HAL open science

A new EM algorithm for underdetermined convolutive blind source separation

Zaher El Chami, Dinh-Tuan Pham, Christine Serviere, Alexandre Guérin

To cite this version:

Zaher El Chami, Dinh-Tuan Pham, Christine Serviere, Alexandre Guérin. A new EM algorithm for underdetermined convolutive blind source separation. EUSIPCO 2009 - 17th European Signal Processing Conference, Aug 2009, Glasgow, United Kingdom. pp.1457-1461. hal-00435933

HAL Id: hal-00435933

https://hal.science/hal-00435933v1

Submitted on 8 Jun 2021

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.

Distributed under a Creative Commons Attribution 4.0 International License

A NEW EM ALGORITHM FOR UNDERDETERMINED CONVOLUTIVE BLIND SOURCE SEPARATION

Zaher El Chami[1], *Antoine Dinh-Tuan Pham*[2], *Christine Servière*[3], *Alexandre Guerin*[1]

[1]Orange Labs - TECH/SSTP, 2 Avenue Pierre Marzin, 22307 Lannion Cedex, France
[2]Laboratory of Modeling and Computation, B.P 53X, 38041 Grenoble cedex 9, France
[3]GIPSA-lab, Department of Images and Signals, BP 46, 38402 St Martin d'Hère Cedex, France
E-mail: {zaher.elchami;alexandre.guerin}@orange-ftgroup.com
dinh-tuan.pham@imag.fr; christine.serviere@inpg.fr

ABSTRACT

This paper presents a new statistical method for separating more than two sound sources from a two-channel recording. It is based on a probabilistic model of the Interchannel Level/Phase Difference presented in [1] and the model parameters are estimated using the maximum likelihood criterion and an Expectation-Maximization algorithm. The source separation task is achieved by soft time-frequency masking of the observation. These masks are derived from the estimated source position model. Algorithm performance is evaluated on the real and synthetic convolutive mixtures data of the first audio source evaluation campaign [2] as well as the Signal Separation Evaluation campaign (SiSEC) [10]. Promising results are obtained when comparing to the other methods presented in these two campaigns.

1. INTRODUCTION

Blind Source Separation (BSS) is a widely used technique that aims at recovering a set of N original sources based only on their M observed mixtures. This task is more difficult when the mixing model is not instantaneous but convolutive and gets even harder in the underdetermined case ($N > M$). Indeed, when N is larger than M, no algebraic linear solution can be found to separate the sources, even if the mixing matrix is identified. Still, with the source sparseness assumption, researchers have found a way to build non linear masks for the separation task.

Most of the BSS methods that consider the sparseness assumption deal with the two-channel case from which cues or features like Interchannel Level/Phase Difference (ILD/IPD) are used. In [3] a KMeans algorithm classifies these features in clusters and binary separating masks are estimated. In [4], rather than clustering the observed (ILD/IPD), the authors propose to model these features as Gaussian variables with the assumption of a dominant path. Then, after estimating the model parameters, soft separating masks have been derived. Nevertheless, in real-world situations, the underlying linear phase assumption deriving from the dominant path hypothesis reveals not applicable due to early reflections and reverberation. Also, the proposed Gaussian model, even if it facilitates the equation computation, does not have any theoretical background.

In this paper a Model Based Underdetermined (blind) Source Separation (MBUSS) is considered. In our previous work [1] a theoretical distribution for the (log(ILD)/IPD) features is presented, but no separation algorithm based on this distribution was explicitly provided. This paper, proposes an estimation procedure for this theoretical distribution parameters, based on an Expectation-Maximization (EM) algorithm where the Maximization step is speeded-up by a Quasi-Newton algorithm. Unlike [3], probabilistic soft masks are computed instead of binary ones, thus less distortion and artifact are audible in the extracted sources. Unlike [4], separation is performed independently in each frequency band, therefore no linear phase assumption is made and wide band sources can be extracted even with the presence of frequency aliasing. Of course separating independently in each frequency band has the traditional permutation problem drawback. This permutation alignment is corrected using the ratio envelope of the extracted sources as in [5].

