Separating Time-Frequency Sources from Time-Domain Convolutive Mixtures Using Non-negative Matrix Factorization

Simon Leglaive, Roland Badeau, Gael Richard

SEPARATING TIME-FREQUENCY SOURCES FROM TIME-DOMAIN CONVOLUTIVE MIXTURES USING NON-NEGATIVE MATRIX FACTORIZATION

Simon Leglaive, Roland Badeau, Gaël Richard

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

ABSTRACT

This paper addresses the problem of under-determined audio source separation in multichannel reverberant mixtures. We target a semi-blind scenario assuming that the mixing filters are known. Source separation is performed from the time-domain mixture signals in order to accurately model the convolutive mixing process. The source signals are however modeled as latent variables in a time-frequency domain. In a previous paper we proposed to use the modified discrete cosine transform. The present paper generalizes the method to the use of the odd-frequency short-time Fourier transform. In this domain, the source coefficients are modeled as centered complex Gaussian random variables whose variances are structured by means of a non-negative matrix factorization model. The inference procedure relies on a variational expectation-maximization algorithm. In the experiments we discuss the choice of the source representation and we show that the proposed approach outperforms two methods from the literature.

Index Terms— Audio source separation, reverberant mixtures, non-negative matrix factorization, variational inference.

1. INTRODUCTION

Multichannel audio source separation consists in recovering several source signals from the observation of a mixture recorded with multiple microphones. In this paper we consider under-determined mixtures where the number of microphones is lower than the number of sources to be estimated. Moreover we focus on separating reverberant (or convolutive) mixtures, assuming a semi-blind scenario where the mixing filters are known.

In an under-determined context, source separation methods commonly work with a time-frequency (TF) representation of the source signals. Indeed, model-based approaches can take advantage of the very particular structure of audio signals in the TF plane [1]. For example, sparse component analysis methods [2] exploit the sparsity of the source signals in the TF domain. They rely on stationary super-Gaussian priors or they make use of deterministic approaches based on sparsity inducing penalties. Another important trend in audio source separation corresponds to the variance modeling framework [3]. Non-negative matrix factorization (NMF) techniques are especially popular for representing the short-term power spectral density of the source signals [4, 5, 6, 7, 8].

Under-determined source separation becomes even more challenging when the mixtures are reverberant. In that case the time-domain source signals are convolved with mixing filters before being added to produce a mixture. While this convolution is simply expressed in the time domain, it is not straightforward to take it into account when working with a TF representation of the mixture signals. Therefore, it is common to approximate the convolutive mixing process as being instantaneous in each frequency band of the short-time Fourier transform (STFT) [9, 10]. This approximation is considered to be valid when the mixing filters are short compared with the STFT analysis window. Source separation performance under this approximation is thus fundamentally limited when the mixture is highly reverberant. To overcome this limitation some methods have investigated more accurate TF mixture models. For example, time-domain convolution is exactly represented as a two-dimensional filtering in the TF domain in [11]. In [12] it is accurately approximated using a convolutive transfer function model.

Another approach introduced in [13] consists in modeling the sources in the TF domain while keeping a time-domain representation of the convolutive mixture. This is the approach we also followed in [14]. In this previous paper the source signals were characterized using the modified discrete cosine transform (MDCT) which is real-valued and critically sampled. In the present paper we generalize this method to a source representation based on the odd-frequency STFT (OFSTFT) which is redundant and complex-valued. This transform is similar to the STFT except that the discrete Fourier transform (DFT) is replaced by the odd-frequency DFT (OFDFT) [15]. Each source coefficient in this domain is modeled as a centered complex Gaussian random variable whose variance is structured by means of an NMF model. We infer the latent source variables using a variational expectation-maximization (VEM) algorithm. We experimentally study the performance of the method according to the choice of the source representation (MDCT or OFSTFT with different redundancy factors). We also show that the proposed approach outperforms two methods from the literature [16, 13] in a semi-blind setting where the mixing filters are known.

