Semi-Blind Student's *t* Source Separation for Multichannel Audio Convolutive Mixtures

Simon Leglaive, Roland Badeau, Gaël Richard

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

***Abstract*—This paper addresses the problem of multichannel audio source separation in under-determined convolutive mixtures. We target a semi-blind scenario assuming that the mixing filters are known. The convolutive mixing process is exactly modeled using the time-domain impulse responses of the mixing filters. We propose a Student's *t* time-frequency source model based on non-negative matrix factorization (NMF). The Student's *t* distribution being heavy-tailed with respect to the Gaussian, it provides some flexibility in the modeling of the sources. We also study a simpler Student's *t* sparse source model within the same general source separation framework. The inference procedure relies on a variational expectation-maximization algorithm. Experiments show the advantage of using an NMF model compared with the sparse source model. While the Student's *t* NMF source model leads to slightly better results than our previous Gaussian one, we demonstrate the superiority of our method over two other approaches from the literature.**

***Index Terms*—Under-determined audio source separation, multichannel convolutive mixture, Student's *t* distribution, non-negative matrix factorization, variational inference.**

I. Introduction

Multichannel audio source separation aims to recover a set of audio source signals from several observed mixtures. We consider an under-determined scenario where the number of sources is greater than the number of microphones. Moreover we focus on modeling reverberant (or convolutive) mixtures. This problem involves two modeling steps: modeling the source signals and the way they are mixed together.

Source modeling is commonly achieved in a time-frequency (TF) domain because it provides a meaningful and sparse representation of the source signals. Sparse component analysis [1] and variance modeling frameworks [2] are two important trends in audio source separation [3]. Within the variance modeling framework, non-negative matrix factorization (NMF) techniques are popular to represent the spectro-temporal characteristics of the sources [4], [5], [6], [7], [8].

Convolutive mixtures are frequently approximated as being instantaneous in the short-time Fourier transform (STFT) domain [9], [10] or modeled by means of a spatial covariance matrix [11], [12]. A less common approximation relies on a subband filtering model [13], which has recently demonstrated its potential for reverberant audio source separation [14], [15].

In this paper we focus on exact time-domain convolutive mixture modeling. We target a semi-blind scenario assuming that the mixing filters are known. Other approaches relying on the same exact convolutive modeling were focusing on sparse TF source models [16], [17]. Note that these methods were also considering a semi-blind setting. Here we propose to incorporate within a same framework a sparse and an NMF-based source model relying on the Student's *t* distribution. A multichannel Student's *t* source separation method has recently been proposed in [18], but only the likelihood was Student's *t*, not the source model. Moreover the convolutive mixing process was approximated in the STFT domain.

Building on our previous Gaussian model [19], the source coefficients in the Modified Discrete Cosine Transform (MDCT) domain are modeled as centered Student's *t* random variables. For one given source, the scale parameters of these random variables are either considered to be equal, leading to a sparsity based model, or TF-dependent but structured by means of an NMF model. We then use the time-domain observations to infer the TF latent source variables. The inference relies on a variational expectation-maximization (VEM) algorithm [20]. Experiments demonstrate the superiority of the Student's *t* NMF model over the sparse one and other state-of-the-art methods.

The models are introduced in section II. The variational inference is presented in section III. Experiments are conducted in section IV and we finally draw conclusions in section V.

II. 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 us 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 a white Gaussian additive noise:

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

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

$$s_j(t) = \sum_{(f,n)\in\mathcal{B}} s_{j,fn}\psi_{fn}(t). \tag{3}$$

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).

