Spatial location priors for Gaussian model-based reverberant audio source separation

Ngoc Q. K. Duong, Emmanuel Vincent, Rémi Gribonval

Spatial location priors for Gaussian model-based reverberant audio source separation

Ngoc Q. K. Duong, Emmanuel Vincent, Rémi Gribonval

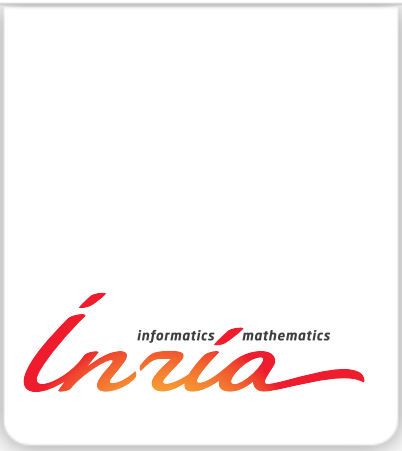
RESEARCH REPORT

N° 8057

September 2012

Project-Teams Metiss

ISSN 0249-6399 ISRN INRIA/RR--8057--FR+ENG



Spatial location priors for Gaussian model-based reverberant audio source separation

Ngoc Q. K. Duong*, Emmanuel Vincent, Rémi Gribonval

Project-Teams Metiss

Research Report n° 8057 — September 2012 — 19 pages

Abstract: This article addresses the under-determined reverberant audio source separation when prior knowledge about the spatial source locations and the room characteristics is available. We consider the Gaussian modeling framework whereby the contribution of each source to all mixture channels in the time-frequency domain is modeled as a zero-mean Gaussian random variable whose covariance represents the spatial characteristics of the source. We advocate the use of rigorous Bayesian estimation by defining three different priors over the spatial parameters, whose means are given by the theory of statistical room acoustics and whose variances are learned from training data. We then derive corresponding Expectation-Maximization (EM) algorithms to estimate the model parameters in the Maximum A Posteriori (MAP) sense. These algorithms provide a principled solution to the well-known permutation problem and two of them achieve better separation performance, as shown in our experiment, than Maximum Likelihood (ML) based EM algorithms exploiting the same prior knowledge.

Key-words: audio source separation, local Gaussian model, probabilistic prior, maximum a posteriori

* N. Q. K. Duong is with Technicolor, Rennes Research & Innovation Centre.

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

A priori de localisation spatiale pour la séparation de sources audio réverbérées par modèle gaussien

Résumé : Cet article porte sur la séparation de mélanges réverbérants sous-déterminés de sources audio lorsque des connaissances *a priori* sont disponibles sur la position spatiale des sources et les caractéristiques de la pièce. Nous considérons le cadre de la modélisation gaussienne, où la contribution de chaque source à l'ensemble des canaux du mélange est représentée comme un vecteur gaussien aléatoire de moyenne nulle et dont la covariance représente les caractéristiques spatiales de la source. Nous promouvons une estimation bayésienne rigoureuse en définissant trois distributions *a priori* différentes sur les paramètres spatiaux, dont la moyenne est donnée par la théorie statistique de l'acoustique des salles et dont les variances sont apprises sur des données d'apprentissage. Nous concevons ensuite les algorithmes d'Espérance-Maximisation (EM) correspondants pour estimer les paramètres du modèle au sens du Maximum *A Posteriori* (MAP). Ces algorithmes fournissent une solution rigoureuse au problème de permutation bien connu et deux d'entres eux fournissent une meilleure performance de séparation que les algorithmes EM basés sur le critère du Maximum de Vraisemblance (MV) et exploitant les mêmes connaissances pour l'initialisation, comme montré par nos expériences.

