MULTICHANNEL AUDIO SOURCE SEPARATION WITH PROBABILISTIC REVERBERATION MODELING

Simon Leglaive, Roland Badeau, Gael Richard

To cite this version:
Simon Leglaive, Roland Badeau, Gael Richard. MULTICHANNEL AUDIO SOURCE SEPARATION WITH PROBABILISTIC REVERBERATION MODELING. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPA), Oct 2015, New Paltz, NY, United States. pp.5. hal-01219635

HAL Id: hal-01219635
https://inria.hal.science/hal-01219635
Submitted on 23 Oct 2015

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 WITH PROBABILISTIC REVERBERATION MODELING

Simon Leglaive, Roland Badeau and Gaël Richard
Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, France
<firstname>.<lastname>@telecom-paristech.fr

ABSTRACT
In this paper we show that considering early contributions of mixing filters through a probabilistic prior can help blind source separation in reverberant recording conditions. By modeling mixing filters as the direct path plus \( R - 1 \) reflections, we represent the propagation from a source to a mixture channel as an autoregressive process of order \( R \) in the frequency domain. This model is used as a prior to derive a Maximum A Posteriori (MAP) estimation of the mixing filters using the Expectation-Maximization (EM) algorithm. Experimental results over reverberant synthetic mixtures and live recordings show that MAP estimation with this prior provides better separation results than a Maximum Likelihood (ML) estimation.

Index Terms— Blind audio source separation, Underdetermined convolutive mixtures, Probabilistic prior, MAP estimation, EM algorithm.

1. INTRODUCTION

Blind audio source separation consists in estimating the \( J \) source signals \( s_j(t), j = 1, ..., J \) from \( T \) observed mixtures \( x_i(t), i = 1, ..., I \). If there are \( J < I \), the problem is under-determined. For data recorded in reverberant conditions, the propagation from a source to a microphone can be represented by a filter, the mixing model is therefore convolutive:

\[
x_i(t) = \sum_{j=1}^{J} [a_{ij} * s_j](t) + b_i(t),
\]

where \( * \) denotes the convolution product and \( a_{ij}(t) \) is the impulse response of the mixing filter between source \( j \) and microphone \( i \). \( b_i(t) \) is an additive noise.

Most approaches for source separation work in the time/frequency (TF) domain. Using the Short-Time Fourier Transform (STFT) and assuming short mixing filters, the convolutive mixing model is approximated by an instantaneous mixing at each TF point \((f, n), f = 1, ..., F, n = 1, ..., N\):

\[
x_{i,fn} = \sum_{j=1}^{J} a_{ij,f} s_{j,fn} + b_{i,fn},
\]

where \( x_{i,fn}, s_{j,fn} \) and \( b_{i,fn} \) are the STFTs of the corresponding time signals and \( a_{ij,f} \) is the frequency response of filter \( a_{ij}(t) \). Equation (2) can be rewritten in matrix form:

\[
x_f = A_f s_f + b_f
\]

with \( x_f = [x_{1,fn}, ..., x_{I,fn}]^T \), \( s_f = [s_{1,fn}, ..., s_{J,fn}]^T \), \( b_f = [b_{1,fn}, ..., b_{I,fn}]^T \) and \( A_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I \times J} \). Operator \((\cdot)^T\) denotes transpose operation.

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).