We start by presenting in Section 2 the OFDFT from which the OFSTFT can be constructed. In Section 3 we introduce the model. Section 4 details the VEM algorithm. The experimental evaluation is presented in Section 5 and we finally conclude in Section 6.

2. THE ODD-FREQUENCY DFT

The OFDFT of a signal $z(t)$, $t = 0, ..., T-1$, is defined for $f = 0, ..., T-1$ as $z_f = \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} z(t) \exp\left(-\imath \frac{2\pi}{T} \left(f + \frac{1}{2}\right) t\right)$ [15], where $\imath = \sqrt{-1}$. It is a particular case of the Generalized DFT [17]. Compared with the standard DFT, we see that it simply corresponds to shifting the frequency index by a factor $1/2$. For a real-valued signal $z(t)$, the following symmetry property holds: $z_{T-f-1} = z_f^*$ where $\cdot^*$ denotes complex conjugation. Compared with the standard DFT where the coefficients at the zero and the Nyquist's frequencies are real-valued, all coefficients of the OFDFT are complex-valued. It is thus more appropriate when the frequency coefficients are modeled as complex-valued random variables, which is common in audio source separation. Moreover

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).

this symmetry property allows us to write a simple expression for the inverse OFDFT involving only the non-redundant coefficients: $z(t) = \frac{2}{\sqrt{T}}\Re\left(\sum_{f=0}^{T/2-1} z_f \exp\left(\imath\frac{2\pi}{T}\left(f+\frac{1}{2}\right)t\right)\right)$, where $\Re(\cdot)$ denotes the real part. The OFSTFT can then be defined similarly as the standard STFT but using the OFDFT. We can mention that in [18] it has been shown that the MDCT and the OFSTFT are more appropriate than the standard STFT for assuming independent TF coefficients, which is common in audio source separation.

3. MODEL

The signal at each microphone $i = 1, ..., I$ is represented as a noisy mixture of J source images $y_{ij}(t)$, $j = 1, ..., J$, for $t = 0, ..., T-1$:

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

where $b_i(t) \sim \mathcal{N}_{\mathbb{R}}(0, \sigma_i^2)$ is a white Gaussian additive noise. The probability density function (pdf) of $\mathcal{N}_{\mathbb{R}}$ is defined in Appendix A.

Each source image $y_{ij}(t)$ corresponds to the discrete convolution of a source signal $s_j(t) \in \mathbb{R}$, $t = 0, ..., L_s - 1$, with a mixing filter $a_{ij}(t) \in \mathbb{R}$, $t = 0, ..., L_a - 1$, (such that $T = L_s + L_a - 1$):

$$y_{ij}(t) = [a_{ij} \star s_j](t). \tag{2}$$

Similarly as in [19], a source signal $s_j(t)$ is represented by a set of TF synthesis coefficients $\{s_{j,fn} \in \mathbb{K} = \mathbb{C} \text{ or } \mathbb{R}\}_{f,n}$ for $(f, n) \in \mathcal{B}$ with $\mathcal{B} = \{0, ..., F-1\} \times \{0, ..., N-1\}$:

$$s_j(t) = \frac{2}{\phi}\Re\left(\sum\nolimits_{(f,n)\in\mathcal{B}} s_{j,fn}\psi_{fn}(t)\right). \tag{3}$$

$\psi_{fn}(t) \in \mathbb{K}$, $t = 0, ..., L_s - 1$, is a TF synthesis atom and $\phi = 1$ if $\mathbb{K} = \mathbb{C}$ or $\phi = 2$ if $\mathbb{K} = \mathbb{R}$. In this work we consider either the MDCT [20] if $\mathbb{K} = \mathbb{R}$ or the OFSTFT if $\mathbb{K} = \mathbb{C}$. For the MDCT, the TF synthesis atom is defined as:

