An EM Algorithm for Audio Source Separation Based on the Convolutive Transfer Function

Xiaofei Li, Laurent Girin, Radu Horaud

AN EM ALGORITHM FOR AUDIO SOURCE SEPARATION BASED ON THE CONVOLUTIVE TRANSFER FUNCTION

Xiaofei Li,[1] *Laurent Girin,*[2,1] *Radu Horaud,*[1]

[1] INRIA Grenoble Rhône-Alpes, 38330 Montbonnot-Saint-Martin, France
{xiaofei.li, radu.horaud}@inria.fr
[2] Univ. Grenoble Alpes, GIPSA-Lab, 38400 Saint-Martin d'Hères, France
laurent.girin@gipsa-lab.grenoble-inp.fr

ABSTRACT

This paper addresses the problem of audio source separation from (possibly under-determined) multichannel convolutive mixtures. We propose a separation method based on the convolutive transfer function (CTF) in the short-time Fourier transform domain. For strongly reverberant signals, the CTF is a much more appropriate model than the widely-used multiplicative transfer function approximation. An Expectation-Maximization (EM) algorithm is proposed to jointly estimate the model parameters, including the CTF coefficients of the mixing filters, and infer the sources. Experiments show that the proposed method provides very satisfactory performance on highly reverberant speech mixtures.

Index Terms— Audio source separation, convolutive transfer function, EM algorithm.

1. INTRODUCTION

In this paper we address the problem of multichannel audio source separation (MASS) from (possibly underdetermined) convolutive mixtures. Most of MASS techniques are designed in the short time Fourier transform (STFT) domain where the convolutive process is generally approximated at each time-frequency (TF) bin by a product between a source STFT coefficient and the Fourier transform of the mixing filter, e.g. [1, 2, 3, 4]. This assumption is here referred to as the multiplicative transfer function (MTF) approximation [5]. It is theoretically valid only if the length of the mixing filter impulse response is smaller than the length of the STFT window. Since the latter is limited to assume signal local stationarity, this is very rarely the case in practice, even for moderately reverberant environments. Hence the MTF can be a poor approximation, fundamentally endangering the separation performance. This is even more critical for strongly reverberant environments.

Yet, the MTF is poorly questioned in the MASS literature, and only a few studies attempted to tackle its limitations. In [6], the use of a full-rank spatial covariance matrix for the source images, instead of the rank-1 matrix corresponding to the MTF model [4], is claimed to overcome to some extent the limitations of MTF. The problem is circumvoluted in [7, 8] by estimating the source signals in the time domain using a Lasso optimization technique. This achieves quite good source separation performance in reverberant environments, at the price of a tremendous computation time. Also, only semi-blind separation with known mixing filters was addressed in [7]. A variational Expectation-Maximization (EM) algorithm was recently proposed in [9] in which the convolutive process is expressed in the time-domain whereas separation is carried out in the TF domain. This led to very interesting results, again at the price of huge computation.

To accurately represent convolution in the STFT domain, cross-band filters (CBFs) were introduced in [10], as an alternative to MTF. Using the CBFs, an output STFT coefficient is represented as a summation over frequency bins of multiple convolutions between the input STFT coefficients and the TF-domain filter impulse response, along the frame axis. Considering only the current frequency bin, i.e. a unique convolution along the STFT frame axis, is a reasonable practical approximation, referred to as the convolutive transfer function (CTF) model [11]. CBFs were considered for solving the MASS problem in [12], in combination with a high-resolution non-negative matrix factorization (HR-NMF) model of the source signal. A variational EM algorithm was proposed to estimate the filters and infer the source signals. Unfortunately, due to the model complexity, this method was observed to perform well only in an oracle setup where both filters and source parameters are initialized from the individual source images. In [13], a STFT-domain convolutive model was used together with an HMM model on source activity. However, the optimization method used to estimate the parameters and infer the source signal was quite complex. In our previous work [14], a Lasso-type optimization was applied for MASS within the CTF framework and led to drastically reduce the computation time compared to [7]. However, again, this was done only in a semi-blind setup.