In order to improve separation performance, an increasing number of methods guide the separation by considering deterministic constraints or probabilistic priors over the source or mixing parameters \([1, 2]\). For example, one can assume that the mixing process is close to the anechoic scenario, in which the propagation between source \( j \) and channel \( i \) corresponds to a delay \( \tau_{ij} \) and an attenuation \( \rho_{ij} \). In this case, as illustrated on figure 1, only the direct path is considered for the mixing filter; early echoes and late reverberation are ignored and the transfer function \( a_{ij,f} \) in (2) is approximated by \( a_{ij,f} = \rho_{ij} b_{ij,f} \) where \( b_{ij,f} = e^{-j\pi\tau_{ij}} \). In \([3]\), the proximity between \( a_{ij,f} \) and \( d_{ij,f} \) is used to solve the permutation ambiguity in a separation approach based on independent component analysis in the frequency domain. This proximity is also used in methods based on TF masking that exploit the spatial diversity of sources and their sparsity in the TF plan \([3, 4]\). In \([5]\), Duong et al. study various mixing models, from the rank-1 anechoic model to the full-rank spatial covariance model. They show that using an anechoic model for the mixing \( a_{ij,f} = d_{ij,f} \) leads to poor performance compared to the convolutive model (2) where \( a_{ij,f} \) is unconstrained. In \([6]\) the authors consider Inverse-Wishart and Gaussian priors over spatial covariance matrices to derive a MAP estimation of the parameters with the EM algorithm. The effectiveness of this approach is shown in a semi-supervised context where the source positions and some room characteristics are known. A complex Wishart prior over the inverses of spatial covariance matrices is introduced in \([7]\). Geometrically calculated or pre-measured steering vectors are considered as hyper-parameters for this prior.

In this work, we will consider a blind scenario. We propose to keep the convolutive model but to consider the early contributions of mixing filters through a probabilistic prior over the mixing matrix \( A_f \). Then, we estimate \( A_f \) in the MAP sense with the EM algorithm, adapting the multichannel separation method defined in \([8]\) and based on a ML estimation of the parameters.

Section 2 briefly introduces the approach \([8]\) which is our baseline. Section 3 shows that considering early contributions of mixing filters can lead to an autoregressive model in the frequency domain. MAP estimation of the mixing parameters is derived in Section 4. Section 5 describes the experimental results. We present our conclusions in section 6.

Figure 1: Schematic illustration of a reverberant mixing filter
2. MODEL AND MAXIMUM LIKELIHOOD ESTIMATION

In this section, we present the model defined by Ozerov and Févotte in [8] and we introduce the ML estimation of the parameters. The STFT of source \( j \) is modeled as a sum of \( \#K_j \) latent Gaussian components, \( \{K_j\}_{j=1} \) being a non-trivial partition of \( K = \{1, ..., K\} \) with \( K \geq J \),

\[
s_{j,f,n} = \sum_{k \in K_j} c_{k,f,n} \quad \text{with} \quad c_{k,f,n} \sim N_c(0, w_{jk} h_{kn})
\]  

(4)

where \( w_{jk}, h_{kn} \in \mathbb{R}^+ = [0, +\infty] \). \( N_c(\mu, \Sigma) \) is the multivariate complex Gaussian distribution, with probability density function:

\[
N_c(x; \mu, \Sigma) = \frac{1}{\det(\pi \Sigma)} \exp[-(x - \mu)^\top \Sigma^{-1}(x - \mu)],
\]  

(5)

where \((\cdot)^\top\) denotes conjugate transpose and \(\det(\cdot)\) is the determinant.

Components \( c_{k,f,n} \) are assumed mutually independent and individually independent over frequency \( f \) and frame \( n \), consequently

\[
s_{j,f,n} \sim N_c(0, \sum_{k \in K_j} w_{jk} h_{kn})
\]  

(6)

The source model (6) is related to Nonnegative Matrix Factorization (NMF). Indeed, it has been shown in [9] that ML estimation of the variance parameters in this model is equivalent to NMF of the source power spectrogram using the Itakura-Saito divergence.

In equation (3), \( b_{f,n} \) is a stationary white Gaussian noise, isotropic in each sub-band:

\[
b_{f,n} \sim N_c(0, \Sigma_{b,f} = \sigma_b^2 I),
\]  

(7)

where \( I_n \) is the identity matrix of size \( n \).