ISBN 978-0-9928626-7-1 © EURASIP 2017

We define the source model in the MDCT domain such that the synthesis atom $\psi_{fn}(t) \in \mathbb{R}$, $t = 0, ..., L_s - 1$, is given by:

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

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, $H = L_w/2$ is the hop size and $F = L_w/2$ is the number of frequency bins. We choose the MDCT because it is critically sampled; it involves less TF coefficients than the STFT which is redundant. Using the MDCT thus allows us to limit the number of source TF coefficients to be estimated, and consequently it also limits the computational load. With this TF representation, a source image can be further written as:

$$y_{ij}(t) = [a_{ij} \star s_j](t) = \sum_{(f,n)\in\mathcal{B}} s_{j,fn}\, g_{ij,fn}(t), \tag{5}$$

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

We assume that the TF synthesis coefficients $s_{j,fn}$ independently follow a Student's t distribution (see Appendix A):

$$s_{j,fn} \sim \mathcal{T}_\alpha(0, \lambda_{j,fn}). \tag{6}$$

α and $\lambda_{j,fn}$ are respectively the shape parameter (also called degrees of freedom) and the scale parameter. This distribution approaches the Gaussian one as α goes to infinity. Smaller values of α yield heavier tails. The particular case $\alpha = 1$ leads to the Cauchy distribution.

Model (6) can be shown to be equivalent to the following hierarchical one:

$$\begin{cases} s_{j,fn}|v_{j,fn} & \sim \mathcal{N}(0, v_{j,fn}); \\ v_{j,fn} & \sim \mathcal{IG}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\lambda_{j,fn}^2\right), \end{cases} \tag{7}$$

where $\mathcal{N}$ denotes the Gaussian distribution and $\mathcal{IG}$ the Inverse-Gamma distribution (see Appendix A). In the following we will consider two particular cases of this source model:

1) **Sparse source model**: The TF synthesis coefficients for one given source are assumed to be identically distributed, i.e. for all f, n:

$$\lambda_{j,fn}^2 = \lambda_j^2. \tag{8}$$

When the shape parameter α is set to a sufficiently small value, this model assumes that the sources are sparse in the MDCT domain. This sparse Student's t source model has already been used for source separation in [21] but for instantaneous and not for convolutive mixtures.

2) **NMF source model**: The squared scale parameters $\lambda_{j,fn}^2$ are structured by means of an NMF model:

$$\lambda_{j,fn}^2 = [\mathbf{W}_j\mathbf{H}_j]_{fn}, \tag{9}$$

with $\mathbf{W}_j = [w_{j,fk}]_{f,k} \in \mathbb{R}_+^{F\times K_j}$ and $\mathbf{H}_j = [h_{j,kn}]_{k,n} \in \mathbb{R}_+^{K_j\times N}$. K_j is the rank of the factorization. Note that a similar but not equivalent source model has recently been proposed in [22]. In this paper a source was modeled by a sum of Gaussian components in the STFT domain. The TF-dependent variances of one component were assumed to follow inverse-gamma priors, whose scale parameters were constrained by a rank-1 NMF model. Here the j-th source is Student's t in the MDCT domain with a rank-K_j NMF model on the scale parameters.

III. Variational Inference

Let $\mathbf{x} = \{x_i(t)\}_{i,t}$ denote the set of observed variables, $\mathbf{z} = \{\mathbf{s} = \{s_{j,fn}\}_{j,f,n}, \mathbf{v} = \{v_{j,fn}\}_{j,f,n}\}$ the set of latent variables and $\boldsymbol{\theta} = \{\boldsymbol{\sigma} = \{\sigma_i^2\}_i, \boldsymbol{\lambda} = \{\lambda_{j,fn}^2\}_{j,f,n}\}$ the parameters to be estimated. We recall that the mixing filters $\{a_{ij}(t)\}_{i,j,t}$ are assumed to be known. We would like to infer the latent variables according to their posterior mean, using a maximum likelihood estimation of the model parameters. However, exact posterior inference with the proposed model is analytically intractable. We thus resort to variational inference. Let $\mathcal{F}$ be a set of probability density functions (pdfs) over the latent variables $\mathbf{z}$. For any pdf $q \in \mathcal{F}$ and any function $f(\mathbf{z})$, we note $\langle f(\mathbf{z})\rangle_q = \int f(\mathbf{z})q(\mathbf{z})d\mathbf{z}$. Then for any $q \in \mathcal{F}$ and parameter set $\boldsymbol{\theta}$, the log-likelihood can be decomposed as $\ln p(\mathbf{x};\boldsymbol{\theta}) = \mathcal{L}(q;\boldsymbol{\theta}) + D_{\text{KL}}(q||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}))$, where $\mathcal{L}(q;\boldsymbol{\theta}) = \langle \ln\left(p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})/q(\mathbf{z})\right)\rangle_q$ and $D_{\text{KL}}(q||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) = \langle \ln\left(q(\mathbf{z})/p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})\right)\rangle_q$ is the Kullback-Leibler (KL) divergence between q and the posterior distribution $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$. $\mathcal{L}(q;\boldsymbol{\theta})$ is called the variational free energy and can be further decomposed as $\mathcal{L}(q;\boldsymbol{\theta}) = E(q;\boldsymbol{\theta}) + H(q)$ where