In the present paper we further extend this work on CTF-based MASS: we plug the CTF model (presented in Section 2) within a MASS probabilistic framework (Section 3) and we propose an EM algorithm (Section 4) for joint estimation of CTF coefficients and source parameters and inference of source STFT coefficients. As most probabilistic EM-based algorithms, the proposed algorithm is not truly blind in the sense that it requires a fairly good initialization to behave efficiently. We show in Section 5 that it provides superior performance to [1, 3, 6, 9, 14] within a semi-blind set-up where the filters and sources parameters are initialized "reasonably". Importantly, compared to the time-domain convolutive model [7, 8, 9], the CTF dramatically decreases the data size to be processed. As a result, the proposed CTF-based methods are much easier to converge, and needs much less iterations.

This research has received funding from the ERC Advanced Grant VHIA (#340113).

2. CONVOLUTIVE TRANSFER FUNCTION

In a reverberant and noise-free environment, a source image $y(n)$ in the time domain is given by the linear convolution between the source signal $s(n)$ and the impulse response of the propagating filter $a(n)$. This convolution is usually approximated in the STFT domain as the product $y(p,k) = a(k)s(p,k)$, where $y(p,k)$ and $s(p,k)$ are the STFT of the corresponding signals, and $a(k)$ is the discrete Fourier transform of $a(n)$, N is the frame length, $k \in [0, N-1]$ is the frequency bin index, and $p \in [1, P]$ is the frame index. As discussed above, this MTF approximation is only valid when $a(n)$ is shorter than the STFT window, which is often questionable. In this paper we therefore use the CTF model, i.e. $y(n)$ is approximated in the STFT domain by [11]:

$$y(p,k) = a(p,k) \star s(p,k) = \sum_{p'} a(p',k)s(p-p',k). \quad (1)$$

The filter CTF is defined as the TF-domain impulse response $a(p,k)$. It is related to the time-domain impulse response $a(n)$ by:

$$a(p,k) = (a(n) \star \zeta_k(n))|_{n=pL}, \quad (2)$$

which represents the convolution with respect to the time index n evaluated at multiples of the frame step L, with

$$\zeta_k(n) = e^{j\frac{2\pi}{N}kn} \sum_{m=-\infty}^{+\infty} \tilde{\omega}_a(m)\, \tilde{\omega}_s(n+m),$$

where $\tilde{\omega}_a(n)$ and $\tilde{\omega}_s(n)$ denote the STFT analysis and synthesis windows, respectively. In summary, within the CTF model, the time-domain convolution is transformed into a TF-domain convolution at every frequency bin k. The corresponding approximation error is (much) lower than the error resulting from the MTF approximation for the reverberant case.

3. MULTICHANNEL AUDIO SOURCE SEPARATION BASED ON CTF: MODEL FORMULATIONS

3.1. Basic mixture model formulation and probabilistic model

We consider a convolutive mixture with J sources and I sensors, possibly underdetermined (i.e. we may have $I < J$). Based on the CTF model (1), for all frames $p \in [1, P]$, and all frequency bins $k \in [0, N-1]$, the observed signal $x_i(p,k)$ at channel $i \in [1, I]$ is given in the STFT domain by:

$$x_i(p,k) = \sum_{j=1}^{J} a_{ij}(p,k) \star s_j(p,k) + e_i(p,k), \quad (3)$$

where $a_{ij}(p,k)$ is the CTF from source j to sensor i, and $e_i(p,k)$ denotes the noise signal.