2. MODEL BASED SOURCE SEPARATION

We describe here briefly the (log(ILD)/IPD) probabilistic model that will be used in the separation task. For more detailed analysis, please refer to [1]. Consider the convolutive two-channel mixture model:

$$x_j(t) = \sum_{i=1}^{N} x_j^{(i)}(t) = \sum_{i=1}^{N} \sum_{k} a_{ji}(k)\, s_i(t-k) \qquad (1)$$

where $x_j^{(i)}(t)$ is the contribution of the ith source to the jth sensor. $s_i(t), i = 1,..,N$ are the sources and $a_{ji}(k)$ is the impulse response of the acoustic channel separating source i from microphone j with $j = 1, 2$. The time-domain observed signals $x_j(t)$ are converted into frequency-domain time-series signals using the Short-Time Fourier Transform (STFT):

$$X_j(t,\omega) = \sum_{k=0}^{L-1} w(k) x_j(t+k) e^{-j\omega k} = \sum_{i=1}^{N} X_j^{(i)}(t,\omega) \qquad (2)$$

where $w(k)$ is a window (e.g. Hanning) and $X_j^{(i)}(t,\omega)$ is the STFT of $x_j^{(i)}(t)$. Sparseness of the sources in the Time Frequency (TF) domain is the key assumption in solving the underdetermined separation problem. It means that each given source is non negligible on only a few number of TF slots. It generally implies that the sources have nearly disjoint supports in the TF domain in the sense that for each TF slot there can be at most one dominant source, and it is this last assumption that will be actually assumed [3]. If q is the index of this dominant source at (t,ω), then the jth observation in eq.(2) can be approximated to $X_j(t,\omega) \approx X_j^{(q)}(t,\omega)$ and the ratio between the two observations at (t,ω) will then be:

$$R(t,\omega) = \frac{X_1(t,\omega)}{X_2(t,\omega)} = \frac{\sum_{i=1}^{N} X_1^{(i)}(t,\omega)}{\sum_{i=1}^{N} X_2^{(i)}(t,\omega)} \approx \frac{X_1^{(q)}(t,\omega)}{X_2^{(q)}(t,\omega)}. \qquad (3)$$

The approximation $X_j^{(i)}(t,\omega) \approx A_{ji}(\omega) S_i(t,\omega)$, with $A_{ji}(\omega)$ as the Fourier transform of $a_{ji}(k)$, can be found in most of

the TF source separation methods. With it, the last term in (3) can be reduced to a constant in each frequency band ω and thus:

$$R(t,\omega) = \frac{X_1(t,\omega)}{X_2(t,\omega)} \approx \frac{X_1^{(q)}(t,\omega)}{X_2^{(q)}(t,\omega)} \approx \frac{A_{1q}(\omega)}{A_{2q}(\omega)}. \quad (4)$$

The above ratio, being (approximately) time independent but frequency and source dependent, has been widely used as dominant source indicator in the TF binary masks approach. In [3], the modulus and argument of $R(t,\omega)$, which are no more than the well-known Interchannel Level Difference (ILD) and Interchannel Phase Difference (IPD), have been clustered by a KMeans algorithm. The cluster members define the binary separating mask and the cluster centers give an estimation of $A_{1q}(\omega)/A_{2q}(\omega)$. But does the above approximation (4) hold when dealing with a long tap impulse response, i.e with a reverberating environment?

2.1 One source Ratio STFT distribution model

To simplify the notations in this section, we will omit the ω parameter. From now, we implicitly work in given frequency band ω. It has been demonstrated in [1] that, even if one source is observed, $R(t)$ is not constant in time. Fig.1 plots the real and imaginary part of $R(t)$ for a 10-second speech source placed in a moderate reverberant conditions (T_{60} = 250ms). This one source observed ratio should be considered as a (complex) random variable, the distribution of which has been derived in [1]. More precisely this paper provides the theoretical joint density of the real part and imaginary parts x and y of $\log R^{(q)}(t) = \log[X_1^{(q)}(t,\omega)/X_2^{(q)}(t,\omega)]$:

$$x = \log|R^{(q)}(t)| = \log \mathrm{ILD}_q, \quad y = \arg R^{(q)}(t) = \mathrm{IPD}_q$$

where $\mathrm{ILD}_q/\mathrm{IPD}_q$ are the ILD/IPD of s_q. Note that we did not considered the traditional feature $R^{(q)}(t)$ because it admits an infinite variance (see [1]). To estimate the joint density of (x,y), we assume that L is large enough so that, by the Central Limit Theorem, the pair $X_1^{(q)}(t)$ and $X_2^{(q)}(t)$ can be considered as Gaussian (complex circular) [1] with variance $\sigma_{1,q}^2(t)$ and $\sigma_{2,q}^2(t)$ and complex cross-correlation β_q. It has been demonstrated in [1] that the variance ratio $\sigma_{1,q}/\sigma_{2,q}$ and β_q does not vary in time and may be considered as specific features for each source s_q, hence of the variable $\log R^{(q)}(t)$. Based on the complex gaussian circular assumption, the couple of variables (x,y) admits the following joint density [1]:

$$p(x,y|r_q,\rho_q) = p_{\rho_q}(x - \log|r_q|, y - \arg r_q) \quad (5)$$

where $r_q = (\sigma_{1,q}/\sigma_{2,q})e^{i\arg\beta_q}$, $\rho_q = |\beta_q|$ and

$$p_\rho(x,y) = \frac{1}{4\pi}\frac{1-\rho^2}{(\cosh x - \rho\cos x)^2}.$$

The parameters r_q and ρ_q are specific to the source s_q position in space. Note that it is the source position in space which is modeled and not the source itself. Thus, this "'space"' position model can be applied independently from the source model (gaussian, laplacian, ...). More in details, r_q corresponds to the mean of $\log R^{(q)}(t)$ and thus depends only on the source position in space where $|r_q|$ is equal to the ILD_q and $\arg(r_q)$ is equal to the IPD_q. As for ρ_q, it will stand for the reverberation degree of the acoustic path separating s_q from the set of microphones. This can be viewed from its definition as the modulus of the cross-correlation between the two observations where high reverberation causes low cross-correlation between microphones and vice-versa. For example, in free field conditions, i.e anechoic environment, the cross-correlation is maximum and $\rho_q = 1$.

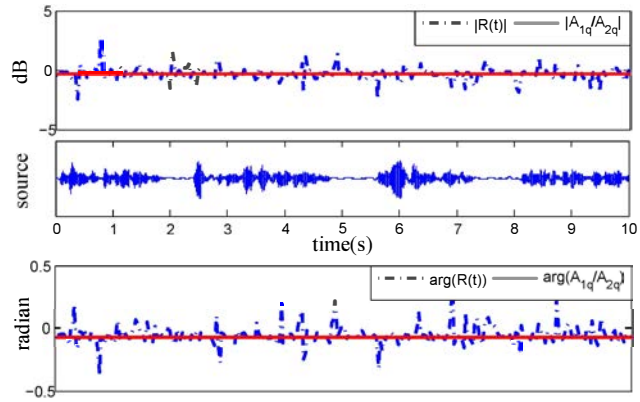

Figure 1: Ideal and observed (IPD,log(ILD)) for a single source mixture at the 100Hz center frequency

2.2 Mixture ratio STFT distribution model

Having the one source log ratio distribution model (5) and under the disjoint assumption, we are led to assume the following distribution model for the real and imaginary parts of the observed log ratio $\log[R(t,\omega)]$:

$$p(x,y|\boldsymbol{\rho},\boldsymbol{r},\boldsymbol{\mu}) = \sum_{i=1}^{N} \mu_i p_{\rho_i}\{x - \log|r_i|, y - \arg r_i\}. \quad (6)$$

This model is given at the frequency ω for the set of considered time points $t \in T$ with $\boldsymbol{\rho} = [\rho_1,..,\rho_N]$, $\boldsymbol{r} = [r_1,..,r_N]$ and $\boldsymbol{\mu} = [\mu_1,..,\mu_N]$ where μ_i is an added parameter that denotes the *a priori* probability of the ith source in the considered frequency band. This parameter reveals necessary since all sources are not equiprobable in a given frequency band, depending for instance on the mean pitch of a person. Simulations with equiprobable hypothesis showed that the estimation of the model parameters $\boldsymbol{\rho}$ and $\boldsymbol{r}$ is biased, hence producing degraded performance in terms of source separation.