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

while for the OFSTFT we have:

$$\psi_{fn}(t) = \sqrt{\frac{1}{L_w}} w(t - nH) \exp\left(\imath\frac{2\pi}{L_w}\left(f + \frac{1}{2}\right)(t - nH)\right). \tag{5}$$

$w(t)$ is a sine window of even length L_w, therefore $F = L_w/2$. The hop size H equals $L_w/2$ when using the MDCT, while for the OFSTFT different values can be chosen so that perfect reconstruction is achieved. Note that compared with [19] we explicitly account for the fact that audio signals are real-valued. Indeed, in the OFSTFT case, $\{s_{j,fn}\}_{f=0}^{F-1}$ corresponds to the set of non-redundant coefficients for each time frame n (see Section 2).

From (2) and (3) a source image can be further written as:

$$y_{ij}(t) = \frac{2}{\phi}\Re\left(\sum\nolimits_{(f,n)\in\mathcal{B}} s_{j,fn} g_{ij,fn}(t)\right), \tag{6}$$

where $g_{ij,fn}(t) = [a_{ij} \star \psi_{fn}](t)$.

The synthesis coefficients $s_{j,fn}$ are then modeled as centered and real Gaussian random variables if $\mathbb{K} = \mathbb{R}$ or complex circularly symmetric Gaussian random variables if $\mathbb{K} = \mathbb{C}$:

$$s_{j,fn} \sim \begin{cases} \mathcal{N}_{\mathbb{R}}(0, v_{j,fn}) & \text{if } \mathbb{K} = \mathbb{R}; \\ \mathcal{N}_{\mathbb{C}}^{p}(0, v_{j,fn}) & \text{if } \mathbb{K} = \mathbb{C}. \end{cases} \tag{7}$$

The pdfs of these distributions are provided in Appendix A. The variances $v_{j,fn} \in \mathbb{R}_+$ are finally structured by means of an NMF model of rank K_j, generally chosen such that $K_j(F+N) \ll FN$:

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

with $\mathbf{W}_j \in \mathbb{R}_+^{F\times K_j}$, $\mathbf{H}_j \in \mathbb{R}_+^{K_j\times N}$.

4. VARIATIONAL INFERENCE

Let $\mathbf{x} = \{x_i(t)\}_{i,t}$ denote the set of observed variables, $\mathbf{s} = \{s_{j,fn}\}_{j,f,n}$ the latent variables and $\boldsymbol{\theta} = \{\{\sigma_i^2\}_i, \{\mathbf{W}_j, \mathbf{H}_j\}_j\}$ the model parameters. Remember that the mixing filters $\{a_{ij}(t)\}_{i,j,t}$ are assumed to be known. Exact posterior inference of the latent variables is here computationally heavy because the time-domain convolution induces complex posterior dependencies between the latent variables. We thus adopt a variational approach to infer the latent variables and estimate the model parameters. Let $q \in \mathcal{F}$ be a pdf over $\mathbf{s}$, where $\mathcal{F}$ is a variational family. Variational inference consists in optimizing a criterion called the variational free energy and defined as [21]:

$$\mathcal{L}(q; \boldsymbol{\theta}) = \langle \ln\left(p(\mathbf{x}, \mathbf{s}; \boldsymbol{\theta}) / q(\mathbf{s})\right)\rangle_q, \tag{9}$$

where $< \cdot >_q$ denotes the mathematical expectation taken with respect to q. More precisely we will use the VEM algorithm that consists in iterating two steps until convergence: the E-step where we compute $q^\star = \arg\max_{q\in\mathcal{F}} \mathcal{L}(q; \boldsymbol{\theta}^\star)$ and the M-step where we compute $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q^\star; \boldsymbol{\theta})$. In practice we will use the mean-field approximation by constraining the variational family $\mathcal{F}$ to the set of pdfs that factorize as $q(\mathbf{s}) = \prod_{j,f,n} q_{jfn}(s_{j,fn})$. Under this approximation we can show that the pdf over $s \in \mathbf{s}$ that maximizes the variational free energy satisfies [21]:

$$\ln q^\star(s) \stackrel{c}{=} \langle \ln p(\mathbf{x}, \mathbf{s}; \boldsymbol{\theta})\rangle_{q(\mathbf{s}\setminus s)}, \tag{10}$$

where $\stackrel{c}{=}$ represents equality up to an additive constant and $\mathbf{s}\setminus s$ denotes the set of all latent variables but s.