$$E(q;\boldsymbol{\theta}) = \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_q, \tag{10}$$

and $H(q) = -\langle \ln q(\mathbf{z})\rangle_q$ is the entropy of the variational distribution q. Since the KL divergence is always non-negative, the variational free energy is a lower bound of the log-likelihood. The VEM algorithm [20] 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 assuming that the pdf q can be factorized as:

$$q(\mathbf{z}) = \prod_{j=1}^{J}\prod_{(f,n)\in\mathcal{B}} q_{jfn}^s(s_{j,fn})q_{jfn}^v(v_{j,fn}). \tag{11}$$

For simplicity of notations we will drop the superscript and the indices when referring to the distributions $q_{jfn}^{(\cdot)}$. Under the mean-field approximation it can be shown that the pdf over a latent variable $z \in \mathbf{z}$ which maximizes the variational free energy satisfies [20]:

$$\ln q^\star(z) \stackrel{c}{=} \langle \ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\rangle_{q(\mathbf{z}\setminus z)}, \tag{12}$$

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

A. Source Estimate Under the Variational Approximation

Under the mean-field approximation, the j-th TF source estimate is given by:

$$\hat{s}_{j,fn} = \langle s_{j,fn}\rangle_q. \tag{13}$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

Algorithm 1: E-step

1: $\delta = \dfrac{\alpha+1}{2}$
2: **for all** j, f, n **do**
3: $\quad \beta_{j,fn} = (\alpha/2)\lambda_{j,fn}^2 + (\hat{s}_{j,fn}^2 + \gamma_{j,fn})/2$
4: $\quad \gamma_{j,fn} = \left[\sum_{i=1}^{I} \dfrac{1}{\sigma_i^2} \sum_{t=0}^{T-1} g_{ij,fn}(t)^2 + \dfrac{\delta}{\beta_{j,fn}}\right]^{-1}$
5: $\quad d_{j,fn} = \hat{s}_{j,fn}\delta/\beta_{j,fn} - \sum_{i=1}^{I} \dfrac{1}{\sigma_i^2} \sum_{t=0}^{T-1} g_{ij,fn}(t)\left(x_i(t) - \sum_{j'=1}^{J} \hat{y}_{ij'}(t)\right)$
6: $\quad \hat{s}_{j,fn} \leftarrow \hat{s}_{j,fn} - \gamma_{j,fn} d_{j,fn}$

The time-domain source estimate $\hat{s}_j(t)$ is then reconstructed by inverse MDCT and the source image $\hat{y}_{ij}(t)$ is obtained by convolution of $\hat{s}_j(t)$ with the associated mixing filter $a_{ij}(t)$.

B. Complete-data Log-likelihood

From (1), (2) and (7), the complete data log-likelihood $\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \ln p(\mathbf{x}|\mathbf{z}; \boldsymbol{\sigma}) + \ln p(\mathbf{s}|\mathbf{v}) + \ln p(\mathbf{v}; \boldsymbol{\lambda})$ writes:

$$\begin{aligned}\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \stackrel{c}{=} &-\frac{1}{2}\sum_{i=1}^{I}\sum_{t=0}^{T-1}\left[\ln(\sigma_i^2) + \frac{1}{\sigma_i^2}\left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t)\right)^2\right] \\ &- \sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\left[\ln(v_{j,fn})\left(\frac{\alpha}{2}+1+\frac{1}{2}\right) + \ln\Gamma\left(\frac{\alpha}{2}\right)\right. \\ &\left. + \frac{1}{v_{j,fn}}\left(\frac{\alpha}{2}\lambda_{j,fn}^2 + \frac{s_{j,fn}^2}{2}\right) + \frac{\alpha}{2}\ln\left(\frac{2}{\alpha\lambda_{j,fn}^2}\right)\right],\end{aligned} \tag{14}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