Probabilistic model As is now classical in many MASS papers, the source signals $s_j(p,k)$ are assumed to be mutually independent, and individually independent across STFT frames and frequency bins. Each coefficient $s_j(p,k)$ is assumed to follow a zero-mean complex Gaussian distribution with variance $v_j(p,k)$ [4, 6], i.e. its probability density function (pdf) is:

$$\mathcal{N}_c(s_j(p,k); 0, v_j(p,k)) = \frac{1}{\pi v_j(p,k)} \exp\big(-\frac{|s_j(p,k)|^2}{v_j(p,k)}\big).$$

The noise signal is assumed to be zero-mean complex Gaussian, stationary, independent to all source signals, and individually independent across STFT frames and frequency bins. However we assume possible inter-channel noise correlation. Defining the noise vector $\boldsymbol{e}(p,k) = [e_1(p,k), \ldots, e_I(p,k)]^\top \in \mathbb{C}^{I\times 1}$, its pdf is:

$$\mathcal{N}_c(\boldsymbol{e}(p,k); 0, \boldsymbol{\Sigma}_{\mathbf{e}}(k)) = \frac{1}{\pi^I |\boldsymbol{\Sigma}_{\mathbf{e}}(k)|} e^{-\boldsymbol{e}(p,k)^H \boldsymbol{\Sigma}_{\mathbf{e}}(k)^{-1} \boldsymbol{e}(p,k)},$$

where ^{H} denotes complex transpose. In this work, $\boldsymbol{\Sigma}_{\mathbf{e}}(k)$ is assumed to be known, though it could easily be included in the parameters to be estimated.

All CTF coefficients $a_{ij}(p,k)$ are considered as (unknown) parameters, and we assume for simplicity that all CTFs have the same length denoted $Q+1$, with $Q \ll P$. Since the mixture model (3) is defined independently at each frequency bin k, and since all signals are assumed independent across frequency bins, the separation process is carried out independently at each frequency. Therefore, from now on, we omit the frequency index k to clarity the presentation.

3.2. Two vector/matrix reformulations

Section 4 will present an EM algorithm for joint estimation of model parameters and source inference. Before that, we present two reformulations of the above model. Formulation 1 enables us to derive the M-step of the EM algorithm in a very compact form, and Formulation 2 enables us to derive the E-step in a very compact form. Both formulations are useful since, as will see, the other way round is not true. Going from the E-step to the M-step and vice-versa will only necessitate to reorganize the variables and parameters in the appropriate vector/matrix form.

Formulation 1: Let us define the following vectors, for $p \in [1, P]$:

$$\boldsymbol{s}_j(p) = [s_j(p), \ldots, s_j(p-q), \ldots, s_j(p-Q)]^\top \in \mathbb{C}^{(Q+1)\times 1},$$
$$\boldsymbol{s}(p) = [\boldsymbol{s}_1(p)^\top, \ldots, \boldsymbol{s}_j(p)^\top, \ldots, \boldsymbol{s}_J(p)^\top]^\top \in \mathbb{C}^{J(Q+1)\times 1},$$

where $\top$ denotes matrix transpose. If $p \leq q$, we set $s_j(p-q) = 0$. We already defined $\boldsymbol{e}(p) \in \mathbb{C}^{I\times 1}$. The microphone signal $\boldsymbol{x}(p) \in \mathbb{C}^{I\times 1}$ can be defined in the same way as $\boldsymbol{e}(p)$, and we have:

$$\boldsymbol{x}(p) = \sum_{j=1}^{J} \mathbf{A}_j \boldsymbol{s}_j(p) + \boldsymbol{e}(p) = \mathbf{A}\boldsymbol{s}(p) + \boldsymbol{e}(p), \quad (4)$$

where $\mathbf{A}_j \in \mathbb{C}^{I\times(Q+1)}$ is the matrix with the i-th row being the CTF of source j to sensor i, and $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_j, \ldots, \mathbf{A}_J] \in \mathbb{C}^{I\times J(Q+1)}$. Note that with Formulation 1, the pdf of the mixture conditioned on the sources is $\mathcal{N}_c(\boldsymbol{x}(p); \mathbf{A}\boldsymbol{s}(p), \boldsymbol{\Sigma}_{\mathbf{e}})$.

