{aligned} \tag{17}$$

where $m_{j,fn}$ is defined in (13), $e_i(t) = \left\langle \left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t)\right)^2 \right\rangle_{q^\star}$ and $\gamma_{j,fn} = \langle (s_{j,fn} - m_{j,fn})^2 \rangle_{q^\star}$. From (6), (12) and (15), $e_i(t)$ is given by:

$$e_i(t) = \left(x_i(t) - \sum_{j=1}^{J} \hat{y}_{ij}(t) \right)^2 + \sum_{j=1}^{J} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \gamma_{j,fn} g_{ij,fn}^2(t). \tag{18}$$

From (12) and after having identified $q(\mathbf{s})^\star$ at the E-step we will have to compute the entropy $H(q^\star) = -\langle \ln(q^\star(\mathbf{s})) \rangle_{q^\star}$.

3.3. E-step

Under the mean-field approximation, we can show that the densities $q_{jfn}(s_{j,fn})$ which maximize the variational free energy satisfy [20]:

$$\ln q_{jfn}(s_{j,fn}) = \langle \ln p(\mathbf{x}, \mathbf{s}; \boldsymbol{\theta}) \rangle_{\left(\prod_{(j',f',n') \neq (j,f,n)} q_{j'f'n'} \right)}. \tag{19}$$

From (16) we can develop this expression. After computation we find that $q^\star_{jfn}(s_{j,fn}) = N(s_{j,fn}; m_{j,fn}, \gamma_{j,fn})$ where:

$$\gamma_{j,fn} = \left(\frac{1}{v_{j,fn}} + \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \sum_{t=0}^{T-1} g_{ij,fn}^2(t) \right)^{-1}; \tag{20}$$

$$m_{j,fn} = m_{j,fn} - \gamma_{j,fn} d_{j,fn}; \tag{21}$$

$$d_{j,fn} = \frac{m_{j,fn}}{v_{j,fn}} - \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \sum_{t=0}^{T-1} g_{ij,fn}(t) \left(x_i(t) - \sum_{j'=1}^{J} \hat{y}_{ij'}(t) \right). \tag{22}$$

Note that the parameters $m_{j,fn}$ have to be updated in turn[2].

Entropy of the distribution: From (12) and the variational distribution $q^\star_{jfn}(s_{j,fn})$ we identified, we can show that:

$$H(q^\star) = \frac{JFN}{2} \left(1 + \ln(2\pi) \right) + \frac{1}{2} \sum_{j=1}^{J} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \ln(\gamma_{j,fn}). \tag{23}$$

Preconditioned conjugate gradient method: We can show that $d_{j,fn} = \partial(-\mathcal{L}(q^\star; \boldsymbol{\theta}))/(\partial m_{j,fn})$. Therefore, we clearly see from (21) that the update of $m_{j,fn}$ corresponds to going in the opposite direction of the derivative $d_{j,fn}$ with a step size $\gamma_{j,fn}$. When the derivative is zero, it is clear that we achieve a fixed point of the algorithm. Therefore, for the sake of computational efficiency, we will use the preconditioned conjugate gradient (PCG) method [21] instead of the coordinate-wise update (21). We define $\mathbf{d}$ and $\mathbf{m}$ the column vectors of size JFN with entries $d_{j,fn}$ and $m_{j,fn}$ respectively. We also define the diagonal preconditioning matrix $\mathbf{D}$ of size $JFN \times JFN$ and entries $\gamma_{j,fn}^{-1}$. The order of the coefficients indexed by j, f, n for constructing these vectors and this diagonal matrix does not matter as long as it is kept identical. The E-step finally corresponds to updating $\gamma_{j,fn}$ according to (20) and updating $m_{j,fn}$ with the PCG method summarized in Algorithm 1.

3.4. M-step

The M-step aims to maximize $E(q^\star; \boldsymbol{\theta})$ in (17) with respect to (w.r.t) the parameter set $\boldsymbol{\theta}$. Zeroing the derivative of $E(q^\star; \boldsymbol{\theta})$ w.r.t σ_i^2 leads to the following update with $e_i(t)$ given by (18):

$$\sigma_i^2 = \frac{1}{T} \sum_{t=0}^{T-1} e_i(t). \tag{24}$$

Up to an additive term which does not depend on the NMF parameters, we can recognize in (17) the Itakura-Saito (IS) divergence [3] between $v_{j,fn} = [\mathbf{W}_j \mathbf{H}_j]_{fn}$ and the posterior mean of the source power spectrogram $\langle s_{j,fn}^2 \rangle_{q^\star} = m_{j,fn}^2 + \gamma_{j,fn}$. Therefore the source parameters are updated by computing an NMF on $\hat{\mathbf{P}}_j = \left[m_{j,fn}^2 + \gamma_{j,fn} \right]_{fn} \in \mathbb{R}_+^{F \times N}$ with the IS divergence. It can be done with the standard multiplicative update rules given in [3].

