MULTICHANNEL AUDIO SOURCE SEPARATION WITH PROBABILISTIC REVERBERATION MODELING

Simon Leglaive, Roland Badeau, Gaël Richard

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, Under-determined convolutive mixtures, Probabilistic prior, MAP estimation, EM algorithm.

1. INTRODUCTION

Blind audio source separation consists in estimating the J source signals $s_j(t)$, $j = 1, ..., J$, $t = 1, ..., T$ from I observed mixtures $x_i(t)$, $i = 1, ..., I$. If there are less mixtures than sources ($I < J$), 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:

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

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), $f = 1, ..., F$, $n = 1, ..., N$:

$$x_{i,fn} = \sum_{j=1}^{J} a_{ij,f} s_{j,fn} + b_{i,fn}, \tag{2}$$

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:

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_{fn} \tag{3}$$

with $\mathbf{x}_{fn} = [x_{1,fn}, ..., x_{I,fn}]^T$, $\mathbf{s}_{fn} = [s_{1,fn}, ..., s_{J,fn}]^T$, $\mathbf{b}_{fn} = [b_{1,fn}, ..., b_{I,fn}]^T$ and $\mathbf{A}_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I \times J}$. 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).

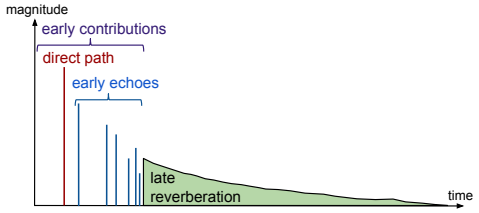

Figure 1: Schematic illustration of a reverberant mixing filter

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 τ_{ij} and an attenuation ρ_{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 $d_{ij,f} = \rho_{ij}\delta_{ij}^f$ where $\delta_{ij} = e^{-j2\pi\tau_{ij}}$. 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 $\mathbf{A}_f$. Then, we estimate $\mathbf{A}_f$ 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.

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 $\#\mathcal{K}_j$ latent Gaussian components, $\{\mathcal{K}_j\}_{j=1}^{J}$ being a non-trivial partition of $\mathcal{K} = \{1, ..., K\}$ with $K \geq J$,

$$s_{j,fn} = \sum_{k \in \mathcal{K}_j} c_{k,fn} \quad \text{with} \quad c_{k,fn} \sim \mathcal{N}_c(0, w_{fk}h_{kn}) \tag{4}$$

where $w_{fk}, h_{kn} \in \mathbb{R}^+ = [0, +\infty[$. $\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate complex Gaussian distribution, with probability density function:

$$N_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\det(\pi\boldsymbol{\Sigma})} \exp[-(\mathbf{x} - \boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})], \tag{5}$$

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

Components $c_{k,fn}$ are assumed mutually independent and individually independent over frequency f and frame n, consequently

$$s_{j,fn} \sim \mathcal{N}_c\Big(0, \sum_{k \in \mathcal{K}_j} w_{fk}h_{kn}\Big). \tag{6}$$

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), $\mathbf{b}_{fn}$ is a stationary white Gaussian noise, isotropic in each sub-band:

$$\mathbf{b}_{fn} \sim \mathcal{N}_c(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b},f} = \sigma_f^2 \mathbf{I}_I), \tag{7}$$

where $\mathbf{I}_n$ is the identity matrix of size n.