Formulation 2: Let $\boldsymbol{s}_j = [s_j(1), \ldots, s_j(P)]^T$, $\boldsymbol{e}_i = [e_i(1), \ldots, e_i(P)]^T$ and $\boldsymbol{x}_i = [x_i(1), \ldots, x_i(P)]^T$ denote the j-th source vector and the i-th noise and microphone vectors, all involving all P frames, hence all in $\mathbb{C}^{P\times 1}$. Let us define the corresponding filter matrix in $\mathbb{C}^{P\times P}$:

$$\mathcal{A}_{ij} = \begin{bmatrix} a_{ij}(0) & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ a_{ij}(Q) & \ddots & a_{ij}(0) & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{ij}(Q) & \cdots & a_{ij}(0) \end{bmatrix},$$

where the flipped filter CTF $\{a_{ij}(p)\}$ is duplicated as the row vectors, with one element shift per row. Then we have:

$$\boldsymbol{x}_i = \sum_{j=1}^{J} \mathcal{A}_{ij}\boldsymbol{s}_j + \boldsymbol{e}_i = \mathcal{A}_i\boldsymbol{s} + \boldsymbol{e}_i, \tag{5}$$

where $\mathcal{A}_i = [\mathcal{A}_{i1}, \ldots, \mathcal{A}_{iJ}] \in \mathbb{C}^{P\times JP}$, and $\boldsymbol{s} = [\boldsymbol{s}_1^\top, \ldots, \boldsymbol{s}_J^\top]^\top \in \mathbb{C}^{JP\times 1}$.[1] If we further concatenate these quantities over sensors as $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_I^\top]^\top \in \mathbb{C}^{IP\times 1}$, $\boldsymbol{e} = [\boldsymbol{e}_1^\top, \ldots, \boldsymbol{e}_I^\top]^\top \in \mathbb{C}^{IP\times 1}$ and $\mathcal{A} = [\mathcal{A}_1^\top, \ldots, \mathcal{A}_I^\top]^\top \in \mathbb{C}^{IP\times JP}$, then we have:

$$\boldsymbol{x} = \mathcal{A}\boldsymbol{s} + \boldsymbol{e}. \tag{6}$$

Using Formulation 2, the source vector $\boldsymbol{s}$ follows a zero-mean complex Gaussian distribution with $JP \times JP$ diagonal covariance matrix $\boldsymbol{\Psi}_{\mathbf{s}}$ where the P first diagonal entries are $v_1(1), \ldots, v_1(P)$, the next P diagonal entries are $v_2(1), \ldots, v_2(P)$, and so on. The noise vector $\boldsymbol{e}$ follows a zero-mean complex Gaussian distribution with $IP \times IP$ covariance matrix $\boldsymbol{\Psi}_{\mathbf{e}}$, with the entries $\boldsymbol{\Psi}_{\mathbf{e}}((i_1-1)P+p_1, (i_2-1)P+p_2)$ being equal to $\boldsymbol{\Sigma}_{\mathbf{e}}(i_1, i_2)$ if $p_1 = p_2$, and 0 otherwise (here the arguments in parentheses denotes the row and column index; $i_1, i_2 \in [1, I], p_1, p_2 \in [1, P]$). The pdf of the mixture conditioned on the sources is $\mathcal{N}_c(\boldsymbol{x}; \mathcal{A}\boldsymbol{s}, \boldsymbol{\Psi}_{\mathbf{e}})$.

4. EM ALGORITHM FOR MASS WITH CTF

Denote $\mathbf{V} = \{v_j(p)\}_{j\in[1,J],p\in[1,P]}$ as the set of source variances. The entire set of parameters of the present problem is $\Theta = \{\mathbf{V}, \mathbf{A}\}$. We present an EM algorithm that was derived to jointly obtain the maximum likelihood estimation of the parameters and the inference of the STFT coefficients $\{s_j(p)\}_{j,p}$ of the source signals.