Source estimate. Under the variational mean-field approximation, the estimate of the j-th source in the TF domain is given by:

$$\hat{s}_{j,fn} = \langle s_{j,fn}\rangle_q. \tag{11}$$

The time-domain signal $\hat{s}_j(t)$ is then reconstructed by inverse TF transform and the source image $\hat{y}_{ij}(t)$ is obtained by convolution with the corresponding mixing filter: $\hat{y}_{ij}(t) = [a_{ij} \star \hat{s}_j](t)$.

Complete-data log-likelihood. From the model introduced in Section 3, the complete-data log-likelihood $\ln p(\mathbf{x}, \mathbf{s}; \boldsymbol{\theta}) = \ln p(\mathbf{x}|\mathbf{s}; \boldsymbol{\theta}) + \ln p(\mathbf{s}; \boldsymbol{\theta})$ can be expressed as:

$$\ln p(\mathbf{x}, \mathbf{s}; \boldsymbol{\theta}) \stackrel{c}{=} -\frac{1}{2}\sum_{i=1}^{I}\sum_{t=0}^{T-1}\left[\ln(\sigma_i^2) + \frac{1}{\sigma_i^2}\left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t)\right)^2\right] - \frac{1}{\phi}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\left[\ln(v_{j,fn}) + \frac{|s_{j,fn}|^2}{v_{j,fn}}\right]. \tag{12}$$

E-step. From (10) and (12) we can show that:

$$q_{jfn}^\star(s_{j,fn}) = \begin{cases} N_{\mathbb{C}}(\rho_{j,fn}, \hat{s}_{j,fn}^r, \hat{s}_{j,fn}^\imath, \gamma_{j,fn}^r, \gamma_{j,fn}^\imath) & \text{if } \mathbb{K} = \mathbb{C}; \\ N_{\mathbb{R}}(\hat{s}_{j,fn}^r, \gamma_{j,fn}^r) & \text{if } \mathbb{K} = \mathbb{R}, \end{cases} \tag{13}$$

where these pdfs are defined in Appendix A and for $\mathfrak{p} \in \{r, \imath\}$ we have:

$$\rho_{j,fn}=\left(\frac{2}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\Im(g_{ij,fn}(t))\right)\Bigg/ \left[\left(\frac{2}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))^2+\frac{1}{\phi v_{j,fn}}\right)\times \left(\frac{2}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))^2+\frac{1}{\phi v_{j,fn}}\right)\right]^{0.5}; \quad (14)$$

$$\gamma_{j,fn}^{\mathfrak{p}}=\left[2(1-\rho_{j,fn}^2)\left(\frac{2}{\phi^2}\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\mathfrak{P}(g_{ij,fn}(t))^2+\frac{1}{\phi v_{j,fn}}\right)\right]^{-1}; \quad (15)$$

$$\hat{s}_{j,fn}^{\mathfrak{p}} = \hat{s}_{j,fn}^{\mathfrak{p}} - \gamma_{j,fn}^{\mathfrak{p}}(1-\rho_{j,fn}^2)d_{j,fn}^{\mathfrak{p}}, \quad (16)$$

with $\mathfrak{P}(\cdot)$ denoting the real part $\Re(\cdot)$ (resp. the imaginary part $\Im(\cdot)$) if $\mathfrak{p} = r$ (resp. $\imath$) and