2.3 Soft mask separation

The set of parameters $\boldsymbol{\rho}$, $\boldsymbol{r}$ and $\boldsymbol{\mu}$, which depend on the frequency band ω, are the parameters of the mixture model and need to be estimated in order to separate the sources. Once the parameters of the probabilistic model given in (6) are estimated, the *a posteriori* probability that i^{th} source is dominant at the TF point (t,ω) can be obtained directly as follows:

$$\pi_i(t) = \frac{\mu_i p_{\rho_i}[\log|R_i(t)|, \arg R_i(t)]}{\sum_{q=1}^{N}\mu_q p_{\rho_q}[\log|R_q(t)|, \arg R_q(t)]} \quad (7)$$

where $R_i(t) = R(t)/r_i$. Then, source separation can be easily performed in each frequency band ω by applying these above probabilities to the observations. Note that sources are independently separated in each frequency band, thus permutation ambiguity remains and needs to be solved. In this paper, the correlation between the ratio envelope is used as in [5]. Note that other methods, based on linear phase assumption [6], can not be used due to reverberation. Fig. 2 shows the flow chart of the Model Based Source Separation approach where $M_i(t,\omega)$ refer to the masks that extract the ith source. $M_i(t,\omega)$ is constructed from the *a posteriori* probabilities (7) after correcting the permutation ambiguities. Thus, $M_i(t,\omega) = \pi_{\Pi_\omega(i)}$ with $\Pi_\omega = [\Pi_\omega(1),\ldots,\Pi_\omega(N)]$

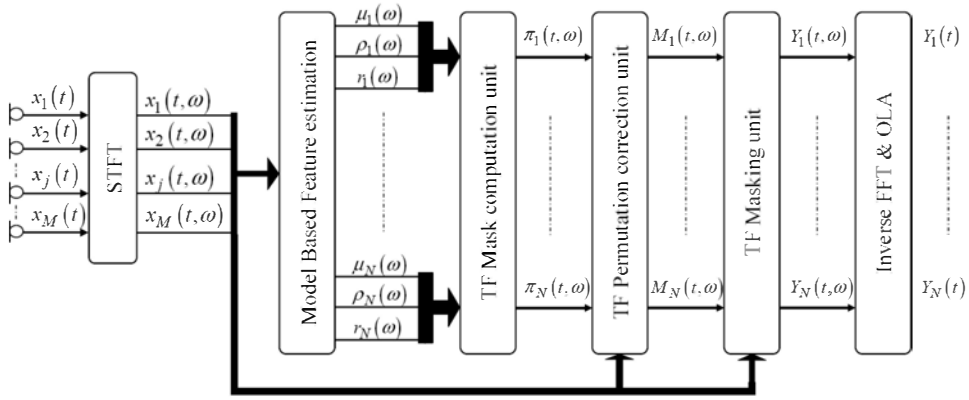

Figure 2: Basic scheme of Model Based Source Separation Approach

the permutation alignment vector estimated by [5] in the frequency band ω.

As said before, to build the soft separating mask, the μ_i, r_i, and ρ_i parameters need to be estimated for each source at each frequency band ω. To do so, we propose to use as criterion the maximum of the log-likelihood of the data $\{\log|R(t)|, \arg R(t)\}, t \in T$. Under the independence assumption between the time set of observations, it is given by:

$$L(\boldsymbol{\rho},\boldsymbol{r},\boldsymbol{\mu}) = \sum_{t\in T} \log\left\{\sum_{i=1}^{N} \pi_i p_{\rho_i}\left[\log|R_i(t)|, \arg R_i(t)\right]\right\}. \quad (8)$$

The above log-likelihood is hard to maximize. However, it may be recasted as the log-likelihood for a model with missing observations or hidden variables. These variables are the indexes that indicate which source is dominant at each (t,ω) point (here ω is fixed and hence not displayed). In this context the log-likelihood can be maximized by the well known Expectation Maximization (EM) algorithm [7].

3. THE EM ALGORITHM

This algorithm operates in two steps as described below.

3.1 The E-step

This step computes the conditional expectation of the full log-likelihood given the data $\{\log|R(t)|, \arg R(t)\}, t \in T$. The expected log-likelihood will be computed at generic new parameters μ_i', r_i' and ρ_i' and the conditional expectation is computed relatively to the model specified by the current parameters μ_i, r_i and ρ_i. The result can be shown to be:

$$\sum_{t\in T}\sum_{i=1}^{N} \pi_i(t) \log\left\{\mu_i' p_{\rho_i'}\left[\log\left|\frac{R(t)}{r_i'}\right|, \arg\frac{R(t)}{r_i'}\right]\right\} \quad (9)$$

where $\pi_i(t)$ is the *a posteriori* probability given in (7) and computed at the current parameter μ_i, r_i and ρ_i.