Let \( \theta = \{A, W, H, \Sigma_b, f\} \) be the set of parameters, with \( A \) the \( I \times J \) matrix with entries \( a_{ij,f} \), \( W \) the \( F \times K \) matrix with entries \( w_{jk} \), \( H \) the \( K \times N \) matrix with entries \( h_{kn} \) and \( \Sigma_b \) the column vector of size \( F \) with entries \( \sigma_b^2 \). The ML estimation of parameters \( \theta \) consists in maximizing the log-likelihood \( \log p(X|\theta) \), by means of the EM algorithm in the specific case of the exponential family. The complete data are \( \{X, C\} \), with \( X \) and \( C \) the two \( I \times F \times N \) and \( K \times F \times N \) tensors of entries \( x_{i,f,n} \) and \( c_{k,f,n} \) respectively. The E-step consists in computing the conditional expectation of the sufficient natural statistics \( R_{ss,f} = \frac{1}{N} \sum_{n} x_{i,f,n} x_{j,f,n}^\top \) and \( R_{sR,f} = \frac{1}{N} \sum_{n} x_{i,f,n} y_{f,n} \) and \( u_{k,f,n} = |c_{k,f,n}|^2 \). The M-step consists in re-estimating the parameters \( \theta \) by minimizing of \( Q_1(\theta|\theta') \), with \( \theta' \) the set of parameters estimated at the previous iteration:

\[
Q_1(\theta|\theta') = -E_{C|X,\theta'}[\log p(X, C|\theta)]
\]

\[
\leq \sum_{f,n} \left[ \log(w_{jk} h_{kn}) + \sum_{k} \frac{\tilde{u}_{k,f,n}}{w_{jk} h_{kn}} \right]
\]

\[
+ N \sum_{f} \left[ \log(\det(\Sigma_{b,f})) + \operatorname{tr}(\Sigma_{b,f}^{-1} R_{ss,f} - \Sigma_{b,f}^{-1} A_f H_{ss,f}^\top)
\right. 
\]

\[
- \left. \Sigma_{b,f}^{-1} H_{ss,f} A_f^\top + \Sigma_{b,f}^{-1} A_f H_{ss,f}^\top) \right]
\]

(8)

\( \equiv \) denotes equality up to a constant, \( \hat{\cdot} \) indicates the conditional expectation of the sufficient natural statistics computed during the E-step and \( \operatorname{tr}(\cdot) \) is the trace. The complete EM algorithm is derived in [8]. After the parameter estimation, the sources are reconstructed by Wiener filtering.

3. A PRIOR FOR MODELING EARLY CONTRIBUTIONS

If we consider that a mixing filter only contains \( R \) early contributions, the propagation from source \( j \) to microphone \( i \) can be modeled by \( R \) attenuation coefficients \( \rho_{kj} \) and delays \( \tau_{kj} \), \( k = 0, ..., R - 1 \), such that \( a_{ij,f} \) is approximated by:

\[
d_{ij,f} = \sum_{k=0}^{R-1} \rho_{kj} \delta_{kj}^{f} \quad \text{with} \quad \delta_{kj}^{f} = e^{-j2\pi \tau_{kj}}.
\]  

(9)

It follows that \( \{d_{ij,f}\}_{f=0}^{R+1,...,F} \) satisfies a recursive equation of the form (see, e.g. [10]):

\[
\sum_{r=0}^{R} \varphi_{ij}^{r} d_{ij,f-r} = 0.
\]  

(10)

Finally, we consider that \( \{a_{ij,f}\}_{f=0}^{R+1,...,F} \) follows model (10) up to a certain deviation \( b_{ij,f} \) such that,

\[
R \sum_{r=0}^{R} \varphi_{ij}^{r} a_{ij,f-r} = b_{ij,f}
\]  

(11)

where \( b_{ij,f} \) is a complex white Gaussian noise with variance \( \sigma_{ij}^2 \). So it is an independent and identically distributed process following the complex Gaussian distribution with probability density function:

\[
N_c(x; 0, \sigma^2) = \frac{1}{\pi \sigma^2} \exp \left( -\frac{|x|^2}{\sigma^2} \right).
\]  

(12)