4.1. M-step

We use here Formulation 1, and we denote the set of observations as $\mathbf{X} = \{\boldsymbol{x}(p)\}_{p\in[1,P]}$, and the set of source signals as $\mathbf{S} = \{\boldsymbol{s}(p)\}_{p\in[1,P]}$. The complete-data likelihood function is:

$$\begin{aligned} p(\mathbf{X}, \mathbf{S}|\Theta) &\propto p(\mathbf{X}|\mathbf{S}, \Theta)p(\mathbf{S}|\Theta) \\ &= \prod_{p=1}^{P} \mathcal{N}_c(\boldsymbol{x}(p); \mathbf{A}\boldsymbol{s}(p), \boldsymbol{\Sigma}_{\mathbf{e}}) \prod_{j=1}^{J}\prod_{p=1}^{P} \mathcal{N}_c(s_j(p); 0, v_j(p)). \end{aligned}$$

Let us denote by Θ^{old} the value of Θ at previous EM iteration. Let us denote by $\mathbb{E}_{\mathbf{S}|\mathbf{X},\Theta^{\text{old}}}[\,.\,]$ the expectation in the sense of the posterior distribution $p(\mathbf{S}|\mathbf{X}, \Theta^{\text{old}})$. The auxiliary function $Q(\Theta, \Theta^{\text{old}}) = \mathbb{E}_{\mathbf{S}|\mathbf{X},\Theta^{\text{old}}}[\log(p(\mathbf{X}, \mathbf{S}|\Theta))]$ is given by:

$$\begin{aligned} Q(\Theta, \Theta^{\text{old}}) = \sum_{p=1}^{P} {\boldsymbol{\Sigma}_{\mathbf{e}}}^{-1} \text{Trace}\Big(&\mathbf{A}\widehat{\boldsymbol{s}}(p)\boldsymbol{x}(p)^H + \boldsymbol{x}(p)\widehat{\boldsymbol{s}}(p)^H\mathbf{A}^H \\ &- \mathbf{A}\widehat{\mathbf{R}}_{\mathbf{s}}(p)\mathbf{A}^H\Big) - \sum_{j=1}^{J}\sum_{p=1}^{P}\Big(\log(v_j(p)) + \frac{\widehat{v}_j(p)}{v_j(p)}\Big) + \text{const}, \end{aligned}$$

where $\widehat{\boldsymbol{s}}(p) = \mathbb{E}_{\mathbf{S}|\mathbf{X},\Theta^{\text{old}}}[\boldsymbol{s}(p)]$, $\widehat{\mathbf{R}}_{\mathbf{s}}(p) = \mathbb{E}_{\mathbf{S}|\mathbf{X},\Theta^{\text{old}}}[\boldsymbol{s}(p)\boldsymbol{s}(p)^H]$, and $\widehat{v}_j(p) = \mathbb{E}_{\mathbf{S}|\mathbf{X},\Theta^{\text{old}}}[|s_j(p)|^2]$ are respectively the posterior mean and the posterior second-order moment matrix of the source vector, and the source-wise posterior second-order moment. These quantities are provided by the preceding E-step, with reorganization of the entries. Note that $\widehat{v}_j(p)$ is the $((j-1)(Q+1)+1)$-th diagonal entry of $\widehat{\mathbf{R}}_{\mathbf{s}}(p)$.

Setting the complex derivative of $Q(\Theta, \Theta^{\text{old}})$ with respect to $\mathbf{A}^*$ (* denotes conjugate) and $v_j(p)$ equal to zero, the update of $\mathbf{A}$ and $v_j(p)$ in the M-step are respectively:

$$\begin{aligned} \mathbf{A}^{\text{new}} &= \Big(\sum_{p=1}^{P} \boldsymbol{x}(p)\widehat{\boldsymbol{s}}(p)^H\Big)\Big(\sum_{p=1}^{P} \widehat{\mathbf{R}}_{\mathbf{s}}(p)\Big)^{-1}, \\ v_j^{\text{new}}(p) &= \widehat{v}_j(p). \end{aligned} \tag{7}$$