3.2 The M-step

This step maximizes the above conditional expectation of the full log-likelihood with respect to the generic parameters μ_i', r_i' and ρ_i'. The maximum point is then taken as the new parameter. It is easily seen that the maximization of (9) with respect to μ_i' (under the constraint $\sum_{i=1}^{N}\mu_i' = 1$) and with respect to the set (r_i', ρ_i') can be performed independently. The first maximization yields the new μ_i: $\mu_i = \sum_{t\in T}[\pi_i(t)]/|T|$ where $|T|$ denotes the number of points in T. The second one is reduced to the maximization of:

$$C(r_i', \rho_i') = \sum_{t\in T} \pi_i(t) \log p_{\rho_i'}\left[\log\left|\frac{R(t)}{r_i'}\right|, \arg\frac{R(t)}{r_i'}\right] \quad (10)$$

for each $i = 1, .., N$ with respect to r_i' and ρ_i'.

3.3 Relaxing the M-step

Maximizing the above expression (10) is not easy and can not be done without using an iterative algorithm. In [8], we overcome this problem by replacing the theoretical density (5) in (10) with another similar and easier one to handle. Thus, we were able to maximize it analytically. In this work, as the theoretical density shall be used, we propose not to maximize the expected full log-likelihood (10) but to just make it *increase*. In fact, from the EM algorithm theory, *increasing* the expected likelihood (10) would be sufficient to increase, also, the marginal likelihood (8). Therefore, given the current parameters r_i and ρ_i, we limit ourselves to finding the new parameters r_i' and ρ_i' that increase the objective function (10). Such new parameters can indeed be found *analytically* as described in the following basic step:

3.3.1 Basic step

Using Jensen's inequality, it can be easily proved that $C(r_i', \rho_i') - C(r_i, \rho_i)$ has the following lower bound:

$$-2\zeta \log \sum_t \frac{w_i(t)\left\{\cosh[\xi_i(t) - \lambda_i] - \rho_i' \cos[\phi_i(t) - \theta_i]\right\}}{\zeta(1-\rho_i'^2)^{1/2}/(1-\rho_i^2)^{1/2}} \quad (11)$$

where

$$\begin{aligned} w_i(t) &= \frac{\pi_i(t)}{\cosh[\log|R_i(t)|] - \rho\cos[\arg R_i(t)]} \\ &= \frac{2|R_i(t)|\pi_i(t)}{|R_i(t)|^2 + 1 - 2\rho\Re[R_i(t)]} \end{aligned}$$

and $\zeta = \sum_u \pi_i(u)$, $\xi_i(t) = \log|R_i(t)|$, $\lambda_i = \log|r_i'/r_i|$, $\phi_i(t) = \arg R_i(t)$, $\theta_i = \arg(r_i'/r_i)$ and $\Re(z)$ is the real part of the complex number z. The maximum of the above equation (11), with respect to (r_i', ρ_i'), is strictly larger than zero (because it is equal to zero when $(r_i', \rho_i') = (r_i, \rho_i)$). Thus, the new couple (r_i', ρ_i') that maximizes (11) will increase, at the same time, the objective function (10). Therefore, by a simple variable substitution, the problem of increasing the likelihood is now reduced to maximize (11) with respect to