Mots-clés : séparation de sources audio, modèle gaussien local, distribution *a priori*, maximum *a posteriori*

1 Introduction

We deal with the separation of individual reverberated audio sources from their recorded mixtures, which usually occur in cocktail party rooms and the problem is well known as blind source separation (BSS). Let us consider a multichannel mixture signal recorded by an array of I microphones and denote it by $\mathbf{x}(t) = [x_1(t), ..., x_I(t)]^T$. This mixture signal can be expressed as [1]

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t) \tag{1}$$

where J denotes the number of sources, and $\mathbf{c}_j(t) = [c_{1j}(t), ..., c_{Ij}(t)]^T$ is the *spatial image* of the j-th source, that is the contribution of this source to all mixture channels. Note that background noise is also considered as a source in the above equation. In the case when the j-th source is a point source, *i.e.* it emits sound from a single position in space, $\mathbf{c}_j(t)$ is characterized as [2]

$$\mathbf{c}_j(t) = \sum_{\tau} \mathbf{h}_j(\tau) s_j(t-\tau) \tag{2}$$

where $\mathbf{h}_j(\tau) = [h_{1j}(\tau), ..., h_{Ij}(\tau)]^T$ are linear mixing filters modeling the acoustic path from the j-th source to all I microphones and $s_j(t)$ is the emitted single-channel source signal. In this paper, we focus on the separation of under-determined mixtures, *i.e.* such that $I < J$, assuming that J is known.

BSS systems usually exploit information about either the source spatial positions or about the source spectral structures, *e.g.* sparsity, harmonicity, smoothness, etc., to discriminate the sources [3]. For that purpose, most existing BSS approaches operate in the time-frequency (T-F) domain via the short-time Fourier transform (STFT) and rely on *narrowband approximation* of the convolutive mixture by complex-valued multiplication in each frequency bin f and time frame n as

$$\mathbf{c}_j(n, f) \approx \mathbf{h}_j(f) s_j(n, f) \tag{3}$$

where the $I \times 1$ mixing vector $\mathbf{h}_j(f)$ is the Fourier transform of $\mathbf{h}_j(\tau)$, $s_j(n, f)$ and $\mathbf{c}_j(n, f) = [c_{1j}(n, f), ..., c_{Ij}(n, f)]^T$ are the STFT coefficients of the sources $s_j(t)$ and their spatial images $\mathbf{c}_j(t)$, respectively. Popular approaches include binary masking [4] and ℓ_1-norm minimization [5]. However, the separation performance achievable by these techniques remains limited for reverberated or diffuse sources, *i.e.* sound from the source coming from all directions, [6, 7] since the narrowband approximation does not hold.

Our work has built upon an emerging nonstationary Gaussian modeling framework [8, 9, 10] whereby the T-F coefficients of the source image $\mathbf{c}_j(n, f)$ are modeled as a zero-mean Gaussian random vector with covariance matrix $\mathbf{\Sigma}_j(n, f) = \mathbb{E}(\mathbf{c}_j(n, f)\mathbf{c}_j^H(n, f))$ factored a

$$\mathbf{\Sigma}_j(n, f) = v_j(n, f)\, \mathbf{R}_j(f) \tag{4}$$

where $v_j(n, f)$ are scalar time-varying *variances* encoding the spectro-temporal power of the sources and $\mathbf{R}_j(f)$ are $I \times I$ *full-rank* time-invariant *spatial covariance matrices* encoding their spatial position and spatial spread. This parameterization is probabilistic in the sense that $\mathbf{c}_j(n, f)$ can not be deterministically computed from the chosen

parameters, but is randomly generated according to the considered Gaussian distribution. Note that this framework does not rely on the point source assumption nor on the narrowband assumption, hence it appears applicable to reverberated or diffuse sources.

Under the classical assumption that the sources are uncorrelated, the vector $\mathbf{x}(n,f)$ of STFT coefficients of the mixture signal is also zero-mean Gaussian with covariance matrix

$$\mathbf{\Sigma_x}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(f). \tag{5}$$

Under this model, source separation can be achieved in two major steps: the model parameters $\theta = \{v_j(n,f), \mathbf{R}_j(f)\}$ are first estimated, the spatial images of all sources are then obtained in the minimum mean square error (MMSE) sense by multichannel Wiener filtering

$$\widehat{\mathbf{c}}_j(n,f) = v_j(n,f)\mathbf{R}_j(f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f). \tag{6}$$