$$d_{j,fn}^{r}=\frac{2}{\phi}\left[\frac{\hat{s}_{j,fn}^{r}}{v_{j,fn}}-\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Re(g_{ij,fn}(t))\left(x_i(t)-\sum_{j'=1}^{J}\hat{y}_{ij'}(t)\right)\right]; \quad (17)$$

$$d_{j,fn}^{\imath}=\frac{2}{\phi}\left[\frac{\hat{s}_{j,fn}^{\imath}}{v_{j,fn}}+\sum_{i=1}^{I}\frac{1}{\sigma_i^2}\sum_{t=0}^{T-1}\Im(g_{ij,fn}(t))\left(x_i(t)-\sum_{j'=1}^{J}\hat{y}_{ij'}(t)\right)\right]. \quad (18)$$

Interestingly, in the complex case ($\mathbb{K} = \mathbb{C}$) $q_{jfn}^{\star}(s_{j,fn})$ is the pdf of a complex Gaussian distribution which is not proper (see Appendix A). It means that the real and imaginary parts of the source coefficients are a posteriori correlated and have different variances. In the real case ($\mathbb{K} = \mathbb{R}$) we obtain the same results as presented in [14]. We have to mention that update (16) for $\mathfrak{p} \in \{r, \imath\}$ holds if the parameters are updated in turn. However we can show that $d_{j,fn}^{\mathfrak{p}} = \partial(-\mathcal{L}(q^\star;\boldsymbol{\theta}))/(\partial\hat{s}_{j,fn}^{\mathfrak{p}})$ where $\mathcal{L}(q^\star;\boldsymbol{\theta})$ is given in the next paragraph. Therefore, (16) corresponds to a coordinate ascent of the variational free energy. In practice, as in [14], we will rather use the conjugate gradient method with diagonal preconditioning [22] for optimizing this criterion with respect to the whole set of coefficients $\{\hat{s}_{j,fn}^{r}, \hat{s}_{j,fn}^{\imath}\}$. This choice allows us to make the E-Step more computationally efficient. Further details on the derivation of the E-step and on this conjugate gradient algorithm can be found in the supporting document [23]. The source estimate is finally given by $\hat{s}_{j,fn} = \hat{s}_{j,fn}^{r} + \imath\hat{s}_{j,fn}^{\imath}$. We also define the following second-order moments that will be used in the sequel:

- ▷ Variance: $\gamma_{j,fn} = \langle|s_{j,fn} - \hat{s}_{j,fn}|^2\rangle_q = \gamma_{j,fn}^{r} + \gamma_{j,fn}^{\imath}$;
- ▷ Pseudo-variance: $\tilde{\gamma}_{j,fn}=\langle(s_{j,fn}-\hat{s}_{j,fn})^2\rangle_q$ $=\gamma_{j,fn}^{r}-\gamma_{j,fn}^{\imath}+2\imath\rho_{j,fn}\sqrt{\gamma_{j,fn}^{r}\gamma_{j,fn}^{\imath}}$.

Variational free energy. Omitting the terms that are independent of the model parameters, the variational free energy can be written from (9), (12) and the E-step as follows:

$$\mathcal{L}(q^\star;\boldsymbol{\theta}) \stackrel{c}{=} -\frac{1}{2}\sum_{i=1}^{I}\sum_{t=0}^{T-1}\left[\ln(\sigma_i^2)+\frac{e_i(t)}{\sigma_i^2}\right] - \frac{1}{\phi}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\left[\ln\left(v_{j,fn}\right)+\frac{|\hat{s}_{j,fn}|^2+\gamma_{j,fn}}{v_{j,fn}}\right], \quad (19)$$

where $e_i(t) = \left\langle\left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t)\right)^2\right\rangle_{q^\star}$ can be further expressed from the mean-field approximation and (6) as:

$$e_i(t) = \left(x_i(t) - \sum_{j=1}^{J}\hat{y}_{ij}(t)\right)^2 + \frac{2}{\phi^2}\sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\left[\Re\left(\tilde{\gamma}_{j,fn}g_{ij,fn}^2(t)\right)+\gamma_{j,fn}|g_{ij,fn}(t)|^2\right]. \quad (20)$$