		4 Men Speech					4 Women Speech					3 Music, No drums				3 Music, with drums				
		S1	S2	S3	S4	Mean	S1	S2	S3	S4	Mean	S1	S2	S3	Mean	S1	S2	S3	Mean	OVAP
new MBUSS	SDR	3,91	2,68	3,71	4,17	3,61	4,46	3,40	4,37	5,27	4,37	2,42	5,80	4,81	4,34	1,65	3,49	0,35	1,83	3,61
	ISR	6,17	5,99	6,52	6,87	6,39	6,58	7,28	7,07	7,52	7,11	4,27	8,61	8,61	7,16	1,76	6,38	0,40	2,85	6,00
	SIR	5,75	2,62	5,37	6,89	5,16	6,75	3,45	6,77	10,13	6,78	4,07	8,80	6,57	6,48	13,38	6,15	11,88	10,47	7,04
	SAR	6,21	5,52	6,19	6,08	6,00	7,78	7,44	6,82	7,72	7,44	5,50	11,24	8,96	8,56	15,90	9,73	7,56	11,06	8,05
old MBUSS [8]	SDR	3,25	1,97	3,69	4,11	3,25	4,43	3,84	4,44	5,50	4,55	1,69	5,87	5,19	4,25	1,64	3,02	0,34	1,67	3,50
	ISR	6,14	4,17	6,87	7,21	6,10	6,34	8,03	7,89	8,65	7,73	4,31	8,30	10,50	7,70	1,80	7,16	0,42	3,12	6,27
	SIR	3,73	2,21	5,20	6,27	4,35	8,45	5,35	7,24	9,53	7,64	3,15	9,78	7,15	6,69	14,85	6,94	11,94	11,24	7,27
	SAR	5,99	5,04	6,25	6,17	5,86	6,78	6,95	6,68	7,29	6,93	5,60	11,26	9,00	8,62	11,59	6,09	6,01	7,89	7,19
Kmeans [3]	SDR	3,50	2,24	3,16	3,57	3,12	4,03	3,10	3,77	5,22	4,03	1,82	5,58	5,15	4,19	1,72	4,62	0,36	2,23	3,42
	ISR	7,64	5,90	7,38	8,65	7,39	6,16	10,55	9,20	10,06	8,99	5,29	8,00	10,28	7,86	1,82	8,75	0,42	3,66	7,15
	SIR	7,01	4,62	6,36	7,08	6,27	12,51	5,52	8,04	9,81	8,97	4,13	8,49	7,05	6,55	14,73	7,77	13,68	12,06	8,34
	SAR	4,58	2,40	4,24	4,54	3,94	4,69	5,65	5,53	6,31	5,54	7,34	11,27	9,68	9,43	16,26	10,92	7,75	11,64	7,22

Table 1: Results for synthetic recording with a 5 cm microphone spacing and two different types of sources: speech and music. the overall performance OVAP of the source separation is presented in the last column. All figures are given in dBs

$\theta_i, \lambda_i, \rho'_i$. By developing the numerator terms $\cosh[\xi_i(t) - \lambda_i]$ and $\cos[\phi_i(t) - \theta_i]$, the maximum is obtained when

$$\begin{aligned}
\lambda_i &= \tanh^{-1} \frac{\sum_t w_i(t) \sinh \xi_i(t)}{\sum_t w_i(t) \cosh \xi_i(t)} \\
\theta_i &= \arg \sum_t w_i(t) \text{sign}\{R_i(t)\} \\
\rho_i'^2 &= \frac{[\sum_t w_i(t) \cos \phi_i(t)]^2 + [\sum_t w_i(t) \sin \phi_i(t)]^2}{[\sum_t w_i(t) \cosh \xi_i(t)]^2 - [\sum_t w_i(t) \sinh \xi_i(t)]^2}
\end{aligned}$$

where $\text{sign}\{z\} = z/|z| = e^{i \arg z}$. Finally, to increase (10), we are led to assume the following *one* iteration basic step:

$$r_i \leftarrow r_i \sqrt{a_i/b_i} \text{ sign } c_i, \qquad \rho_i \leftarrow |c_i|/\sqrt{a_i b_i}$$

where $a_i = \sum_t \chi_i(t)|R_i(t)|^2$, $b_i = \sum_t \chi_i(t)$, and $c_i = \sum_t \chi_i(t) R_i(t)$ with $\chi_i(t) = \frac{1}{2} w_i(t)/|R_i(t)|$.

3.3.2 Combining with the Quasi Newton step

The basic step could lead to a slower convergence of the algorithm as the obtained increase of the expected log-likelihood (10) can be much less than the maximum achievable. To overcome this problem we will consider a Quasi-Newton (Q-N) algorithm. In fact, when the old parameter r_i and ρ_i are close to the maximum solution of (10), the Quasi-Newton algorithm would have a quadratic convergence to the point maximizing (10). Therefore, the algorithm should converge within one single Q-N iteration. However, in contrast to the basic step, the Q-N step does not guarantee the increase of the expected log-likelihood. It does not even guarantee the new ρ'_i to be in the interval (0,1). Thus, the following strategy is adopted in the M-step:

- Compute the new parameter ρ'_i and r'_i based on the basic step as in section 3.3.1
- Compute the other new estimation of $\hat{\rho}'_i$ of the Q-N step and test if it belongs to the interval (0,1)
- If not, adopt the new parameter ρ'_i and r'_i of the basic step, otherwise compute the other new parameter $\hat{r}'_i$ of the Q-N step and test if it and $\hat{\rho}'_i$ of the Q-N step lead to a larger increase of the expected log-likelihood than the basic step; if so, adopt these parameter $\hat{r}'_i$ and $\hat{\rho}'_i$ otherwise adopt those of the basic step ρ'_i and r'_i.