C. E-Step

From (11), (12) and (14) we can identify the variational distribution that maximizes the variational free energy: $q^\star(v_{j,fn}) = IG(\delta, \beta_{j,fn})$ and $q^\star(s_{j,fn}) = N(\hat{s}_{j,fn}, \gamma_{j,fn})$, where IG and N are the pdfs of the Inverse-Gamma and Gaussian distributions respectively, defined in Appendix A. The E-Step consists in updating the parameters of these distributions as given in Algorithm 1. It can be shown that $d_{j,fn} = \partial(-\mathcal{L}(q^\star; \boldsymbol{\theta}))/(\partial \hat{s}_{j,fn})$ where $\mathcal{L}(q^\star; \boldsymbol{\theta})$ is given in the next section. The updates in lines 5 and 6 of Algorithm 1 thus correspond to a coordinate-wise minimization of the negative variational free energy. For the sake of computational efficiency we will rather use the preconditioned conjugate gradient (PCG) method for updating all the parameters $\{\hat{s}_{j,fn}\}_{j,f,n}$ at once. The PCG method being very similar to the one described in [19], it is not detailed here.

D. Variational Free Energy

Using the results from the E-step, we can compute the variational free energy $\mathcal{L}(q^\star; \boldsymbol{\theta}) = E(q^\star; \boldsymbol{\theta}) + H(q^\star)$. From

Algorithm 2: M-step

1: **for all** i, j **do**
2: $\quad$ **switch** *source model* **do**
3: $\quad\quad$ **case** *sparse*
4: $\quad\quad\quad \lambda_j^2 = \left(\dfrac{1}{FN}\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left(\dfrac{\beta_{j,fn}}{\delta}\right)^{-1}\right)^{-1}$
5: $\quad\quad$ **case** *NMF*
6: $\quad\quad\quad \mathbf{W}_j \leftarrow \mathbf{W}_j \odot \dfrac{[\mathbf{W}_j\mathbf{H}_j]^{\odot -1}\,\mathbf{H}_j^T}{(\mathbf{B}_j/\delta)^{\odot -1}\,\mathbf{H}_j^T}$
7: $\quad\quad\quad \mathbf{H}_j \leftarrow \mathbf{H}_j \odot \dfrac{\mathbf{W}_j^T\,[\mathbf{W}_j\mathbf{H}_j]^{\odot -1}}{\mathbf{W}_j^T\,(\mathbf{B}_j/\delta)^{\odot -1}}$
8: $\quad \sigma_i^2 = \frac{1}{T}\sum_{t=0}^{T-1} e_i(t)$, with $e_i(t)$ defined in (16)

equations (10) and (14), the energy term writes:

$$\begin{aligned}E(q^\star; \boldsymbol{\theta}) \stackrel{c}{=} &-\frac{1}{2}\sum_{i=1}^{I}\sum_{t=0}^{T-1}\left[\ln(\sigma_i^2) + \frac{e_i(t)}{\sigma_i^2}\right] \\ &- \sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\left[(\ln(\beta_{j,fn}) - \mathrm{di}\Gamma(\delta))\left(\frac{\alpha}{2}+1+\frac{1}{2}\right) + \ln\Gamma\left(\frac{\alpha}{2}\right)\right. \\ &\left. + \frac{\delta}{\beta_{j,fn}}\left(\frac{\alpha}{2}\lambda_{j,fn}^2 + \frac{\hat{s}_{j,fn}^2 + \gamma_{j,fn}}{2}\right) + \frac{\alpha}{2}\ln\left(\frac{2}{\alpha\lambda_{j,fn}^2}\right)\right],\end{aligned} \tag{15}$$