Let $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{W}, \mathbf{H}, \boldsymbol{\Sigma}_\mathbf{b}\}$ be the set of parameters, with $\mathbf{A}$ the $I \times J \times F$ tensor with entries $a_{ij,f}$, $\mathbf{W}$ the $F \times K$ matrix with entries w_{fk}, $\mathbf{H}$ the $K \times N$ matrix with entries h_{kn} and $\boldsymbol{\Sigma}_\mathbf{b}$ the column vector of size F with entries σ_f^2. The ML estimation of parameters $\boldsymbol{\theta}$ consists in maximizing the log-likelihood $\log p(\mathbf{X}|\boldsymbol{\theta})$, by means of the EM algorithm in the specific case of the exponential family. The complete data are $\{\mathbf{X}, \mathbf{C}\}$, with $\mathbf{X}$ and $\mathbf{C}$ the two $I \times F \times N$ and $K \times F \times N$ tensors of entries $x_{i,fn}$ and $c_{k,fn}$ respectively. The E-step consists in computing the conditional expectation of the sufficient natural statistics $\mathbf{R}_{\mathbf{xx},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{x}_{fn}^H$, $\mathbf{R}_{\mathbf{xs},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{s}_{fn}^H$, $\mathbf{R}_{\mathbf{ss},f} = \frac{1}{N}\sum_n \mathbf{s}_{fn}\mathbf{s}_{fn}^H$ and $u_{k,fn} = |c_{k,fn}|^2$. The M-step consists in re-estimating the parameters $\boldsymbol{\theta}$ by minimization of $Q_1(\boldsymbol{\theta}|\boldsymbol{\theta}')$, with $\boldsymbol{\theta}'$ the set of parameters estimated at the previous iteration:

$$\begin{aligned} Q_1(\boldsymbol{\theta}|\boldsymbol{\theta}') &= -\mathbb{E}_{\mathbf{C}|\mathbf{X},\boldsymbol{\theta}'}[\log p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta})] \\ &\stackrel{c}{=} \sum_{f,n}\Big[\sum_k \log(w_{fk}h_{kn}) + \sum_k \frac{\hat{u}_{k,fn}}{w_{fk}h_{kn}}\Big] \\ &+ N\sum_f \Big[\log(\det(\boldsymbol{\Sigma}_{\mathbf{b},f})) + \mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\hat{\mathbf{R}}_{\mathbf{xx},f} - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\hat{\mathbf{R}}_{\mathbf{xs},f}^H\right. \\ &\left. - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\hat{\mathbf{R}}_{\mathbf{xs},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\hat{\mathbf{R}}_{\mathbf{ss},f}\mathbf{A}_f^H\right)\Big]. \end{aligned} \tag{8}$$

$\stackrel{c}{=}$ denotes equality up to a constant, $\hat{(\cdot)}$ indicates the conditional expectation of the sufficient natural statistics computed during the E-step and $\mathrm{tr}(\cdot)$ 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 attenuations ρ_{kij} and delays τ_{kij}, $k = 0, ..., R-1$, such that $a_{ij,f}$ is approximated by:

$$d_{ij,f} = \sum_{k=0}^{R-1} \rho_{kij}\delta_{kij}^f \quad \text{with} \quad \delta_{kij} = e^{-j2\pi\tau_{kij}}. \tag{9}$$

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

$$\sum_{r=0}^{R} \varphi_{rij} d_{ij,f-r} = 0. \tag{10}$$

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

$$\sum_{r=0}^{R} \varphi_{rij} a_{ij,f-r} = b_{ij,f} \tag{11}$$

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

$$N_c(x; 0, \sigma^2) = \frac{1}{\pi\sigma^2} \exp\Big(\frac{-|x|^2}{\sigma^2}\Big). \tag{12}$$

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 φ_{0ij} to be equal to one for all $i = 1, .., I$, $j = 1, ..., J$. Finally, we can write the probability of the sequence of $a_{ij,f}$ for $f = 1, ..., F$:

$$p(\{a_{ij,f}\}_f) = p(a_{ij,1}...a_{ij,R})\prod_{f=R+1}^{F} p(a_{ij,f}|a_{ij,f-1}...a_{ij,f-R}) \tag{13}$$

with

$$p(a_{ij,f}|a_{ij,f-1}...a_{ij,f-R}) = \frac{1}{\pi\sigma_{ij}^2}\exp\left[\frac{-\Big|\sum_{r=0}^{R}\varphi_{rij}a_{ij,f-r}\Big|^2}{\sigma_{ij}^2}\right]. \tag{14}$$