From (11), we see that \( \{a_{ij,f}\}_{f} \) is an autoregressive process of order \( R \). Without loss of generality, we force the first prediction coefficient \( \varphi_{ij} \) to be equal to one for all \( i = 1, ..., J, j = 1, ..., J \). Finally, we can write the equations of the sequence of \( a_{ij,f} \) for \( f = 1, ..., F \):

\[
p\{a_{ij,f}\}_{f} = p(a_{ij,1} ... a_{ij,r}) \prod_{f=R+1}^{F} p(a_{ij,f} | a_{ij,f-1} ... a_{ij,f-R})
\]  

(13)

with

\[
p(a_{ij,f} | a_{ij,f-1} ... a_{ij,f-R}) = \frac{1}{\pi \sigma_{ij}^2} \exp \left[ -\frac{\sum_{r=0}^{R} \varphi_{ij}^{r} a_{ij,f-r}^2}{\sigma_{ij}^2} \right].
\]  

(14)

4. MAXIMUM A POSTERIORI ESTIMATION

MAP estimation of the parameters \( \theta \) consists in maximizing the a posteriori log-probability \( \log p(\theta|X) \). By means of the EM algorithm, we obtain the same procedure as introduced in section 2 and detailed in [8], except for the update of the mixing matrix \( A_f \) at the M-step. Indeed, we have to minimize the opposite of the conditional expectation of the a posteriori log-probability of complete data with respect to \( A_f \). Using Bayes’ rule, this is equivalent to minimizing

\[
Q_2(\theta|\theta') = Q_1(\theta|\theta') - \log p(A)
\]  

(15)

where \( Q_1(\theta|\theta') \) is defined in equation (8). By considering non-informative priors on \( a_{ij,f} \) for \( f = 1, ..., R \) and assuming that mixing filters are independent over \( i \) and \( j \), we compute \( -\log p(A) \) from (13) and (14) and we obtain:

\[
-\log p(A) \equiv \sum_{i,j}(F-R) \log \sigma_{ij}^2 + \frac{1}{\sigma_{ij}^2} \sum_{r=0}^{R} \sum_{i,j} \varphi_{ij}^{r} a_{ij,f-r}^2.
\]  

(16)
Let $\mathbf{S} = [\frac{1}{T}]_{ij} \in (\mathbb{R}^+)^{J \times J}$ and $\Phi_r = [\varphi_{rij}]_{ij} \in \mathbb{C}^{J \times J}$, $r = 0, ..., R$. The new estimation of $\mathbf{A}_f$ in the M-step is obtained by zeroing the gradient of $Q_2(\theta^0)$ with respect to $\mathbf{A}_f$. We obtain for $f = 1, ..., R$,

$$\text{vec}(\mathbf{A}_f) = \left[ \frac{8}{T} \text{vec}(\mathbf{S} \mathbf{R}_{ij} F_{ij})^T \right]^{(f)}$$

and

$$\text{vec} \left( \mathbf{N} \mathbf{R}_{xs,f} - \mathbf{S}_f \sum_{r=0}^{R} \Phi_r \sum_{n=0}^{N} \Phi_{n,n} \mathbf{A}_{f+r-n} \right);$$

for $f = R + 1, ..., F - R$,

$$\text{vec}(\mathbf{A}_f) = \left[ \frac{8}{T} \text{vec}(\mathbf{S} \mathbf{R}_{ij} F_{ij})^T \right]^{(f)}$$

and for $f = F - R + 1, ..., F$,

$$\text{vec}(\mathbf{A}_f) = \left[ \frac{8}{T} \text{vec}(\mathbf{S} \mathbf{R}_{ij} F_{ij})^T \right]^{(f)}$$

where $(\cdot)^*$ denotes complex conjugate, $\otimes$ and $\bullet$ are respectively the Kronecker product, and $|$ is the element-wise modulus. $\mathbf{1}_{mn}$ is a column vector of length $m \times n$ whose entries are all equal to 1.