4.2. E-step

The E-step is efficiently derived using Formulation 2. Using the current parameter estimates Θ (given in the preceding M-step by (7)), we construct the filter matrix $\mathcal{A}$ and $\boldsymbol{\Psi}_{\mathbf{s}}$ following Formulation 2. The posterior probability distribution of the source vector $\boldsymbol{s}$ writes $p(\boldsymbol{s}|\boldsymbol{x}, \Theta) \propto p(\boldsymbol{x}|\boldsymbol{s}, \Theta)p(\boldsymbol{s}|\Theta)$, and we have seen that both $p(\boldsymbol{x}|\boldsymbol{s}, \Theta)$ and $p(\boldsymbol{s}|\Theta)$ are Gaussian. Therefore, $p(\boldsymbol{s}|\boldsymbol{x}, \Theta)$ is Gaussian and the posterior mean $\widehat{\boldsymbol{s}}$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}}$ can be derived by reorganizing the quadratic and linear forms in $\boldsymbol{s}$ in the exponent of the distribution. We obtain:

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}} &= \big(\mathcal{A}^H {\boldsymbol{\Psi}_{\mathbf{e}}}^{-1}\mathcal{A} + {\boldsymbol{\Psi}_{\mathbf{s}}}^{-1}\big)^{-1}, \\ \widehat{\boldsymbol{s}} &= \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}}\mathcal{A}^H {\boldsymbol{\Psi}_{\mathbf{e}}}^{-1}\boldsymbol{x}. \end{aligned} \tag{8}$$

Eq. (8) has the classical form of source estimation with Wiener filtering, as obtained with Gaussian models in the MTF framework [4]. However, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}}$ is here $JP \times JP$ and $\widehat{\boldsymbol{s}}$ is $JP \times 1$, hence the inference is performed jointly on all frames of the whole sequence, exploiting interframe information in $\boldsymbol{x}$, and thus performing separation through multichannel deconvolution. More precisely, the structure of $\mathcal{A}$ enables to exploit $Q+1$ frames of $\boldsymbol{x}$ in this process.

Reformulation (for the next M-step): Let the subscript $_{\{j,p\}}$ denote "the j-th source at p-th frame" within a vector, or within a row or a column of a matrix. Using Formulation 2, the index of $\widehat{\boldsymbol{s}}_{\{j,p\}}$ is $\{(j-1)P+p\}$, and this is also valid for the column and row index of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}}$. Using Formulation 1, the index of $\widehat{\boldsymbol{s}}(p)_{\{j,p-q\}}$ is $\{(j-1)(Q+1)+q+1\}$ ($q \in [0, Q]$), and this is also valid for the column and row index of $\widehat{\mathbf{R}}_{\mathbf{s}}(p)$. In the next M-step, $\widehat{\boldsymbol{s}}(p)_{\{j,p-q\}}$ is given by $\widehat{\boldsymbol{s}}_{\{j,p-q\}}$ from the previous E-step, and the entries of $\widehat{\mathbf{R}}_{\mathbf{s}}(p)$ are computed from $\widehat{\boldsymbol{s}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}}$ from the previous E-step by:

$$\begin{aligned} \widehat{\mathbf{R}}_{\mathbf{s}}(p)_{\{j_1,p-q_1\},\{j_2,p-q_2\}} = \widehat{\boldsymbol{s}}_{\{j_1,p-q_1\}}\widehat{\boldsymbol{s}}^*_{\{j_2,p-q_2\}} + \\ \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}\{j_1,p-q_1\},\{j_2,p-q_2\}}. \end{aligned} \tag{9}$$

5. EXPERIMENTS

Simulation set-up: To test the efficiency of the proposed source separation method, experiments were conducted with simulated binaural signals, under various acoustic conditions. Binaural room

[1] Note that $\boldsymbol{s}(p)$ from Formulation 1 is not a subvector of $\boldsymbol{s}$ from Formulation 2, and the same for the other vectors. Going from Formulation 1 to Formulation 2 necessitates to reorganize the vector entries.