4. MAXIMUM A POSTERIORI ESTIMATION

MAP estimation of the parameters $\boldsymbol{\theta}$ consists in maximizing the a posteriori log-probability $\log p(\boldsymbol{\theta}|\mathbf{X})$. 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 $\mathbf{A}_f$ 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 $\mathbf{A}_f$. Using Bayes' rule, this is equivalent to minimizing

$$Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}') = Q_1(\boldsymbol{\theta}|\boldsymbol{\theta}') - \log p(\mathbf{A}) \tag{15}$$

where $Q_1(\boldsymbol{\theta}|\boldsymbol{\theta}')$ 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 $-\log p(\mathbf{A})$ from (13) and (14) and we obtain:

$$-\log p(\mathbf{A}) \stackrel{c}{=} \sum_{i,j}\Big((F-R)\log\sigma_{ij}^2 + \frac{1}{\sigma_{ij}^2}\sum_{f=R+1}^{F}\Big|\sum_{r=0}^{R}\varphi_{rij}a_{ij,f-r}\Big|^2\Big). \tag{16}$$

Let $\mathbf{S} = [\frac{1}{\sigma_{ij}^2}]_{ij} \in (\mathbb{R}^+)^{I\times J}$ and $\mathbf{\Phi}_r = [\varphi_{rij}]_{ij} \in \mathbb{C}^{I\times J}$, $r = 0, ..., R$. The new estimation of $\mathbf{A}_f$ in the M-step is obtained by zeroing the gradient of $Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}')$ with respect to $\mathbf{A}_f$. We obtain for $f = 1, ..., R$,

$$\mathrm{vec}(\mathbf{A}_f) = \Big[N\hat{\mathbf{R}}_{\mathbf{ss},f}^T \otimes \mathbf{I}_I + (\mathbf{I}_J \otimes \mathbf{\Sigma}_{\mathbf{b},f}) \bullet \Big(\mathbf{1}_{IJ}\mathrm{vec}(\mathbf{S} \bullet \sum_{r=R-f+1}^{R} |\mathbf{\Phi}_r|^{\cdot 2})^T\Big)\Big]^{-1}$$
$$\times \mathrm{vec}\left(N\hat{\mathbf{R}}_{\mathbf{xs},f} - \mathbf{\Sigma}_{\mathbf{b},f}(\mathbf{S} \bullet \sum_{r=R-f+1}^{R} \mathbf{\Phi}_r^* \bullet \sum_{\substack{n=0\\n\neq r}}^{R} \mathbf{\Phi}_n \bullet \mathbf{A}_{f+r-n})\right); \quad (17)$$

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

$$\mathrm{vec}(\mathbf{A}_f) = \Big[N\hat{\mathbf{R}}_{\mathbf{ss},f}^T \otimes \mathbf{I}_I + (\mathbf{I}_J \otimes \mathbf{\Sigma}_{\mathbf{b},f}) \bullet \Big(\mathbf{1}_{IJ}\mathrm{vec}(\mathbf{S} \bullet \sum_{r=0}^{R} |\mathbf{\Phi}_r|^{\cdot 2})^T\Big)\Big]^{-1}$$
$$\times \mathrm{vec}\left(N\hat{\mathbf{R}}_{\mathbf{xs},f} - \mathbf{\Sigma}_{\mathbf{b},f}(\mathbf{S} \bullet \sum_{r=0}^{R} \mathbf{\Phi}_r^* \bullet \sum_{\substack{n=0\\n\neq r}}^{R} \mathbf{\Phi}_n \bullet \mathbf{A}_{f+r-n})\right); \quad (18)$$

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

$$\mathrm{vec}(\mathbf{A}_f) = \Big[N\hat{\mathbf{R}}_{\mathbf{ss},f}^T \otimes \mathbf{I}_I + (\mathbf{I}_J \otimes \mathbf{\Sigma}_{\mathbf{b},f}) \bullet \Big(\mathbf{1}_{IJ}\mathrm{vec}(\mathbf{S} \bullet \sum_{r=0}^{F-f} |\mathbf{\Phi}_r|^{\cdot 2})^T\Big)\Big]^{-1}$$
$$\times \mathrm{vec}\left(N\hat{\mathbf{R}}_{\mathbf{xs},f} - \mathbf{\Sigma}_{\mathbf{b},f}(\mathbf{S} \bullet \sum_{r=0}^{F-f} \mathbf{\Phi}_r^* \bullet \sum_{\substack{n=0\\n\neq r}}^{R} \mathbf{\Phi}_n \bullet \mathbf{A}_{f+r-n})\right), \quad (19)$$