We also have to minimize $-\log p(\mathbf{A})$ with respect to the hyperparameters of the prior for all $i$ and $j$, that are $\sigma_{ij}^2$ and $\varphi_{ij} = [\varphi_{0ij}, ..., \varphi_{Rij}]^T$ under the constraint $\varphi_{0ij} = 1$. For $\sigma_{ij}^2$ we obtain

$$\sigma_{ij}^2 = \frac{1}{(F - R)} \sum_{r=R+1}^{R} \left| \varphi_{rij} - \sigma_{ij}^{-2} \right|^2.$$  

We compute $\varphi_{ij}$ using Lagrange multipliers and rewriting the prior in equation (16) as

$$-\log p(\mathbf{A}) = \sum_{i,j} \left( (F - R) \log \sigma_{ij} + \frac{1}{\sigma_{ij}^2} (\mathbf{A}_{ij} \varphi_{ij} - 1)^H (\mathbf{A}_{ij} \varphi_{ij}) \right)$$

where

$$\mathbf{A}_{ij} = \begin{pmatrix} a_{ij,R+1} & a_{ij,R} & \cdots & a_{ij,1} \\ a_{ij,R+2} & a_{ij,R+1} & \cdots & a_{ij,2} \\ \vdots & \vdots & & \vdots \\ a_{ij,F} & a_{ij,F-1} & \cdots & a_{ij,F-R} \end{pmatrix}.$$  

We obtain, with $\mathbf{e}_1 = [1, 0, ..., 0]^T$ a column vector of length $R + 1$,

$$\varphi_{ij} = \frac{1}{\mathbf{e}_1^T (\mathbf{A}_{ij}^H \mathbf{A}_{ij})^{-1} \mathbf{e}_1} (\mathbf{A}_{ij}^H \mathbf{A}_{ij})^{-1} \mathbf{e}_1.$$  

The algorithm for MAP estimation is thus quite similar to the algorithm detailed in [8]. We only modify the M-step for the estimation of $\mathbf{A}_f$, in order to take the prior into account. Contrary to the approach in [6], the separation is here weakly guided. Indeed, the hyperparameters of the prior are not fixed but have to be estimated.

5. EXPERIMENTS

In this section, we compare the separation performance of the algorithms with and without prior, which correspond respectively to the MAP and ML estimations. In order to evaluate the performance, we use 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 [11]. We used the BSS Eval Toolbox available at [12]. We run our experiments on synthetic and live-recorded convolutive stereo mixtures with two microphone setups. For each type of mixture (synthetic or live-recorded) and microphone setup, we used two 10 second-long excerpts sampled at 16KHz involving three musical sources. We thus have a total of 8 excerpts. It corresponds to the musical development data for the 2007 Stereo Audio Source Separation Evaluation Campaign [13]. Live recordings were acquired by playing the source signals through loudspeakers in a room with a reverberation time of 250 ms. The recording setup consisted in a pair of omnidirectional microphones with 1 m or 5 cm spacing. Synthetic convolutive mixtures were obtained by filtering the sources with synthetic filters, corresponding to the same setup as live-recorded mixtures. The distances between the sources and the center of the microphone pair vary between 80 cm and 1.2 m and the source angles of arrivals vary between -60° and +60° with a minimal spacing of 15°. We used a 128 ms half-overlapping sine window to compute the STFTs.