impulse responses (BRIR) were generated with the ROOMSIM simulator [15] and with the head related impulse response (HRIR) of a KEMAR dummy head [16]. 16-kHz Speech signals from the TIMIT dataset [17] were convolved with the simulated BRIRs to generate sensor signals. The duration of each sensor signal was 3s. The speech sources were located with azimuth directions varying from -90° to 90°, spaced by 5°, and an elevation of 0°. The anechoic case and three reverberation times, $T_{60} = 0.22$ s, 0.5 s and 0.79 s, were tested. A set of underdetermined mixtures with 3, 4 and 5 sources were processed. For each experiment, 20 mixtures were generated. The STFT window was a Hamming window of $1{,}024$ samples (64 ms), with 75% overlap. The noise covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e}}(k)$ is set as an isotropic diagonal matrix for all frequencies, with 1% of the power of microphone signals.

EM initialization: As is usual with EM algorithms, the initialization of our EM is critical, and can be conducted by initializing either the E-step or the M-step. In this study, we consider a semi-blind set-up, where the CTFs are computed by (2) from the known time-domain filters. The CTF-based Lasso-type method (CTF-Lasso) proposed in [14] is then applied to obtain a first source vector estimate. Briefly stated, it solves the following optimization problem:

$$\min_{\boldsymbol{s}} \parallel \mathcal{A}\boldsymbol{s} - \boldsymbol{x} \parallel_2^2 + \lambda \parallel \boldsymbol{s} \parallel_1,$$

(see [14] for details). Then the magnitude square of each source coefficient estimate is taken as the initialization of source variance. The number of iteration is empirically set to a constant in this work, i.e. 7 (convergence criteria will be investigated in future work).

Baseline methods: For comparison, we tested six state-of-the-art source separation methods, all set in the same semi-blind configuration as the proposed EM method (i.e. with known mixing filters): i) the CTF-Lasso used for initialization[2], ii) DUET [1], iii) the ℓ_1-MIN method of [3], iv) the full-rank spatial covariance matrix (FR-SCM) method [6], v) the wideband Lasso (W-Lasso) method [7] with a sparsity regularization factor of 10^{-5}, and vi) the variational EM method of [9] (VEM). For the latter, the NMF parameters were initialized using the output of CTF-Lasso. DUET and ℓ_1-MIN are based on the MTF approximation. Since the BRIRs are longer than the STFT window, they have to be truncated to generate the TF-domain mixing matrix. However we obtained better results using the Fourier transform of the HRIRs. For FR-SCM, the SCMs were individually estimated using each separate source image, and then kept fixed during the EM, following the line of the semi-oracle experiments in [6].

Results: The signal-to-distortion ratio (SDR) [18] in decibels (dB), averaged over 20 mixtures for each condition, is used as the separation performance metric. Fig. 1(left) plots the SDR obtained for 3-source mixtures and for the 4 reverberation times. It can be seen in these plots that all seven methods achieve high SDR in the anechoic case. As T_{60} increases, the SDR of DUET and ℓ_1-MIN dramatically decreases, since the MTF approximation is no longer suitable when the filter impulse response is (much) longer than the STFT window. FR-SCM mitigates the problem to a quite limited extend. In contrast to these three methods, W-Lasso, VEM, CTF-Lasso and the proposed method achieve remarkable performances: the SDR actually increases with T_{60}, which is a bit surprising at first sight. The reason is possibly that given a good fit of the mixture model, the longer the filter, the more information is available to discriminate and separate different sources. For the 3-source mixtures, W-Lasso performs the best. Due to the CTF approximation error, CTF-Lasso is below W-Lasso by 4 to 5 dB. Taking the output of CTF-Lasso as an initial point, VEM improves the SDR by $\approx$ 1 dB for low T_{60}, while reducing the SDR a bit for high T_{60}. In contrast, the proposed EM algorithm refines the source estimate and improves the SDR by 3 to 4 dB over CTF-Lasso for every T_{60}. Possible reasons are i) VEM is the combination of an exact model (time-domain convolution) with an approximate algorithm (variational EM) while the proposed method is a combination of an approximate model (CTF) with an exact EM algorithm, hence the "loss in the model" may be lower than the "loss in the algorithm", and ii) the source variance is modeled by NMF in VEM, which may not be sufficiently accurate for speech signals, while the proposed method keeps a free source variance parameter for each TF bin.