where $\mathrm{di}\Gamma(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function and $e_i(t) = \left\langle \left(x_i(t) - \sum_{j=1}^{J} y_{ij}(t)\right)^2 \right\rangle_{q^\star}$ can be developed from (5) and (11) as follows:

$$e_i(t) = \left(x_i(t) - \sum_{j=1}^{J}\hat{y}_{ij}(t)\right)^2 + \sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}} \gamma_{j,fn}\, g_{ij,fn}^2(t). \tag{16}$$

From the mean-field approximation (11), the entropy of the variational distribution writes:

$$\begin{aligned}H(q^\star) \stackrel{c}{=} &\; JFN\left[\delta - (1+\delta)\,\mathrm{di}\Gamma(\delta) + \ln\left(\Gamma\left(\delta\right)\right)\right] \\ &+ \sum_{j=1}^{J}\sum_{(f,n)\in\mathcal{B}}\left[\frac{1}{2}\ln(\gamma_{j,fn}) + \ln(\beta_{j,fn})\right].\end{aligned} \tag{17}$$

E. M-step

The M-step aims to maximize (or only increase) the energy $E(q^\star; \boldsymbol{\theta})$ in equation (15) with respect to the parameter set $\boldsymbol{\theta}$. The resulting updates are given in Algorithm 2. The updates for λ_j^2 and σ_i^2 are obtained by zeroing the derivative of $E(q^\star; \boldsymbol{\theta})$. Interestingly, for the sparse source model, λ_j^2 is updated according to the harmonic mean of $\{\beta_{j,fn}/\delta\}_{f,n}$, thus providing a kind of robustness to outliers. The multiplicative NMF update rules in lines 6 and 7 (where $\mathbf{B}_j = [\beta_{j,fn}]_{fn} \in \mathbb{R}_+^{F\times N}$) are obtained by using a majorization-minimization approach (see Appendix B for details). Note that they can be repeated several times within the M-Step. Moreover, as explained in Appendix B, these updates differ from what one can usually find in the NMF literature.

ISBN 978-0-9928626-7-1 © EURASIP 2017

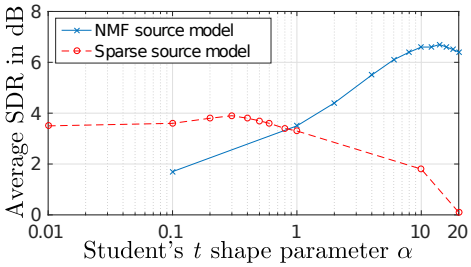

Fig. 1. Average SDR in dB as a function of the shape parameter α.

IV. Experiments

Our experiments are conducted from audio tracks provided by the Musical Audio Signal Separation (MASS) dataset [23]. We consider 8 stereo mixtures sampled at 16 kHz and obtained by simulating mixing filters with the Roomsimove toolbox [24]. The mixtures duration ranges from 12 to 28 seconds. The reverberation time[1] is set to 256 ms. 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 are spatially disjoint and their directions of arrival range from -45° to 45°. We evaluate the source separation performance in terms of reconstructed monophonic sources. We use standard energy ratios: the Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR) and Signal-to-Artifact Ratio (SAR). These criteria expressed in decibels (dB) are defined in [25]. We used the BSS Eval Toolbox available at [26] to compute these measures. For all the methods compared in this section, we used a half-overlapping TF analysis/synthesis sine window of 128 ms. For the NMF-based methods, the factorization rank was arbitrarily fixed to $K_j = 10$ for every source. For all experiments the source parameters are blindly initialized while the mixing filters are fixed to the true ones.

We start by comparing the Student's t sparse and NMF-based source models. Fig. 1 represents the SDR averaged over all the sources in the dataset as a function of the shape parameter α for the two approaches. We notice that the sparse source model requires a Student's t distribution with heavier tails than the NMF source model. We also clearly observe the advantage of modeling the spectro-temporal characteristics of the sources by means of an NMF model.

