Multichannel audio source separation: variational inference of time-frequency sources from time-domain observations
Simon Leglaive, Roland Badeau, Gaël Richard

To cite this version:
Simon Leglaive, Roland Badeau, Gaël Richard. Multichannel audio source separation: variational inference of time-frequency sources from time-domain observations. 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Mar 2017, La Nouvelle Orléans, LA, United States. hal-01416347

HAL Id: hal-01416347
https://hal.archives-ouvertes.fr/hal-01416347
Submitted on 11 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
MULTICHANNEL AUDIO SOURCE SEPARATION: VARIATIONAL INFERENCE OF TIME-FREQUENCY SOURCES FROM TIME-DOMAIN OBSERVATIONS

Simon Leglaive, Roland Badeau, Gaël Richard

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

ABSTRACT

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

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

1. INTRODUCTION

Audio source separation is the task that aims to recover a set of audio source signals from one or several mixtures. When the sources are recorded in an enclosed space, the reflections of sound on the surfaces and objects of the room induce reverberation in the recordings. Modeling such reverberant mixtures is a challenge that involves two main stages: modeling the source signals and the way they are mixed togethe.
Note that e.g. method [18], the authors also use an MDCT along with a Gaussian TF dictionary for all the sources. The solution of (8)-(9) can be found by means of an Expectation-Maximization (ML) sense:

\[ s_{j,t} \sim N(0, \sigma_s^2), \quad \theta^* = \arg \max \ p(\mathbf{x} | \theta). \] (9)

where \( \mathbf{x} \) denotes latent source variables, and \( \theta \) is the set of parameters \( \{\sigma_s^2\}_i, \{\mathbf{W}_j\}_i, \{\mathbf{H}_j\}_i \) and \( \{a_{ij}(t)\}_t \). Our objective is to estimate the latent variables in a Minimum Mean Square Error (MMSE) sense:

\[ \hat{s} = \mathbb{E}_{\mathbf{x}|\mathbf{s}, \theta} \cdot [s], \] (8)

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

\[ \theta^* = \arg \max \ p(\mathbf{x} | \theta). \] (9)

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

\[ \ln p(\mathbf{x} | \theta) = \mathcal{L}(q; \theta) + D_{KL}(q || p(\mathbf{s} | \theta)). \] (10)

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

\[ E(q; \theta) = \langle \ln p(\mathbf{x}, \mathbf{s} | \theta) \rangle_q, \] (11)

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

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

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

3.1. Source estimate under the variational approximation

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

\[ m_{jfn} = \langle s_{jfn} \rangle_{q^*}. \] (13)

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

\[ \hat{s}_{j}(t) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{jfn} \hat{v}_{jfn}(t). \] (14)

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

\[ \hat{y}_{j}= [a_{ij} * \hat{s}_{j}] (t) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{jfn} g_{jfn}(t). \] (15)

3.2. Variational free energy

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

\[ \ln p(\mathbf{x}, \mathbf{s} | \theta) = -\frac{1}{2}(IT + JFN) \ln(2\pi) - \frac{1}{2} \sum_{i=1}^T \left[ \ln(\sigma_i^2) + \frac{1}{\alpha_i^2} \left( x_i(t) - \sum_{j=1}^I y_{ij}(t) \right)^2 \right] \] (16)

\[ - \frac{1}{2} \sum_{j=1}^J \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \left[ \ln(v_{jfn}) + \frac{s_{jfn}^2}{v_{jfn}} \right]. \] (16)

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

\[ E(q^*; \theta) = -\frac{1}{2}(IT + JFN) \ln(2\pi) - \frac{1}{2} \sum_{i=1}^T \left[ \ln(\sigma_i^2) + \frac{1}{\alpha_i^2} \right] - \frac{1}{2} \sum_{j=1}^J \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \left[ \ln(v_{jfn}) + \frac{m_{jfn}^2 + \gamma_{jfn}}{v_{jfn}} \right]. \] (17)
where \( m_{j,f,n} \) is defined in (13), \( e_i(t) = (\bar{x}_i(t) - \sum_{j=1}^J y_{ij}(t))^2 \) and 
\( \gamma_{j,f,n} = (\langle s_{j,f,n} - m_{j,f,n} \rangle^2)^{\frac{1}{2}} \). From (6), (12) and (15), \( e_i(t) \) is given by:

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

From (12) and after having identified \( q(s)^* \) at the E-step we will have to compute the entropy \( H(q^*) = -\langle \ln(q^*(s)) \rangle_q \).

### 3.3. E-step

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

\[
\ln q_{j,f,n}(s_{j,f,n}) = \langle \ln p(x,s;\theta) \rangle_{q_j \cdots q_m, \gamma_{j,f,n}} (19).
\]

From (16) we can develop this expression. After computation we find that
\[
d_{j,f,n}(L_{j,f,n}) = N(s_{j,f,n}; m_{j,f,n}; \gamma_{j,f,n})
\]

where:

\[
(1) \quad \gamma_{j,f,n} = \left( \frac{1}{v_{j,f,n}} + \sum_{i=1}^I \int_{t=0}^{T-1} \bar{g}_{ij,f,n}(t) \right)^{-1};
\]

\[
(2) \quad m_{j,f,n} = m_{j,f,n} - \gamma_{j,f,n} d_{j,f,n};
\]

\[
(22) \quad d_{j,f,n} = \frac{m_{j,f,n}}{v_{j,f,n}} - \sum_{i=1}^I \int_{t=0}^{T-1} \bar{g}_{ij,f,n}(t) \left( x_i(t) - \sum_{j'=1}^J \hat{y}_{ij'}(t) \right).
\]

Note that the parameters \( m_{j,f,n} \) have to be updated in turn².