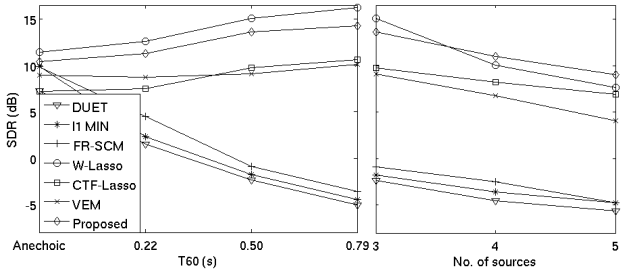

Figure 1: Source separation performance (SDR) **left** for 3-source mixtures as a function of reverberation time, and **right** for reverberation time of 0.5 s as a function of number of sources.

Fig. 1(right) plots the SDR for various number of sources, for $T_{60} = 0.5$ s. As expected, the SDR of all methods degrades when the number of sources increases. The time-domain signals (and filters) have a much larger data size than the STFT-domain signals (and filters). Thence W-Lasso and VEM have more difficulties to converge. For example, in this experiment, W-Lasso, VEM , CTF-Lasso and the proposed method respectively ran $20{,}000$, 100, ≈ 50 and 7 iterations. As a result, when the number of sources increase, the convergence of W-Lasso and VEM becomes more difficult, and the SDR scores have a larger degradation with the number of sources compared to the CTF-based methods. The proposed EM algorithm achieves an SDR improvement of about 3 dB and 2 dB over CTF-Lasso for the case of 4 sources and 5 sources, respectively. In these settings, it performs the best of all methods (though W-Lasso was best for 3 sources).

6. CONCLUSION

In this paper, an EM algorithm was proposed for MASS based on CTF. Two convolution formulations were used to respectively derive the M-step and E-step in a compact form. Overall, the proposed method, VEM and the two Lasso methods perform prominently better than the MTF-based methods by circumventing the narrowband approximation. The proposed EM algorithm is efficient to refine the source estimate and improve the performance measures starting with the output of CTF-Lasso. This scheme achieves the best performance for the cases of 4 sources and 5 sources, and achieves close performance to the best, i.e. W-Lasso, for the case of 3 sources, with a much lower number of iterations.

[2]To avoid the frequency aliasing caused by the decimation of STFT, and to reduce the CTF length, the STFT for CTF-lasso uses a window of 1024 samples with 75% overlap in this work, rather than 512 samples with 50% overlap in [14], which leads to the different results from [14].

7. REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[2] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[3] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.

[4] 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.

[5] 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.

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

[7] 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.

[8] S. Arberet, P. Vandergheynst, J.-P. Carrillo, R. E.Thiran, and Y. Wiaux, "Sparse reverberant audio source separation via reweighted analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1391–1402, 2013.

[9] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation: variational inference of time-frequency sources from time-domain observations," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[10] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[11] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

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

[13] T. Higuchi and H. Kameoka, "Joint audio source separation and dereverberation based on multichannel factorial hidden Markov model," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[14] X. Li, L. Girin, and R. Horaud, "Audio source separation based on convolutive transfer function and frequency-domain lasso optimization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017.

[15] D. R. Campbell, "The ROOMSIM user guide (v3.3)," 2004. [Online]. Available: http://media.paisley.ac.uk/_campbell/Roomsim

[16] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR dummy-head microphone," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology, Gaithersburgh, MD, Tech. Rep. 107, 1988.

[18] 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.