While BSS requires to recover either the source signals or their spatial images from a given mixture without any other knowledge, in many practical situations useful information such as the geometric setting and the room acoustical characteristics can be known in advance and would benefit the source separation performance. Such situations happen, for instance, in a formal meeting where the position of each delegate is fixed, or in a car where the positions of the driver and the passengers are also fixed. This additional information has been exploited in the literature: the source direction of arrival (DOA) has been used to solve the well-known *permutation problem* in the frequency-domain BSS, when the parameters are independently estimated in each frequency bin, so as the parameters are aligned to correspond to the same source across all frequency bins f [11, 12]. DOAs have also been used by geometrically constrained independent component analysis (ICA) [13] and by adaptive beamforming incorporated before BSS algorithm [14, 15] to provide better estimation of the parameters. However all these exploitations lack some theoretical justification for the use of the additional geometric information. In this paper, we advocate the use of rigorous Bayesian estimation in the local Gaussian modeling framework by defining suitable prior distributions for the spatial parameters. These priors all rely on the theory of statistical room acoustics in order to express the mean and/or the variance of the prior as a function of the geometric setting and the room characteristics. We then extend two different EM algorithms presented in [10] and [16], respectively, so as the spatial parameters are estimated in the maximum a posteriori (MAP) sense. The resulting MAP algorithms offer an acoustically principled solution to the estimation of the model parameters and to the permutation problem. Most importantly, they provide a proof of concept of the benefit of the proposed priors towards their future use in a BSS context.

The structure of the rest of the article is as follows. We introduce the source image-based EM algorithms exploiting inverse-Wishart and Wishart spatial location priors in Section 2. We then present the source-based EM algorithms exploiting Gaussian prior in Section 3. We compare the source separation performance achieved by each MAP-based EM algorithm to that of ML-based EM algorithm exploiting the same prior knowledge in Section 4. Finally we conclude and discuss further research directions in Section 5.

2 Source image-based EM algorithms

In this section we first present the general EM algorithm under the source image-based mixing model (1) and (5) where the parameters can be estimated either in the ML or in the MAP sense. We then define two different priors for the spatial covariance matrices $\mathbf{R}_j(f)$ and derive the corresponding MAP parameter updates.

2.1 General EM implementation

EM algorithm for the ML parameter estimation has been presented in [10] where the updates are separately derived for each frequency bin f from the *complete data* $\{\mathbf{c}_j(n,f)\}_{n,f}$, that is the set of hidden T-F coefficients of the spatial images of all sources on all time frames. Since the spatial priors considered in this paper do not change the E-step and the update of $v_j(n,f)$ in the M-step of the algorithm, we use the same EM algorithm as described in [10] except that the update of $\mathbf{R}_j(f)$ in the MAP sense will be derived with the use of the spatial priors.

The general EM algorithm is summarized in in Algorithm 1. In the E-step, the Wiener filter $\mathbf{W}_j(n,f)$ and the covariance $\widehat{\mathbf{\Sigma}}_j(n,f)$ of the spatial image of the jth source are computed. Then in the M-step, $v_j(n,f)$ and $\mathbf{R}_j(f)$ are alternatingly updated.

Algorithm 1 General source image-based EM algorithm

E step:

$$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\mathbf{R}_j(f) \tag{7}$$

$$\mathbf{W}_j(n,f) = \mathbf{\Sigma}_j(n,f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f) \tag{8}$$

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_j(n,f) = \mathbf{W}_j(n,f)\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{W}_j^H(n,f) + \\ + (\mathbf{I} - \mathbf{W}_j(n,f))\mathbf{\Sigma}_j(n,f)\end{aligned} \tag{9}$$

M step:

$$v_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)) \tag{10}$$

$$\text{Update } \mathbf{R}_j(f). \tag{11}$$

where $\mathrm{tr}(.)$ denotes the trace of a square matrix, $\mathbf{I}$ is the identity matrix, and $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$ is computed by locally averaging over the neighborhood of each T-F bin as [17]

$$\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f) = \sum_{n',f'} w_{nf}^2(n',f')\mathbf{x}(n',f')\mathbf{x}^H(n',f') \tag{12}$$

where w_{nf} is a bi-dimensional window specifying the shape of the neighborhood such that $\sum_{n',f'} w_{nf}^2(n',f') = 1$. Note that, strictly speaking, this algorithm is a generalized form of EM [18], since the M-step increases but does not maximize the likelihood of the complete data due to the interleaving of the updates of $v_j(n,f$ and $\mathbf{R}_j(f)$.