We can re-write the function to be maximized w.r.t the mixing filters as:

$$C(\mathbf{a}_{ij}) = -\frac{1}{2} \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \left[\left\| \mathbf{x}_i - \sum_{j=1}^{J} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{j,fn} \mathbf{T}_{fn} \mathbf{a}_{ij} \right\|_2^2 + \sum_{j=1}^{J} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \gamma_{j,fn} \left\| \mathbf{T}_{fn} \mathbf{a}_{ij} \right\|_2^2 \right]. \tag{25}$$

with $\mathbf{T}_{fn} \in \mathbb{R}^{T \times L_a}$ a Toeplitz matrix [21] containing the TF atoms $\psi_{fn}(t)$, $\mathbf{x}_i$ a column vector of size T and entries $x_i(t)$ and $\mathbf{a}_{ij}$ a column vector of size L_a and entries $a_{ij}(t)$. The convolution operation is thus expressed thanks to the multiplication of a vector with a Toeplitz matrix. Zeroing the gradient of $C(\mathbf{a}_{ij})$ is equivalent to solving the following positive definite linear system of equations:

$$\begin{aligned} &\left[\left(\sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{j,fn} \mathbf{T}_{fn} \right)^T \left(\sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{j,fn} \mathbf{T}_{fn} \right) \right. \\ &\left. + \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \gamma_{j,fn} \mathbf{T}_{fn}^T \mathbf{T}_{fn} \right] \mathbf{a}_{ij} \\ &= \left[\sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{j,fn} \mathbf{T}_{fn}^T \left(\mathbf{x}_i - \sum_{j' \neq j} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{j',fn} \mathbf{T}_{fn} \mathbf{a}_{ij'} \right) \right]. \end{aligned} \tag{26}$$

Let introduce the following definitions:

- ▷ $\epsilon_{ij}(t) = x_i(t) - \sum_{j' \neq j} \hat{y}_{ij'}(t)$ for $t = 0, ..., T-1$;
- ▷ $T_M\{\tau(k)\}$: A symmetric Toeplitz matrix of size $M \times M$ formed from the sequence $\{\tau(k)\}_{k=0}^{M-1}$.

Algorithm 1: PCG method for the update of $m_{j,fn}$ at the E-step

Initialize $\mathbf{d}$ from (22) and $\boldsymbol{\omega} = \mathbf{D}^{-1}\mathbf{d}$
while stopping criterion not reached **do**
  Compute $\boldsymbol{\kappa}$ the column vector of size JFN with entries
  $\kappa_{j,fn} = \frac{\omega_{j,fn}}{v_{j,fn}} + \sum_{i=1}^{I} \frac{1}{\sigma_i^2} \sum_{t=0}^{T-1} g_{ij,fn}(t) \times \sum_{j'=1}^{J} \sum_{f'=0}^{F-1} \sum_{n'=0}^{N-1} \omega_{j',f'n'} g_{ij',f'n'}(t)$
  $\mu = (\boldsymbol{\omega}^T \mathbf{d})/(\boldsymbol{\omega}^T \boldsymbol{\kappa})$
  $\mathbf{m} \leftarrow \mathbf{m} - \mu \boldsymbol{\omega}$
  Compute $\mathbf{d}$ from (22)
  $\mathbf{d}_p = \mathbf{D}^{-1}\mathbf{d}$
  $\beta = -(\boldsymbol{\kappa}^T \mathbf{d}_p)/(\boldsymbol{\omega}^T \boldsymbol{\kappa})$
  $\boldsymbol{\omega} \leftarrow \mathbf{d}_p + \beta \boldsymbol{\omega}$
end while

[2] $m_{j,fn}$ appears several times in the right-hand side of (21) and it can be shown that its contributions add up to zero.