M-step. The M-step consists in maximizing (or only increasing) $\mathcal{L}(q^\star;\boldsymbol{\theta})$ in (19) with respect to $\boldsymbol{\theta}$. Zeroing the derivative of this criterion with respect to σ_i^2 leads to the following update:

$$\sigma_i^2 = \frac{1}{T}\sum_{t=0}^{T-1} e_i(t). \quad (21)$$

For the NMF parameters, we can recognize in (19) the Itakura-Saito divergence [4] between the posterior mean of the source power spectrogram $\langle|s_{j,fn}|^2\rangle_{q^\star} = |\hat{s}_{j,fn}|^2 + \gamma_{j,fn}$ and $v_{j,fn} = [\mathbf{W}_j\mathbf{H}_j]_{fn}$ (up to an additive constant). Therefore the NMF parameters can be updated using the multiplicative update rules given in [4].

5. EXPERIMENTS

We perform the experiments using audio source signals provided by the MTG MASS database [24]. We created 8 stereo mixtures sampled at 16 kHz using room impulse responses from the RWCP (JR2) database [25]. These room responses were recorded in a real room with a reverberation time of 470 ms. Each mixture contains between 3 and 5 spatially disjoint sources and the duration ranges from 12 to 28 seconds.

All experiments are performed with the true mixing filters known and fixed. We evaluate the quality of the separation in terms of reconstructed mono sources. We use standard energy ratios defined in [26] and expressed in decibels (dB): the Signal-to-Distortion (SDR), Artifact (SAR) and Interference (SIR) Ratios. These measures are computed using the BSS Eval toolbox [27]. We also consider perceptually motivated objective measures introduced in [28, 29]: the Overall (OPS), Target-related (TPS), Interference-related (IPS) and Artifact-related (APS) Perceptual Scores. They are expressed in percentage and computed using the PEASS toolbox [30]. For all experiments we use an analysis/synthesis sine window of 128 ms. For the methods relying on NMF, the factorization rank is arbitrarily fixed to 10 for all sources.

We refer the reader to our web page for listening to audio examples illustrating the results discussed below[1]. Matlab code implementing the proposed method is also available.

We first evaluate the separation according to the TF source representation used in the proposed framework. More precisely, we study the influence of the redundancy of the transform. The MDCT is critically sampled which means that there are as many TF coefficients as time-domain samples. On the contrary, the OFSTFT is a redundant transform, the redundancy being controlled by the overlap size. For example, a 50% overlap leads to a number of (real) TF coefficients which is twice the number of time-domain samples. The average separation results for the different transforms

[1] https://perso.telecom-paristech.fr/leglaive/demo-waspaa17.html

are shown in Table 1 (lines 2 to 5), higher overlap means higher redundancy. The VEM algorithm was run for 200 iterations. We first observe that according to the SDR, SIR and SAR, the more redundancy we use, the better the results are. However this improvement is not so clear by listening to the separated sources. Therefore we also computed perceptually motivated objective measures. As can be seen from Table 1, the overall separation quality as measured by the OPS is much less dependent on the TF representation. We even obtain the best performance with the MDCT which is critically sampled. These results seem to be more consistent with the perceived separation quality. Moreover we have to mention that by increasing the redundancy we increase the number of latent TF source variables, so the separation is computationally more expensive. Finally we can see that according to the IPS and APS, increasing the redundancy seems to help reducing interferences to some extent, but it induces more artifacts. The overlap-add of multiple incoherent short-term estimates of the source signals at the synthesis stage may explain this phenomenon.