2.2 MAP spatial parameter update exploiting an inverse- Wishart prior

2.2.1 inverse-Wishart prior

According to the theory of statistical room acoustics [19, 20], for a given microphone spacing and source position relative to the microphones, the mean spatial covariance matrix over all possible microphone positions is given by

$$\mathbf{R}_j(f) = \mathbf{h}_j^{\text{ane}}(f)\big(\mathbf{h}_j^{\text{ane}}\big)^H(f) + \sigma_{\text{rev}}^2\mathbf{\Omega}(f) \tag{13}$$

where σ_{rev}^2 is the variance of the reverberant part, $\Omega_{il}(f)$ is a function of the microphone directivity pattern and the distance d_{il} between the i-th and the l-th microphone such that $\Omega_{ii}(f) = 1$, and $\mathbf{h}_j^{\text{ane}}(f)$ is mixing vector coefficients modeling the direct path from j-th source to all microphones. $\mathbf{h}_j^{\text{ane}}(f)$ can be parameterized as [20]

$$\mathbf{h}_j^{\text{ane}}(f) = \begin{pmatrix} \frac{1}{\sqrt{4\pi}r_{ij}}e^{-2i\pi f\frac{r_{ij}}{c}} \\ \vdots \\ \frac{1}{\sqrt{4\pi}r_{Ij}}e^{-2i\pi f\frac{r_{Ij}}{c}} \end{pmatrix} \tag{14}$$

where c is the sound velocity and r_{ij} is the distance from the j-th source to the i-th microphone. Assuming that the reverberant part is diffuse, *i.e.* its intensity is uniformly distributed over all possible directions, for omni-directional microphones its normalized cross-correlation can be shown to be real-valued and equal to [19]

$$\Omega_{il}(f) = \frac{\sin(2\pi f d_{il}/c)}{2\pi f d_{il}/c}. \tag{15}$$

Moreover, the power of the reverberant part within a parallelepipedic room with dimensions L_x, L_y, L_z is given by

$$\sigma_{\text{rev}}^2 = \frac{4\beta^2}{\mathcal{A}(1-\beta^2)} \tag{16}$$

where $\mathcal{A}$ is the total wall area and β the wall reflection coefficient computed from the room reverberation time T_{60} via Eyring's formula [20]

$$\beta = \exp\left\{-\frac{13.82}{(\frac{1}{L_x}+\frac{1}{L_y}+\frac{1}{L_z})cT_{60}}\right\}. \tag{17}$$

This model assumes that the reverberation recorded at all microphones has the same power but is correlated as characterized by $\Psi_{il}(f)$. It has been employed for single source localization in [20] and for source separation in our earlier work [21]. However, our preliminary experiments have confirmed that the actual value of $\mathbf{R}_j(f)$ varies depending on the microphone positions and cannot be set to the fixed value. Therefore, we propose to model $\mathbf{R}_j(f)$ as

$$p(\mathbf{R}_j(f)) = \mathcal{IW}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big) \tag{18}$$

where

$$\mathcal{IW}(\mathbf{R}|\mathbf{\Psi}, m) = \frac{|\mathbf{\Psi}|^m|\mathbf{R}|^{-(m+I)}e^{-\text{tr}(\mathbf{\Psi}\mathbf{R}^{-1})}}{\pi^{I(I-1)/2}\prod_{i=1}^{I}\Gamma(m-i+1)} \tag{19}$$

is the inverse Wishart density [22] over a Hermitian positive definite matrix $\mathbf{R}$ with positive definite inverse scale matrix $\mathbf{\Psi}$, m degrees of freedom and mean $\mathbf{\Psi}/(m-I)$ [22], with Γ the gamma function. This is the *conjugate prior* for the likelihood of the considered Gaussian observation model, so that it results in simple closed-form parameter updates as shown in Section 2.2.3. This distribution, its mean, and its variance are finite for $m > I-1$, $m > I$, and $m > I+1$ respectively. We define the inverse scale matrix $\mathbf{\Psi}_j(f)$ as

$$\mathbf{\Psi}_j(f) = (m-I)\Big(\mathbf{h}_j^{\text{ane}}(f)\big(\mathbf{h}_j^{\text{ane}}(f)\big)^H + \sigma_{\text{rev}}^2\mathbf{\Omega}(f)\Big) \tag{20}$$

so that the mean of $\mathbf{R}_j(f)$ is coherent with (13).