We can show that (26) can be re-written as:

$$\left[T_{L_a}\{\hat{r}_j^{ss}(k)\} + T_{L_a}\left\{\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\gamma_{j,fn}\hat{r}_{fn}^{\psi\psi}(k)\right\}\right]\mathbf{a}_{ij} = \hat{\mathbf{r}}_{ij}^{s\epsilon}, \quad (27)$$

where $\hat{r}_j^{ss}(k) = \sum_{t=0}^{L_s-1-k}\hat{s}_j(t)\hat{s}_j(t+k)$, $\hat{r}_{ij}^{s\epsilon}(k) = \sum_{t=0}^{L_s-1}\hat{s}_j(t)\epsilon_{ij}(t+k)$, $\hat{\mathbf{r}}_{ij}^{s\epsilon} = \left[\hat{r}_{ij}^{s\epsilon}(0), \hat{r}_{ij}^{s\epsilon}(1), \cdots, \hat{r}_{ij}^{s\epsilon}(L_a-1)\right]^T$ and $\hat{r}_{fn}^{\psi\psi}(k) = \sum_{t=0}^{L_s-1-k}\psi_{fn}(t)\psi_{fn}(t+k)$. Compared with (26), this formulation involves simpler and faster operations such as inverse MDCTs, convolutions and cross-correlations. The linear system (27) can be solved by matrix inversion. However, for the sake of computational efficiency in the case of long mixing filters and numerical stability we choose to use the PCG method [21]. As it is rather common for solving positive definite linear systems and due to space limitations we do not detail the algorithm here but its structure is identical to Algorithm 1.

4. EXPERIMENTS

Our experiments are conducted from the audio tracks without effects provided by the Musical Audio Signal Separation (MASS) dataset [22]. We consider 8 stereo mixtures created by simulating mixing filters with the Roomsimove toolbox [23]. The mixtures duration ranges from 12 to 28 seconds. We considered several reverberation times[3]: $T_{60} = 32, 64, 128, 256$ or 512 ms. It results in a total number of 40 mixtures. The number of sources per mixture ranges from 3 to 5. The omnidirectional microphone spacing was set to 1 m, and the distance between the source and the center of the microphone pair to 2 m. The sources directions of arrival range from -45° to 45°. 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.

We compare our approach with the baseline method [8][4]. Indeed, our work is comparable to [8] in the sense that it performs multichannel audio source separation using an NMF source model and a punctual convolutive mixture model. The key difference is that in this method, the authors infer the TF latent sources from TF observations (with an EM algorithm and using the STFT). Therefore they approximate the convolutive mixing process in the STFT domain assuming short mixing filters. In our work, the inference is done from the time-domain observations, the convolutive mixture modeling is thus exact. For both methods we use a half-overlapping TF analysis/synthesis sine window of 128 ms (2048 points at a sampling rate of 16 kHz). The NMF order is arbitrarily fixed to $K_j = 10$ for all the sources. For both algorithms the parameters are initialized from oracle values computed on the true source signals and from the true mixing filters. Indeed, this work aims to compare the best performance achievable by the two approaches. The (V)EM algorithms are run for 100 iterations from the oracle initializations. The PCG algorithms for the proposed method are run for 10 iterations. 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 [24]. We used the BSS Eval Toolbox available at [25] to compute these measures.

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

[4]Except that the NMF parameters are updated as in [9] using multiplicative update rules thanks to a change in the choice of the latent variables.

	SDR		ISR		SIR		SAR	
T_{60} (ms)	ref.	new	ref.	new	ref.	new	ref.	new
32	**16.7**	16.0	**24.3**	23.0	**24.4**	23.0	**18.6**	**18.6**
64	14.9	**15.6**	21.5	**22.6**	22.4	**22.6**	17.1	**18.5**
128	11.8	**15.3**	17.6	**22.3**	18.8	**22.6**	14.6	**18.2**
256	8.5	**13.8**	13.7	**20.5**	14.7	**21.3**	12.0	**16.7**
512	6.3	**11.8**	10.9	**18.1**	11.8	**19.2**	10.1	**14.6**

Table 1. Average source separation results in dB according to the reverberation time T_{60}. "Ref." denotes the baseline approach with TF approximation of the convolutive mixing process and "new" is the proposed method with exact convolutive mixture modeling.

The results are presented in Table 1. Interestingly, we see that when $T_{60} = 32$ ms the baseline approach leads to slightly better performance than the proposed one, the short mixing filters assumption being verified. In [8] the filters length is assumed to be equal to the STFT analysis window length. Therefore the filters are here overparametrized which can be favorable in an oracle setting. Nevertheless, for all the other reverberation times, the proposed method with exact convolutive mixture modeling performs better. The longer the reverberation time, the better it performs compared with the baseline. Indeed, for $T_{60} = 64$ ms the proposed method leads to an SDR improvement of 0.7 dB while for $T_{60} = 512$ ms the improvement reaches 5.5 dB. Informal listening tests confirm this improvement. While for the baseline method the reverberation seems to be spread over the estimated sources, leading to strong interferences at high T_{60}, the separation quality of the proposed method is much more constant when the reverberation time increases. Audio examples and Matlab code are available from our demo web page [26].