The EM algorithm is very sensitive to the parameter initialization. In order to obtain satisfactory separation results, we have to provide a “good initialization”. As in [5], we choose to initialize the mixing system $\mathbf{A}$ using the hierarchical clustering-based algorithm presented in [14]. This method relies on the spatial diversity of sources and their sparsity in the TF plan. It assumes that the mixture STFT coefficients $x_{f,n}$ cluster around the direction of the associated mixing vector $[a_{ij,f}, ..., a_{ij,F}]^T$ in the time frames $n$ where the $j$th source is predominant. Once the frequency-dependent mixing matrix $\mathbf{A}_f$ is estimated by hierarchical clustering, we perform a first estimation of the sources via projection of the mixture over the source directions and binary masking in the TF plan. The source parameters $\mathbf{W}$ and $\mathbf{H}$ are then initialized by NMF of the power spectrograms of the separated sources. We perform 100 iterations of the multiplicative algorithm with Kullback-Leibler divergence [15]. As in [8], the noise parameters $\sigma_r^2$ in (7) are initialized to the average channel empirical variance in each frequency band divided by 100, i.e., $100\sigma_r^2 = \sum_{f=0}^{F-1} |X_{f,n}|^2 / (IN)$. We use $\#K_j = 4$ latent components for each source. For MAP and ML estimations, we run 400 iterations of the EM algorithm (our implementation is based on [16]). Each separation is evaluated with and without prior, from the same initialization. For the MAP estimation, we obtain satisfactory results with an autoregressive model of order $R = 6$. The results are presented in table 1 in the columns “ML” and “MAP”. For each type of mixture and spacing between the microphones, the results are averaged over all the separated sources for 2 excerpts.

We see that for 1 m spacing between the microphones, MAP estimation improves all the measures compared to ML estimation, especially for synthetic mixtures. For 5 cm spacing, the superiority of MAP estimation is not so clear. With this setup the two algorithms provide very similar results and for synthetic mixtures, MAP estimation performs worse. By looking at the evolution of parameters and measures along the iterations, we observed that MAP and ML estimations evolved in a similar way. This may be due to very close trajectories in the parameter search space. It means that the
prior is not strong enough to significantly influence the estimation of $A_{ij}$ through the minimization of $Q_2(\theta | \theta^{'})$ at M-step. We see in equation (16) that the hyper-parameters $\sigma^2_{ij}$ control the strength of the prior. Indeed, as they tend to increase, the contribution of mixing filters in (16) decreases, so the minimization of $Q_2(\theta | \theta^{'})$ in equation (15) with respect to $A_{ij}$ tends to be equivalent to the minimization of $Q_1(\theta | \theta^{'})$; MAP estimation turns to ML estimation. A way to control the strength of the prior is then to constrain the value of the variances $\sigma^2_{ij}$. To illustrate this principle, we represent on figure 2 the average SDR over iterations for one excerpt of the database. For this example we set the variances $\sigma^2_{ij}$ for all $i, j$ at 500, except at iterations 200 and 300 where it is fixed at 0.1 during 10 iterations.

With this setting, the search space for the mixing matrix is more constrained at the beginning of the EM algorithm than at the end. We present in the column "MAP with CSP" of table 1 the separation performance over the whole database with this Control of the Strength of the Prior (CSP). We see that for each type of mixture and microphone spacing, all evaluation measures are improved compared to ML estimation. We also notice the effectiveness of this strategy compared to MAP estimation with unconstrained variances, except in terms of average SDR for synthetic mixtures with 1 m spacing between microphones.

### 6. CONCLUSIONS
By considering early contributions of mixing filters, we presented in this paper a new probabilistic prior to estimate mixing parameters in the MAP sense with the EM algorithm. This prior is based on an autoregressive modeling of mixing filters in the frequency domain. Experimental results over reverberant synthetic mixtures and live recordings for two microphone spacings have shown the superiority of this approach compared to ML estimation. This prior acts as a constraint on the mixing matrix, its strength is controlled through the hyper-parameters $\sigma^2_{ij}$. We showed that source separation performance is improved by setting a strong prior at the beginning of the EM algorithm and progressively weakening it.

Even if this prior comes from the consideration of early contributions in mixing filters, the autoregressive model also fits late reverberation. We believe that we could improve the effectiveness of the prior by better modeling the early contributions while keeping the autoregressive model for late reverberation. Moreover, such an approach for modeling diffuse part of room impulse responses would be consistent with Schroeder’s results about frequency correlation functions of room frequency responses [17]. Spatial correlations will also have to be considered. We will investigate this approach in future works.
7. REFERENCES