This prior distribution extends the parameterization in (13) by allowing deviations of the spatial covariance matrices around their mean controlled by the number of degrees of freedom m. Such deviations occur since (13) is only valid on average when considering a large number of sources, or for instance when the source or the microphones are close to the walls, resulting in a strong directional early echo. Note that the choices of the inverse-Wishart prior in this section and the Wishart prior in Section 2.3.1 are due in particular to the *engineering* constraints such that they can:

- apply to Hermitian matrices $\mathbf{R}_j(f)$
- have a closed-form expression for the mean
- result in close-form parameter updates.

2.2.2 Learning the hyper-parameter

In order to obtain the best fit between the prior distribution and the actual spatial covariance matrices, we learn the number of degrees of freedom m of the inverse-Wishart prior (18) from training data. For a given geometric setting and reverberation time, training signals are generated by convolving training source signals with mixing filters generated for many source and microphone positions p. The corresponding spatial covariance matrices $\mathbf{R}_p(f)$ are then estimated via the oracle estimator, by alternating (10) and (25) with $\gamma = 0$ and $\widehat{\mathbf{\Sigma}}_p(n,f)$ is computed from training signals similar to (12). Since $\mathbf{R}_p(f)$ can be measured only up to an arbitrary scaling factor $\alpha_p(f)$, the number of degrees of freedom m can be estimated in the ML sense by maximizing

$$\mathcal{L}_{IW} = \prod_p \prod_f p\big(\mathbf{R}_p(f)|\alpha_p(f), \mathbf{\Psi}_p(f), m\big) \tag{21}$$

where $p\big(\mathbf{R}_p(f)|\alpha_p(f), \mathbf{\Psi}_p(f), m\big) = J_{\alpha_p(f)}\mathcal{IW}\big(\alpha_p(f)\mathbf{R}_p(f)|\mathbf{\Psi}_p(f), m\big)$, $J_{\alpha_p(f)} = \alpha_p^{I^2}(f)$ is the Jacobian of the scaling transform, and $\mathbf{\Psi}_p(f)$ was computed by (20) for each geometric setting p. The optimal value of $\alpha_p(f)$ is chosen such that the likelohood $p\big(\mathbf{R}_p(f)|\alpha_p(f), \mathbf{\Psi}_p(f), m\big)$ is maximized, that is

$$\alpha_p(f) = \frac{\text{tr}\left(\mathbf{\Psi}_p(f)\mathbf{R}_p^{-1}(f)\right)}{Im} \tag{22}$$

By replacing (22) into (21) and discarding constants, the log-likelihood to be max-

imized is

$$\log \mathcal{L}_{IW} \stackrel{c}{=} \sum_{p,f} -Im \log\big(\alpha_p(f)\big) + m \log|\mathbf{\Psi}_p(f)| \\ - (m+I)\log|\mathbf{R}_p(f)| - Im - \sum_{i=1}^{I} \log\Gamma(m-i+1). \tag{23}$$

Given $\mathbf{\Psi}_p(f)$ and $\mathbf{R}_p(f)$ for all p, f, $\log \mathcal{L}_{IW}$ is then maximized using Matlab's fmincon Newton-based optimizer. As a result, the optimal value of m is found, which increases with the reverberation time as will be shown in Table 1.

2.2.3 MAP spatial parameter update

Given the hyper-parameters $\mathbf{\Psi}_j(f)$ and m, the spatial covariance matrices $\mathbf{R}_j(f)$ can be estimated in the MAP sense in the M-step of the Algorithm 1 by maximizing the expectation of the log-posterior of the complete data

$$Q_{\mathcal{IW}}(\theta|\theta^{\text{old}}) = \sum_{j,f} \Big(\sum_{n} \log p\big(\mathbf{c}_j(n,f)|\mathbf{0}, \mathbf{\Sigma}_j(n,f)\big) \\ + \gamma \log \mathcal{IW}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big)\Big) \tag{24}$$

where γ is a tradeoff hyper-parameter determining the strength of the prior, and $\mathcal{IW}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big)$ is defined in (19). By computing the partial derivatives of $Q_{\mathcal{IW}}(\theta|\theta^{\text{old}})$ with respect to each entry of $\mathbf{R}_j(f)$ and equating them to zero, we obtain the $\mathbf{R}_j(f)$ update in (11) as

$$\mathbf{R}_j(f) = \frac{1}{\gamma(m+I)+N}\Big(\gamma \mathbf{\Psi}_j(f) + \sum_{n=1}^{N} \frac{\widehat{\mathbf{\Sigma}}_j(n,f)}{v_j(n,f)}\Big) \tag{25}$$

where N is the total number of time frames. Note that when $\gamma = 0$, $i.e.$ the contribution of the prior is excluded, (25) becomes equal to the ML update.