We then compare the proposed methods with three other ones from the literature in the same semi-blind setting: **(M1)** the Gaussian NMF-based method [10] where the convolutive mixing process is approximated as being instantaneous in the STFT domain[2]. Since the impulse responses of the mixing filters are longer than the STFT analysis window, they are truncated before computing the frequency responses required by this method; **(M2)** the Lasso method [16] with ℓ_1 regularization on the source TF coefficients; **(M3)** our previous method [19] with Gaussian NMF source modeling in the MDCT domain. M2 and M3 both rely on exact time-domain convolutive mixture modeling.

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

[2]The NMF parameters are actually updated as in [27] using multiplicative update rules.

	SDR	SIR	SAR
M1 [10]	1.7	8.5	4.9
M2 [16]	5.5	11.7	8.8
M3 [19]	**6.7**	12.5	9.5
Student's t sparse ($\alpha = 0.4$)	3.8	11.8	5.9
Student's t NMF ($\alpha = 14$)	**6.7**	**12.7**	**10.0**

TABLE I
Average source separation results in dB.

As can be seen in Table I, M1 performs the worst due to the STFT approximation of the convolutive mixing process. We observe that the two sparsity based methods, M2 and the proposed Student's t one, lead to the same performance in terms of interferences. However M2 is superior in terms of artifacts rejection and global quality. We finally notice that the the proposed Student's t NMF method performs the best, even though the improvement compared with our previous Gaussian approach M3 is small. Matlab code for the proposed VEM algorithm and audio examples are available at [28].

V. Conclusion

This paper introduced a multichannel audio source separation method based on exact convolutive mixture modeling and Student's t source modeling. Within this framework we compared a sparse and an NMF-based source model. The semi-blind experimental evaluation demonstrated the importance of modeling the spectro-temporal characteristics of the sources instead of only assuming sparsity. Future work will focus on developing a fully blind source separation method that exploits priors on the impulse response of the mixing filters. We could for example consider similar priors as in [29] for promoting sparsity and exponentially decaying envelop.

Appendix A
Standard probability distributions

Let $\mathcal{T}_\nu(\mu, \lambda)$ denote the Student's t distribution over a real-valued random variable (r.v.). ν, μ, λ are respectively the shape, location and scale parameters. Its probability density function (pdf) is given by:

$$T_\nu(x; \mu, \lambda) = \frac{1}{\sqrt{\nu\pi\lambda^2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{1}{\nu}\frac{(x-\mu)^2}{\lambda^2}\right)^{-\frac{\nu+1}{2}}, \tag{18}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

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

$$N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{19}$$

Let $\mathcal{IG}(\alpha, \beta)$ denote the Inverse-Gamma distribution over a positive r.v., its pdf is given by:

$$IG(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(\frac{-\beta}{x}\right). \tag{20}$$

The Inverse-Gamma distribution has the following properties:

$$\mathbb{E}[\ln(x)] = \ln(\beta) - \mathrm{di}\Gamma(\alpha) \quad \text{and} \quad \mathbb{E}[x^{-1}] = \frac{\alpha}{\beta},$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

where $\mathrm{di}\Gamma(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function.

APPENDIX B
M-STEP FOR THE STUDENT'S t NMF SOURCE MODEL

When $\lambda_{j,fn}^2 = [\mathbf{W}_j\mathbf{H}_j]_{fn}$, it can be shown that maximizing $E(q^\star;\boldsymbol{\theta})$ with respect to (w.r.t) $\mathbf{W}_j, \mathbf{H}_j$ under a non-negativity constraint is equivalent to minimizing the following cost function under the same constraint:

$$C(\mathbf{W}_j,\mathbf{H}_j) = \sum_{(f,n)\in\mathcal{B}} \left[\frac{[\mathbf{W}_j\mathbf{H}_j]_{fn}}{\beta_{j,fn}/\delta} - \ln\left([\mathbf{W}_j\mathbf{H}_j]_{fn}\right) \right]. \tag{21}$$