As in this work we added the time-domain dimension in the formulation of the source separation problem, our method is more computational demanding. Moreover the computational load increases with the length of the mixing filters. For example, on a 28 seconds long mixture of 3 sources, for the previously mentioned reverberation times in ascending order, one iteration of the VEM algorithm takes about 33, 37, 47, 69 and 111 seconds respectively, with a 3.70 GHz processor. While one iteration of the EM algorithm for the baseline method takes around 1 second.

5. CONCLUSIONS

In this paper we presented a new method for multichannel audio source separation based on exact convolutive mixture modeling. Within this framework, TF latent sources are inferred from the time-domain mixture observations. We showed that this approach, under significant reverberation, considerably improved oracle performance compared with the usual TF approximation of the convolutive mixture. Future work will aim to confirm these results on live-recordings and also evaluate the method on professionally produced music. We will also have to investigate a more realistic blind initialization procedure. The initialization is indeed crucial for an EM algorithm.

As shown in [16] for single channel source separation, using the generative time-domain source model (2)-(4) allows one to use TF dictionaries with multiple resolutions for modeling the sources. This could be useful for harmonic/percussive multichannel source separation for example. This extension is let for future work.

We also believe that modeling the mixture in the time-domain is a promising approach for incorporating probabilistic priors on the mixing filters. Indeed, they exhibit a simple specific structure in time as they correspond to room impulse responses [27].

6. REFERENCES

[1] Emmanuel Vincent, Maria G. Jafari, Samer A. Abdallah, Mark D. Plumbley, and Mike E. Davies, "Probabilistic modeling paradigms for audio source separation," *Machine Audition: Principles, Algorithms and Systems*, pp. 162–185, 2010.

[2] Alexey Ozerov, Emmanuel Vincent, and Frédéric 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.

[3] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[4] Tuomas Virtanen, Ali Taylan Cemgil, and Simon Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 1825–1828.

[5] Antoine Liutkus, Derry Fitzgerald, and Roland Badeau, "Cauchy nonnegative matrix factorization," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015, pp. 1–5.

[6] Umut Şimşekli, Antoine Liutkus, and Ali Taylan Cemgil, "Alpha-stable matrix factorization," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2289–2293, 2015.

[7] Lucas Parra and Clay Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[8] Alexey Ozerov and Cédric 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.

[9] Alexey Ozerov, Cédric Févotte, Raphaël Blouet, and Jean-Louis Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 257–260.

[10] Ngoc Q. K. Duong, Emmanuel Vincent, and Rémi 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.

[11] Matthieu Kowalski, Emmanuel Vincent, and Rémi Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1818–1829, 2010.

[12] Hagai Attias, "New EM algorithms for source separation and deconvolution with a microphone array.," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, 2003, pp. 297–300.

[13] Roland Badeau and Mark D. Plumbley, "Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 11, pp. 1670–1680, 2014.

[14] Alexey Ozerov, Cagdas Bilen, and Patrick Pérez, "Multichannel audio declipping," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016.

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

[16] Cédric Févotte and Matthieu Kowalski, "Low-rank time-frequency synthesis," in *Proc. of Advances in Neural Information Processing Systems*, 2014, pp. 3563–3571.

[17] Fangchen Feng and Matthieu Kowalski, "Hybrid model and structured sparsity for under-determined convolutive audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 6682–6686.

[18] Alexey Ozerov, Antoine Liutkus, Roland Badeau, and Gaël Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1699–1712, 2013.

[19] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (Methodological)*, pp. 1–38, 1977.

[20] Matthew James Beal, *Variational algorithms for approximate Bayesian inference*, Ph.D. thesis, Univ. College of London, 2003.

[21] Gene H. Golub and Charles F. Van Loan, *Matrix computations*, Johns Hopkins University Press, 1996.

[22] Marc Vinyes, "MTG MASS dataset," `http://mtg.upf.edu/download/datasets/mass`, 2008.

[23] Emmanuel Vincent and Douglas R. Campbell, "Roomsimove," `http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip`, 2008.

[24] Emmanuel Vincent, Shoko Araki, Fabian J. Theis, Guido Nolte, Pau Bofill, Hiroshi Sawada, Alexey Ozerov, B. Vikrham Gowreesunker, Dominik Lutter, and Ngoc Q. K. Duong, "The Signal Separation Evaluation Campaign (2007-2010): Achievements and Remaining Challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

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

[26] "Demo web page," `http://perso.telecom-paristech.fr/leglaive/demo-icassp17.html`.

[27] Emanuël A. P. Habets, "Speech dereverberation using statistical reverberation models," in *Speech Dereverberation*, pp. 57–93. Springer, 2010.