Each iteration of the EM update mostly involves the computation of $(N+1)FJ$ inversions and $5NFJ$ multiplications of $I \times I$ matrices (see Algorithm 1). The overall computational complexity of one iteration is therefore $O(6NFJI^3)$. It is linear as a function of the number of sources and the duration of the signal, and cubic as a function of the number of channels. This MAP spatial parameter update does not increase the computational complexity compared to the ML case.

2.3 MAP spatial parameter update exploiting a Wishart prior

2.3.1 Wishart prior

We consider, as an alternative to the inverse-Wishart prior, a Wishart distribution over each spatial covariance matrix $\mathbf{R}_j(f)$ as

$$p(\mathbf{R}_j(f)) = \mathcal{W}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m\big) \tag{26}$$

where

$$\mathcal{W}(\mathbf{R}|\mathbf{\Psi}, m) = \frac{|\mathbf{\Psi}|^{-m}|\mathbf{R}|^{(m-I)}e^{-\text{tr}(\mathbf{\Psi}^{-1}\mathbf{R})}}{\pi^{I(I-1)/2}\prod_{i=1}^{I}\Gamma(m-i+1)} \tag{27}$$

is the Wishart density over a Hermitian positive definite matrix $\mathbf{R}$ with positive definite scale matrix $\mathbf{\Psi}$, m degrees of freedom and mean $m\mathbf{\Psi}$ [22]. This distribution, its mean, and its variance are finite for $m > I-1$, $m > I$, and $m > I+1$ respectively. We define the scale matrix $\mathbf{\Psi}_j(f)$ as

$$\mathbf{\Psi}_j(f) = \frac{\mathbf{h}_j^{\text{ane}}(f)\big(\mathbf{h}_j^{\text{ane}}(f)\big)^H + \sigma_{\text{rev}}^2\mathbf{\Omega}(f)}{m} \tag{28}$$

so that the mean of $\mathbf{R}_j(f)$ is coherent with (13).

2.3.2 Learning the hyper-parameter

Similarly to the inverse-Wishart prior, the number of degrees of freedom m, which determines the deviation of $\mathbf{R}_j(f)$ from its mean, is learned from training data by maximizing the log-likelihood (21), where $p\big(\mathbf{R}_p(f)|\alpha_p(f),\mathbf{\Psi}_p(f),m\big) = J_{\alpha_p(f)}\mathcal{W}\big(\alpha_p(f)\mathbf{R}_p(f)|\mathbf{\Psi}_p(f),m\big)$. The optimum value of $\alpha_p(f)$ is chosen in the ML sense similarly to the inverse-Wishart prior by

$$\alpha_p(f) = \frac{Im}{\text{tr}\left(\mathbf{\Psi}_p^{-1}(f)\mathbf{R}_p(f)\right)} \tag{29}$$

By replacing (29) into (21) and discarding constants, the log-likelihood to be maximized is

$$\begin{aligned}\log\mathcal{L}_W \stackrel{\text{c}}{=} &\sum_{p,f} Im\log\big(\alpha_p(f)\big) - m\log|\mathbf{\Psi}_p(f)| \\ &+ (m-I)\log|\mathbf{R}_p(f)| - Im - \sum_{i=1}^{I}\log\Gamma(m-i+1).\end{aligned} \tag{30}$$

Given $\mathbf{\Psi}_p(f)$ and $\mathbf{R}_p(f)$ for all p, f, $\log\mathcal{L}_W$ is then maximized using Matlab's fmincon Newton-based optimizer. As a result, the optimal value of m is found, which is equal to that of the inverse-Wishart prior, and shown in Table 1.