where $(\cdot)^*$ denotes complex conjugate, $\otimes$ and $\bullet$ are respectively the Kroenecker and the element-wise product, $\mathrm{vec}(\cdot)$ concatenates the columns of a matrix into a single column vector and $|\cdot|^{\cdot 2}$ is the element-wise squared 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 hyper-parameters of the prior for all i and j, that are σ_{ij}^2 and $\boldsymbol{\varphi}_{ij} = [\varphi_{0ij}, ..., \varphi_{Rij}]^T$ under the constraint $\varphi_{0ij} = 1$. For σ_{ij}^2 we obtain

$$\sigma_{ij}^2 = \frac{1}{(F-R)} \sum_{f=R+1}^{F} \Big| \sum_{r=0}^{R} \varphi_{rij} a_{ij,f-r} \Big|^2. \quad (20)$$

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

$$-\log p(\mathbf{A}) \stackrel{c}{=} \sum_{i,j} \Big((F-R)\log\sigma_{ij}^2 + \frac{1}{\sigma_{ij}^2}(\mathbf{\Lambda}_{ij}\boldsymbol{\varphi}_{ij})^H(\mathbf{\Lambda}_{ij}\boldsymbol{\varphi}_{ij})\Big) \quad (21)$$

$$\text{where } \mathbf{\Lambda}_{ij} = \begin{pmatrix} a_{ij,R+1} & a_{ij,R} & \cdots & a_{ij,1} \\ a_{ij,R+2} & a_{ij,R+1} & \cdots & a_{ij,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{ij,F} & a_{ij,F-1} & \cdots & a_{ij,F-R} \end{pmatrix}. \quad (22)$$

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

$$\boldsymbol{\varphi}_{ij} = \frac{1}{\mathbf{e}_1^T(\mathbf{\Lambda}_{ij}^H\mathbf{\Lambda}_{ij})^{-1}\mathbf{e}_1}(\mathbf{\Lambda}_{ij}^H\mathbf{\Lambda}_{ij})^{-1}\mathbf{e}_1. \quad (23)$$

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 hyper-parameters 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 $\mathbf{x}_{fn}$ cluster around the direction of the associated mixing vector $[a_{1j,f}, ..., a_{Ij,f}]^T$ in the time frames n where the jth 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 σ_f^2 in (7) are initialized to the average channel empirical variance in each frequency band divided by 100, *i.e.*, $100\sigma_f^2 = \sum_{i,n} |x_{i,fn}|^2/(IN)$. We use $\#\mathcal{K}_j = 4$ 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

mixture	live recordings						synthetic mixtures					
microphone spacing	1 m			5 cm			1 m			5 cm		
estimation	ML	MAP	MAP with CSP	ML	MAP	MAP with CSP	ML	MAP	MAP with CSP	ML	MAP	MAP with CSP
SDR	0.36	0.60	**0.87**	0.70	0.75	**0.87**	-1.68	**-0.02**	-0.26	0.26	0.25	**1.05**
SIR	-0.43	0.25	**0.96**	-0.19	-0.09	**-0.06**	-0.37	-0.29	**2.80**	1.69	1.67	**2.62**
SAR	5.87	6.68	**7.19**	5.87	6.00	**6.36**	8.75	9.65	**9.74**	7.97	7.83	**8.52**
ISR	3.61	3.92	**4.20**	4.73	4.80	**4.88**	1.59	3.65	**3.93**	3.62	3.59	**4.58**

Table 1: Results of separation averaged over all the sources for the 2 excerpts per type of mixture and spacing between the microphones.

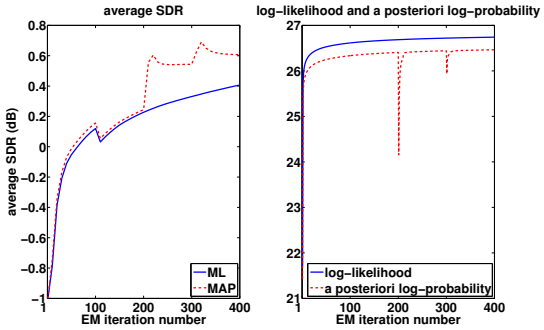

Figure 2: Average SDR, log-likelihood and a posteriori log-probability over iterations. σ_{ij}^2 is fixed for all i, j at 500, except at iterations 200 and 300 where it is fixed at 0.1 during 10 iterations.