Interestingly, up to an additive constant independent of the NMF parameters, $C(\mathbf{W}_j,\mathbf{H}_j)$ is equal to $d_{IS}\left(\mathbf{W}_j\mathbf{H}_j, \mathbf{B}_j/\delta\right)$ where d_{IS} denotes the Itakura-Saito (IS) divergence [4] and $\mathbf{B}_j = [\beta_{j,fn}]_{fn} \in \mathbb{R}_+^{F\times N}$. Note that compared with standard IS-NMF [4], the NMF term here appears in the first argument of the IS divergence (which is not symmetric) instead of the second one. Based on Jensen's inequality, we can show that for any set $\{c_{jk,fn} \in [0,1]\}_{k=1}^{K_j}$ such that $\sum_{k=1}^{K_j} c_{jk,fn} = 1$:

$$-\ln\left([\mathbf{W}_j\mathbf{H}_j]_{fn}\right) \leq -\sum_{k=1}^{K_j} c_{jk,fn} \ln\left(\frac{w_{j,fk}h_{j,kn}}{c_{jk,fn}}\right), \tag{22}$$

where equality holds if and only if $c_{jk,fn} = w_{j,fk}h_{j,kn}/\left[\mathbf{W}_j\mathbf{H}_j\right]_{fn}$. It follows from (21) and (22):

$$C(\mathbf{W}_j,\mathbf{H}_j) \leq \sum_{(f,n)\in\mathcal{B}} \frac{[\mathbf{W}_j\mathbf{H}_j]_{fn}}{\beta_{j,fn}/\delta} - \sum_{k=1}^{K_j} c_{jk,fn} \ln\left(\frac{w_{j,fk}h_{j,kn}}{c_{jk,fn}}\right). \tag{23}$$

Letting the partial derivatives of this upper bound w.r.t $w_{j,fk}$ and $h_{j,kn}$ be zero, and replacing $c_{jk,fn}$ with the expression that equalizes the cost and its upper bound, we obtain the multiplicative update rules given in lines 6 and 7 of Algorithm 2. $\odot$ denotes entry-wise operation and division is taken entry-wise. The non-negativity constraint is satisfied provided that the NMF parameters are initialized with non-negative entries.

REFERENCES

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

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

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

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

[5] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 1825–1828.

[6] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2015, pp. 1–5.

[7] U. Şimşekli, A. Liutkus, and A. T. Cemgil, "Alpha-stable matrix factorization," *IEEE Signal Process. Lett.*, 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. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, 2016, pp. 51–55.

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

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

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

[12] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[13] H. Attias, "New EM algorithms for source separation and deconvolution with a microphone array." in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, 2003, pp. 297–300.

[14] R. Badeau and M. D. Plumbley, "Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 11, pp. 1670–1680, 2014.

[15] X. Li, L. Girin, and R. Horaud, "Audio source separation based on convolutive transfer function and frequency-domain lasso optimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New Orleans, LA, USA, 2017.

[16] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1818–1829, 2010.

[17] S. Arberet and P. Vandergheynst, "Reverberant audio source separation via sparse and low-rank modeling," *IEEE Signal Process. Lett.*, vol. 21, no. 4, pp. 404–408, 2014.

[18] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, 2016, pp. 1–5.

[19] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation: variational inference of time-frequency sources from time-domain observations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New Orleans, LA, USA, 2017.

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

[21] C. Fevotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2174–2188, 2006.

[22] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "An inverse-gamma source variance prior with factorized parameterization for audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, 2016, pp. 136–140.

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

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

[25] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Process.*, vol. 92, pp. 1928–1936, 2012.

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

[27] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, 2011, pp. 257–260.

[28] https://perso.telecom-paristech.fr/leglaive/demo-eusipco17.html

[29] A. Benichoux, L. S. R. Simon, E. Vincent, and R. Gribonval, "Convex regularizations for the simultaneous recording of room impulse responses," *IEEE Trans. Signal Process.*, vol. 62, no. 8, pp. 1976–1986, 2014.

ISBN 978-0-9928626-7-1 © EURASIP 2017