2.3.3 MAP spatial parameter update

The expectation of the log-posterior of the complete data is defined similar to (24) as

$$\begin{aligned}Q_{\mathcal{W}}(\theta|\theta^{\text{old}}) = \sum_{j,f}\Big(&\sum_n \log p\big(\mathbf{c}_j(n,f)|\mathbf{0},\mathbf{\Sigma}_j(n,f)\big) \\ &+ \gamma\log\mathcal{W}\big(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f),m\big)\Big)\end{aligned} \tag{31}$$

By computing the partial derivatives of $Q_{\mathcal{W}}(\theta|\theta^{\text{old}})$ with respect to each entry of $\mathbf{R}_j(n,f)$ and equating them to zero, we obtain a quadratic matrix equation. By solving this equation with a positive definite solution constraint, we obtain the spatial covariance update as

$$\mathbf{R}_j(f) = \frac{1}{2}\mathbf{T}^{-1/2}\big(-b\mathbf{I} + (b^2\mathbf{I} - 4\mathbf{T}^{1/2}\mathbf{C}\mathbf{T}^{1/2})^{1/2}\big)\mathbf{T}^{-1/2} \tag{32}$$

where $(.)^{1/2}$ denotes the square root of a Hermitian matrix, and

$$\begin{aligned} \mathbf{T} &= \gamma \mathbf{\Psi}_j^{-1}(f) \\ b &= -\gamma(m-I)+N \\ \mathbf{C} &= \sum_{n=1}^{N} -\frac{\widehat{\mathbf{\Sigma}}_j(n,f)}{v_j(n,f)}. \end{aligned} \tag{33}$$

3 Source-based EM algorithms

Motivated by the alternative implementation in [16] of the full-rank source separation framework, we investigate in this section a third MAP parameter update exploiting Gaussian prior. We first present the general source-based EM implementation introduced in [16], we then define a Gaussian prior over the mixing parameter and derive the corresponding MAP parameter update.

3.1 General EM implementation

This implementation generalizes the Gaussian modeling framework [9, 10] that applicable to both the full-rank and the rank-1 spatial covariance matrices. It is based on the non-unique representation of the spatial covariance matrix as [16]

$$\mathbf{R}_j(f) = \mathbf{A}_j(f)\mathbf{A}_j^H(f) \tag{34}$$

where $\mathbf{A}_j(f)$ is an $I \times R_j$ complex-valued matrix of rank R_j. For j-th source STFT coefficients $s_j(n,f)$, R_j independent Gaussian random variables $s_{jr}(n,f), r = 1, ..., R_j$ distributed as $s_{jr}(n,f) \sim \mathcal{N}_c(0, v_j(n,f))$ are introduced. With these notations, each j-th source is a mixture of R_j point *sub-sources* with mixing matrix $\mathbf{A}_j(f)$. Overall the mixture STFT coefficients are written as

$$\mathbf{x}(n,f) = \mathbf{A}(f)\mathbf{s}(n,f) \tag{35}$$

where $\mathbf{s}(n,f) = [s_{11}(n,f), ..., s_{1R_1}(n,f), ..., s_{JR_j}(n,f)]^T$ is an $R \times 1$ vector of sub-source coefficients with $R = \sum_{j=1}^{J} R_j$, and $\mathbf{A}(f) = [\mathbf{A}_1(f), ..., \mathbf{A}_J(f)]$ an $I \times R$ mixing matrix.

Similarly to [23], the EM parameter estimation is derived from the noisy mixture model, since the estimated mixing vectors remain fixed to their initial value when considering 35, that is

$$\mathbf{x}(n,f) = \mathbf{A}(f)\mathbf{s}(n,f) + \mathbf{b}(n,f) \tag{36}$$

where $\mathbf{b}(n,f)$ represents some additive zero-mean Gaussian noise with covariance matrix $\mathbf{\Sigma}_\mathbf{b}(n,f) = \sigma_b^2(n,f)\mathbf{I}$. EM is separately derived for each frequency bin f for the *complete data* $\{\mathbf{x}(n,f), s_j(n,f)\}_{j,n}$ that is the set of observed mixture STFT coefficients and hidden source STFT coefficients of all time frames. The details of one iteration are summarized in Algorithm 2 where in the E-step, the Wiener filter $\mathbf{W}_j(n,f)$ and the conditional expectations of the natural statistics $\widehat{\mathbf{R}}_{\mathbf{ss}}(n,f)$, $\widehat{\mathbf{R}}_{\mathbf{xs}}(n,f)$ are computed, then in the M-step, $v_j(n,f)$ and $\mathbf{A}(f)$ are alternatingly updated.