**Entropy of the distribution:** From (12) and the variational distribution \( q_{j,f,n}(s_{j,f,n}) \) we identified, we can show that:

\[
H(q^*) = \frac{JFN}{2} \left( 1 + \ln(2\pi) \right) + \frac{1}{2} \sum_{j=1}^J \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \ln(\gamma_{j,f,n}).
\]

### 3.4. M-step

The M-step aims to maximize \( E(q^*; \theta) \) in (17) with respect to (w.r.t) the parameter set \( \theta \). Zeroining the derivative of \( E(q^*; \theta) \) w.r.t \( \sigma_i^2 \) leads to the following update with \( e_i(t) \) given by (18):

\[
\sigma_i^2 = \frac{1}{T} \sum_{t=0}^{T-1} e_i(t).
\]

To an additive term which does not depend on the NMF parameters, we can recognize in (17) the Itakura-Saito (IS) divergence [3] between \( v_{j,f,n} = [W_j^k h_n^j] \) and the posterior mean of the source power spectrogram \( \langle s_{j,f,n}^2 \rangle_q^* = m_{j,f,n}^2 + \gamma_{j,f,n} \). Therefore the source parameters are updated by computing an NMF on

\[
P_j = \left[ m_{j,f,n} + \gamma_{j,f,n} \right] \in \mathbb{R}^{F \times N} \text{ with the IS divergence. It can be done with the standard multiplicative update rules given in [3].}
\]

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

\[
C(a_{ij}) = -\frac{1}{2} \sum_{j=1}^J \sum_{i=1}^I \left[ \| x_i - \sum_{j=1}^J \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{j,f,n} T_{f,n} a_{ij} \|_2^2 + \sum_{j=1}^J \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \gamma_{j,f,n} \| T_{f,n} a_{ij} \|_2^2 \right].
\]

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

\[
\sum_{f=0}^{F-1} \sum_{n=0}^{N-1} m_{j,f,n}^T (x_i - \sum_{j'=1}^J \sum_{f'=0}^{F-1} \sum_{n'=0}^{N-1} m_{j',f,n} T_{f',n'} a_{ij'})
\]

\[
+ \sum_{j=1}^J \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \gamma_{j,f,n} \| T_{f,n} a_{ij} \|_2^2 \right].
\]

Let introduce the following definitions:

\( \triangleright e_{ij}(t) = x_i(t) - \sum_{j' \neq j} \hat{y}_{ij'}(t) \) for \( t = 0, ..., T - 1 \);
\( \triangleright T_M(\tau(k)) \): A symmetric Toeplitz matrix of size \( M \times M \) formed from the sequence \( \tau(k) \) \( k=0 \).
We can show that (26) can be re-written as:

\[
T_{Lk}\{\hat{r}^{(s)}_{ij}(k)\} + T_{nk}\{\sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \gamma_{j,f,s} \hat{a}^{(s)}_{fn}(k)\} = \hat{r}^{(s)}_{ij}, \quad (27)
\]

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

4. EXPERIMENTS

Our experiments are conducted from the audio tracks without effects provided by the Musical Audio Signal Separation (MASS) dataset [22]. We consider 8 stereo mixtures created by simulating mixing filters with the Roomsimove toolbox [23]. The mixtures duration ranges from 12 to 28 seconds. We considered several reverberation times\(^3\): \(T_{60} = 32, 64, 128, 256\text{ or } 512\text{ ms}\). It results in a total number of 40 mixtures. The number of sources varies between 3 and 5. The omnidirectional microphone spacing was set to 1 m, and the distance between the source and the center of the microphone pair to 2 m. The sources directions of arrival range from 45° to 45°. As the MASS dataset provides stereo sources, each one is first converted to mono, downsampled to 16 kHz and filtered with the associated RIRs to create a source image. We finally sum all the source images to create a mixture.

We compare our approach with the baseline method [8]\(^4\). Indeed, our work is comparable to [8] in the sense that it performs multichannel audio source separation using an NMF source model and a punctual convolutive mixture model. The key difference is that in this method, the authors infer the TF latent sources from TF observations (with an EM algorithm and using the STFT). Therefore they approximate the convolutive mixing process in the STFT domain assuming short mixing filters. In our work, the inference is done from the time-domain observations, the convolutive mixture model is thus exact. For both methods we use a half-overlapping TF analysis/synthesis sine window of 128 ms (2048 points at a sampling rate of 16 kHz). The NMF order is arbitrarily fixed to \(K_{j} = 10\) for all the sources. For both algorithms the parameters are initialized from oracle values computed on the true source signals and from the true mixing filters. Indeed, this work aims at comparing the best performance achievable by the two approaches. The (V)EM algorithms are run for 100 iterations from the oracle initializations. The PCG algorithms for the proposed method are run for 10 iterations. We evaluate the source separation performance in terms of reconstructed source images. We use standard energy ratios: the Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ratio (SAR) and source Image-to-Spatial distortion Ratio (ISR). These criteria expressed in decibels (dB) are defined in [24]. We used the BSS Eval Toolbox available at [25] to compute these measures.

\(^3\)The reverberation time is defined as the time it takes for the sound energy to decrease by 60 dB after extinction of the source.

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

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

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

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

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

5. CONCLUSIONS

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

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

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