We also compare our approach with two methods from the literature. The first one was introduced in [16]. It also relies on a local Gaussian source model based on NMF[2] but the convolutive mixing process is approximated as being instantaneous in each frequency band of the STFT. Inference is performed with an EM algorithm that was run for 200 iterations in this experiment. We see from Table 1, line 6, that this method obtains much lower scores than the proposed one. This is due to the fact that reverberation is not accurately represented by the approximate mixture model in the STFT domain. This experiment thus demonstrates the usefulness of representing the convolutive mixture in the time domain. The second method we consider for this evaluation was introduced in [13]. It also relies on exact time-domain modeling of the convolutive mixing process but it uses a sparse source model based on ℓ_1 regularization of the STFT source coefficients. This approach results in a Lasso problem that is solved with the FISTA algorithm. As proposed in the paper by the authors we run this algorithm for 20000 iterations. Source separation results with this method are given in the last line of Table 1. According to the SDR, SIR and SAR, the proposed method performs better only with the OFSTFT and an overlap of 50 or 75%. Nevertheless the perceptually motivated measures show that the source separation quality is also improved with less redundant representations, even with the critically sampled MDCT. This is confirmed when listening to the separated sources. Comparing our approach with this method shows that not only the exact convolutive mixture modeling is important but also the NMF-based source model. We can also mention that evaluating the separation quality in terms of reconstructed stereo source images instead of mono sources leads to the same conclusions.

To conclude this experimental evaluation we give some indications on the computational complexity of the different methods for one of the mixture that contains 3 sources and lasts for 12 seconds. We present the computational time normalized by the one obtained with the proposed method when using the MDCT. The results are given in the last column of Table 1. As expected, the more redundant the TF transform, the higher the computational time. Ozerov and Févotte's method [16] is clearly the fastest one because it does not rely on time-domain convolutive mixture modeling. The computational time for the method by Kowalski et al. [13] is similar to the one obtained with the proposed method when using the MDCT.

[2]As proposed later by the authors in [31], the NMF parameters are updated differently from [16] by using multiplicative update rules.

	SDR	SIR	SAR	OPS	TPS	IPS	APS	NCT
MDCT [14]	4.8	10.9	8.2	**38.9**	**66.5**	65.7	**38.2**	1.00
OFSTFT - overlap 25%	6.5	12.9	9.5	38.7	63.4	67.7	35.7	2.83
OFSTFT - overlap 50%	7.6	14.9	10.4	37.9	64.6	**68.6**	35.1	4.07
OFSTFT - overlap 75%	**9.7**	**17.9**	**12.1**	36.4	62.8	67.4	34.1	7.79
Ozerov and Févotte [16]	-2.4	4.3	2.1	22.5	45.3	63.9	9.1	0.01
Kowalski et al. [13]	7.5	14.1	10.1	29.1	63.2	60.7	23.1	1.13

Table 1: Source separation results averaged over all the sources in the dataset and normalized computational time (NCT) for one of the mixtures containing 3 sources and lasting for 12 seconds.

6. CONCLUSION

In this paper we generalized our previous work [14] to the use of a source representation based on the OFSTFT. We experimentally studied the impact of using this complex-valued and redundant TF transform on the source separation performance compared with the use of the MDCT. We also showed that the proposed approach outperforms two standard methods from the literature in a semi-blind setting where the mixing filters are known. This experimental evaluation demonstrated the importance of jointly modeling the convolutive mixing process in the time domain and the source signals in the TF domain by means of an NMF model.

Future work will focus on developing a fully blind source separation method where the mixing filters will also be estimated. Using probabilistic priors on the mixing filters could help us to reach this objective [32]. Indeed, the mixing filters are room responses so they exhibit a simple specific structure in the time domain that could be used to guide their estimation.