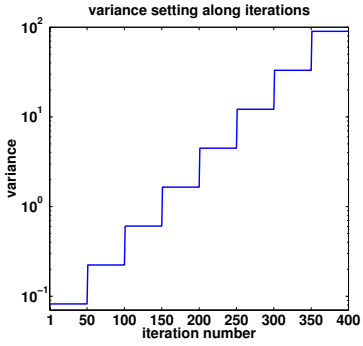

Figure 3: Strategy to control the strength of the prior through the setting of σ_a^2 according to equation (24).

prior is not strong enough to significantly influence the estimation of $\mathbf{A}_f$ through the minimization of $Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}')$ at M-step. We see in equation (16) that the hyper-parameters σ_{ij}^2 control the strength of the prior. Indeed, as they tend to increase, the contribution of mixing filters in (16) decreases, so the minimization of $Q_2(\boldsymbol{\theta}|\boldsymbol{\theta}')$ in equation (15) with respect to $\mathbf{A}_f$ tends to be equivalent to the minimization of $Q_1(\boldsymbol{\theta}|\boldsymbol{\theta}')$; MAP estimation turns to ML estimation. A way to control the strength of the prior is then to constrain the value of the variances σ_{ij}^2. 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 σ_{ij}^2 for all i, j at 500, except at iterations 200 and 300 where it is fixed at 0.1 during 10 iterations. We observe that during the 200 first iterations, with a high variance, MAP and ML estimations lead to the same performance. When we fix the variance to a low value we see that MAP estimation starts getting away from ML estimation. The prior becomes effective and can help to escape from local minima. We also represent the evolution of the log-likelihood and the a posteriori log-probability over iterations, they are the criteria we want to maximize in ML and MAP estimations respectively. When the prior is activated by setting a low variance, we observe a brief decrease of the a posteriori probability, but it results in an increase of the SDR, which indicates that the EM algorithm has jumped to a region of the parameter space leading to better results.

We conducted several experiments in order to find a good strategy to adjust the strength of the prior, fixing the same variance for all source/microphone pairs, *i.e.* for all i, j, $\sigma_{ij}^2 = \sigma_a^2$. Empirically, we found that a good setting is to force a strong prior at the beginning of the EM algorithm and to weaken it over iterations $m = 1, ..., 400$ such that

$$\sigma_a^2(m) = \exp\left[\frac{\lfloor m-1 \rfloor}{50} - 2.5\right], \lfloor\cdot\rfloor \text{ is the floor function.} \quad (24)$$

We represent on figure 3 the evolution of σ_a^2 according to (24). 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 σ_{ij}^2. 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

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

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

[3] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, July 2007.

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

[5] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Underdetermined 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, July 2010.

[6] ——, "Spatial location priors for Gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, p. 149, September 2013.

[7] T. Otsuka, K. Ishiguro, T. Yoshioka, H. Sawada, and H. G. Okuno, "Multichannel sound source dereverberation and separation for arbitrary number of sources based on bayesian nonparametrics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2218–2232, 2014.

[8] 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, March 2010.

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

[10] R. Kumaresan, "On the zeros of the linear prediction-error filter for deterministic signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 31, no. 1, pp. 217–220, February 1983.

[11] 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 Processing*, vol. 92, pp. 1928–1936, March 2012.

[12] "BSS Eval Toolbox Version 3.0 for Matlab," http://bass-db.gforge.inria.fr/bss_eval/, [Online].

[13] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *7th International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, London, United Kingdom, September 2007, pp. 552–559.

[14] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and $\ell 1$-norm minimization." *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[15] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, 2001, pp. 556–562.

[16] "Matlab implementation of [8]," http://www.irisa.fr/metiss/ozerov/, [Online].

[17] M. R. Schroeder, "Frequency-correlation functions of frequency responses in rooms," *The Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1819–1823, 1962.