Computations (not detailed here) show that the Q-N step is given by:

$$\begin{aligned}
\hat{\rho}'_i &= \rho_i - \frac{\rho_i(a_i + b_i) - 2\Re(c_i)}{\sum_t \chi^2(t)\{[|R_i(t)|^2 + 1]^2 - 4\Re[R_i(t)]^2\}/\pi_i(t)} \\
\hat{r}'_i &= r_i \exp[(3/2)(a_i + b_i)/(a_i - b_i)] \text{sign}[3c_i - \Re(c_i)]
\end{aligned}$$

Note that these formula involve the already computed quantities a_i, b_i, c_i.

3.4 EM Initialization and Stop Criteria

A simple way to initialize our algorithm is to choose randomly N times points t_1,.., t_N and initialize r_i by $r_i = R(t_i)$, $i = 1..N$. The *a posteriori* probability $\pi_i(t)$ can be initialized by formula (7) using the density (5) with ρ_i set to one, leading to:

$$\pi_i(t) = \frac{d[R(t), r_i]}{\sum_{i=1}^N d[R(t), r_i]}$$

where $d[R(t), r_i] = \cosh \log [R(t)/r_i] - \cos \arg [R(t)/r_i]$. Since the log-likelihood increases monotonically at each EM iteration, the EM iteration process is stopped when the increase of log-likelihood becomes insignificant(e.g 10^{-6}).

4. EXPERIMENTS AND RESULTS

In order to evaluate our algorithm, we simulated speech and music mixtures in a reverberant noise-free environment by convolving speech and music samples with filter impulse responses coming from the first audio source separation campaign [2]. Also, it was recently tested in the Signal Separation Evaluation Campaign (SiSEC) [10] where results showed that our algorithm compares favorably with the others presented in this campaign in most of the real and synthetic convolutive situation. As these results can be accessed at [10], only the results for the first audio source separation campaign will be presented. Signals are 16 kHz-sampled and have a 10s duration. Four types of mixtures were generated: 4 female speakers, 4 male speakers, 3 musics with one of them is drums and 3 other musics (no drums). Two different sets of source positions are used, one for speech and another for music mixtures. Angles and distances of these positions are given in Table 2. Our algorithm uses Hanning windowed 2048 sample frames, and reconstruction is achieved using the overlap and add method with 75% overlap.

	Speech Sources				Music Sources		
	S1	S2	S3	S4	M1	M2	M3
Distance (m)	1,2	1,1	1	0,8	1,1	0,9	1
Angle (deg)	50	-15	-45	15	45	-30	5

Table 2: Source positions from the set of microphone

4.1 Comparison Algorithm and performance measurement

The proposed algorithm (referred to as new MBUS) is compared with two other algorithms: the first one (referred to

as old MBUS) is our previous method based on the same model-based EM approach, but with a simplifying probability distribution of the couple of variable (IPD,log(ILD)) [8]. The second one (referred to as Kmeans) is the KMeans underdetermined source separation presented in [3], in which the ILD/IPD observations are grouped into N clusters using a KMeans algorithm: each cluster center gives an estimation of the mixing matrix and the cluster point sets give the binary TF separating masks. The separation performance was evaluated for each estimated source i by the same criteria used in SiSEC: Signal to Interference Ratio (SIR_i), Image to Signal Ratio (ISR_i), Signal to Distortion Ratio (SDR_i) and Signal to Artifact Ratio (SAR_i). For a detail description of these criteria and of their computation, the reader may refer to [10].

4.2 Results

Detailed performance are given in Table 1 whereas the mean of each case and the overall performance OVAP (mean performance on all mixture types) are plotted in Figure 3. When looking at the SDR which computes the global separation performance, OVAP column, the proposed MBUSS gives the best results as compared to the others. It shows a slight advantage for the new MBUSS 3.61dB compared to the old one 3.5dB and a more efficient as compared to KMeans 3.42dB. Furthermore, the intermediate errors show that the compromise operated by the algorithms is different.