A. GAUSSIAN PROBABILITY DISTRIBUTIONS

Let $\mathcal{N}_{\mathbb{R}}(x;\mu,\sigma^2)$ denote the Gaussian distribution over a real-valued random variable (r.v.) x. Its pdf is given by:

$$N_{\mathbb{R}}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{22}$$

Let $\mathcal{N}_{\mathbb{C}}(x;\rho,\mu_{x_r},\mu_{x_i},\sigma^2_{x_r},\sigma^2_{x_i})$ denote the Gaussian distribution over a complex-valued r.v. $x = x_r + \imath x_\imath$. Its pdf is given by [33]:

$$\begin{aligned} N_{\mathbb{C}}(x;\rho,\mu_{x_r},\mu_{x_\imath},\sigma^2_{x_r},\sigma^2_{x_\imath}) &= \frac{1}{2\pi\sigma_{x_r}\sigma_{x_\imath}\sqrt{1-\rho^2}} \\ &\times \exp\Bigg[-\frac{1}{2(1-\rho^2)}\Bigg(\frac{(x_r-\mu_{x_r})^2}{\sigma^2_{x_r}} + \frac{(x_\imath-\mu_{x_\imath})^2}{\sigma^2_{x_\imath}} \\ &\qquad - \frac{2\rho(x_r-\mu_{x_r})(x_\imath-\mu_{x_\imath})}{\sigma_{x_r}\sigma_{x_\imath}}\Bigg)\Bigg], \end{aligned} \tag{23}$$

where $\rho = \mathbb{E}[(x_r-\mu_{x_r})(x_\imath-\mu_{x_\imath})]/(\sigma_{x_r}\sigma_{x_\imath}) \in [-1,1]$. The particular case $\mathcal{N}_{\mathbb{C}}(x;0,\mu_{x_r},\mu_{x_\imath},\sigma^2/2,\sigma^2/2)$ corresponds to the proper complex Gaussian distribution. It is denoted by $\mathcal{N}^p_{\mathbb{C}}(x;\mu,\sigma^2)$ where $\mu = \mu_{x_r} + \imath\mu_{x_\imath}$ and $\sigma^2 = 2\sigma^2_{x_r} = 2\sigma^2_{x_\imath}$. In this case the pdf gets simplified to:

$$N^p_{\mathbb{C}}(x;\mu,\sigma^2) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|x-\mu|^2}{\sigma^2}\right). \tag{24}$$

Finally, the complex Gaussian distribution is circularly symmetric if it is proper and $\mu = 0$.

B. REFERENCES

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

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

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

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

[5] T. Virtanen, A. T. Cemgil, and S. 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.

[6] A. Liutkus, D. Fitzgerald, and R. 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.

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

[8] K. Yoshii, K. Itoyama, and M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 51–55.

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

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

[11] R. Badeau and M. 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.

[12] X. Li, L. Girin, and R. Horaud, "Audio source separation based on convolutive transfer function and frequency-domain lasso optimization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017.

[13] M. Kowalski, E. Vincent, and R. 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.

[14] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation: Variational inference of time-frequency sources from time-domain observations," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 26–30.

[15] J. L. Vernet, "Real signals fast Fourier transform: Storage capacity and step number reduction by means of an odd discrete Fourier transform," *Proc. of the IEEE*, vol. 59, no. 10, pp. 1531–1532, 1971.

[16] 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.

[17] G. Bongiovanni, P. Corsini, and G. Frosini, "One-dimensional and two-dimensional generalised discrete Fourier transforms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 1, pp. 97–99, 1976.

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

[19] C. Févotte and M. Kowalski, "Low-rank time-frequency synthesis," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 3563–3571.

[20] H. S. Malvar, *Signal Processing with Lapped Transforms*. Artech House, 1992.

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

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

[23] "Supporting document," http://perso.telecom-paristech.fr/leglaive/documents/supportingDocumentWaspaa2017.pdf.

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

[25] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition." in *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, 2000, pp. 965–968.

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

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

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

[29] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*, 2012, pp. 430–437.

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

[31] 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. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 257–260.

[32] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation with probabilistic reverberation priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2453–2465, 2016.

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