Comparing the new MBUSS to the old one, results show that the proposed algorithm gives better results in terms of SAR (SAR=8.05 and 7.19 respectively) and, almost, the same in terms of SIR (SIR=7.04 and 7.27 respectively), which justifies the use of the theoretical distribution and not the approximated one as in [8]. Comparing the new MBUSS with the KMeans, Fig. 3 (a),(b) and (c) show that the MBUSS method performs better for speech mixture, whereas KMeans performs better on music ones. One explanation could lie in the nature of the signal themselves: music sources are very resonant, hence the energy is highly concentrated on some (usually harmonic) frequencies, which ensures a quasi-disjoint spectro-temporal supports between instruments. Thus, hard masking would be more suitable filter for this type of mixtures than a soft filter. The OVAP on Fig. 3 (d) also shows that the KMeans algorithm favors weak interferences (SIR=8.34) to the cost of more degraded separated speech (SAR=7.22), in contrast to our method that presents stronger interference (SIR=7.04) but with a clearer separated speech signal (SAR=8.05).

These results are confirmed by informal listening tests: less audible artifact are noticed in the proposed MBUSS, at the expense of more interference, which reveals nicer to listen to since our auditory system is very sensitive to artifact (gurgling noise) and less to interference. One explanation (partially) lies in the nature of the mask: as compared to the binary mask used in the KMeans method, the soft mask slightly smooths the output, limiting the isolated errors (the artifact like musical noise), but favoring the presence of interference.Note that the mask is used "as is": we are inclined to think that some "wise" smoothing (linear or nonlinear) could give better objective and subjective results, reducing even more the artifact.

5. CONCLUSION

In this paper, a new method to solve the underdetermined blind separation of audio mixtures problem has been presented. It is based on the sparseness assumption and on a theoretical model for the interchannel cues (log(ILD),IPD) given in [1]. By an EM algorithm, we were able to estimate the parameters of this model and then build the time frequency separating masks. The algorithm demonstrated its ability to separate undetermined reverberant mixtures where in terms of objective criteria it gives the best results in terms of artifact and distortion. Nevertheless, more studies need to be done, especially on the robustness of the model over the source positions and room reverberation.

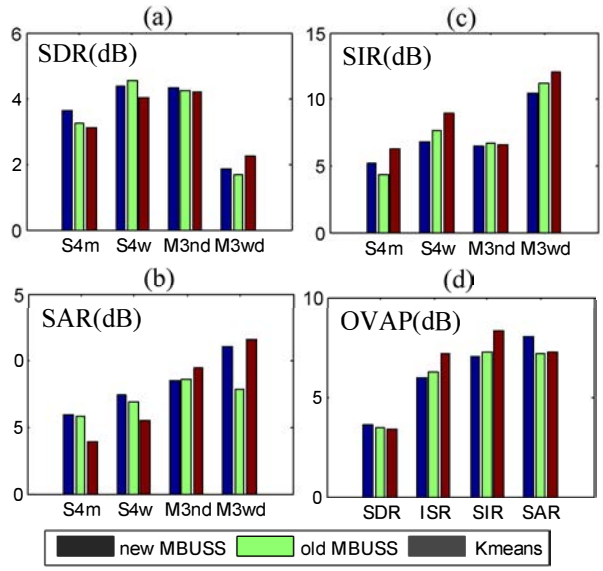

Figure 3: (a),(b),(c) presents respectively the SDR, SAR and SIR performance for each type of mixture; (d) presents the OVAP

REFERENCES

[1] A. Pham, Z. El-Chami, C. Serviere, and A. Guerin, "Modeling the short time fourier transform ratio and application to underdetermined audio source separation,"in *ICA 2009*, Brazil.

[2] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J.P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results,"*ICA 2007*, UK.

[3] S. Araki, H. Sawada and S. Makino (2007). "K-means based underdetermined blind speech separation."In *Blind speech separation*: 243–270. S. Makino, Te-Won Lee and H. Sawada Editors, Springer: New-York.

[4] Michael I. Mandel, Daniel P. W. Ellis, "EM localization and separation using interaural level and phase cues, "*WASPAA 2007*, Japan.

[5] H. Sawada, S. Araki, S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS,"ISCAS 2007, USA

[6] F. Nesta, M. Omologo, and P. Svaizer, "A novel robust solution to the permutation problem based on a joint multiple TDOA estimation "*IWAENC 2008*, USA.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm,"*Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[8] Z. El-Chami, D.-T. Pham, Ch. Servière and A. Guérin, "A new-model based underdetermined source separation,"*IWAENC 2008*, USA.

[9] http://sassec.gforge.inria.fr/

[10] http://sisec.wiki.irisa.fr/tiki-index.php