where $[]_{kk}$ denotes k-th diagonal entry in a matrix, $\tilde{v}_r(n,f) = v_j(n,f$ if and only if $r \in \mathcal{R}_j$, where $\mathcal{R}_j$ denotes the set of sub-source indices associated with j-th source. In the ML parameter estimation, $\mathbf{A}(f)$ is updated in (43) as

$$\mathbf{A}(f) = \Big(\sum_n \widehat{\mathbf{R}}_{\mathbf{xs}}(n,f)\Big)\Big(\sum_n \widehat{\mathbf{R}}_{\mathbf{ss}}(n,f)\Big)^{-1} \tag{44}$$

Algorithm 2 General source-based EM algorithm

E step:

$$\boldsymbol{\Sigma}_{\mathbf{s}}(n,f) = \mathrm{diag}\left([\tilde{v}_r(n,f)]_{r=1}^{R}\right) \tag{37}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}}(n,f) = \mathbf{A}(f)\boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{A}^H(f) + \boldsymbol{\Sigma}_{\mathbf{b}}(n,f) \tag{38}$$

$$\mathbf{W}(n,f) = \boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{A}^H(f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f) \tag{39}$$

$$\begin{aligned}\widehat{\mathbf{R}}_{\mathbf{ss}}(n,f) = {} & \mathbf{W}(n,f)\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{W}^H(n,f) \\ & + (\mathbf{I}_R - \mathbf{W}(n,f)\mathbf{A}(f))\boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\end{aligned} \tag{40}$$

$$\widehat{\mathbf{R}}_{\mathbf{xs}}(n,f) = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{W}^H(n,f) \tag{41}$$

M step:

$$v_j(n,f) = \frac{1}{R_j}\sum_{r\in\mathcal{R}_j}[\widehat{\mathbf{R}}_{\mathbf{ss}}(n,f)]_{rr} \tag{42}$$

$$\text{Update } \mathbf{A}(f). \tag{43}$$

In the following, we will derive a new update for $\mathbf{A}(f)$ when considering a Gaussian prior.

3.2 MAP spatial parameter update exploiting a Gaussian prior

3.3 Gaussian prior

We model each column vector $\mathbf{a}_{jr}(f)$, $r = 1,...,R_j$ of $\mathbf{A}_j(f)$ as a multivariate complex-valued Gaussian distribution

$$p\big(\mathbf{a}_{jr}(f)\big) = \mathcal{N}_c\big(\bar{\mathbf{a}}_{jr}(f), \boldsymbol{\Sigma}_{jr}(f)\big) \tag{45}$$

where $\bar{\mathbf{a}}_{jr}(f)$ and $\boldsymbol{\Sigma}_{jr}(f)$ denote the mean and the covariance matrix of $\mathbf{a}_{jr}(f)$, respectively, which are computed from the theory of statistical room acoustics given the geometric setting. Note that in the particular case when $R_j = 1$, $\mathbf{A}_j(f)$ becomes a mixing vector and $\mathbf{R}_j(f)$ becomes a rank-1 matrix as parameterized under the narrowband approximation. Therefore, as opposed to the inverse-Wishart and Wishart priors investigated for the source image-based algorithms, this prior has the advantage of being valid both for the rank-1 and full-rank parameterizations of $\mathbf{R}_j(f)$. However since the full-rank spatial covariance matrices were shown to better approximate the reverberant mixing process [10], we will not consider the rank-1 case in this paper.

Let us consider a rank-2 case, *i.e.* $R_j = 2$, $\forall j, f$, where $\mathbf{a}_{j1}(f)$ represents the contribution of the direct part and the early echoes from the j-th source to all microphones while $\mathbf{a}_{j2}(f)$ represents the contribution of the late reverberant part. Since the early echoes and the late reverberation can be assumed to be diffuse, the mean coefficients of their mixing vectors are zeros. In other words, $\bar{\mathbf{a}}_{j1}(f)$ are then equal $\mathbf{h}_j^{\text{ane}}(f)$ computed in (14) and $\bar{\mathbf{a}}_{j2}(f) = \mathbf{0}$.

The covariance matrices $\boldsymbol{\Sigma}_{j1}(f)$ and $\boldsymbol{\Sigma}_{j2}(f)$ can also be inferred from the theory

of statistical room acoustics as

$$\begin{aligned}\boldsymbol{\Sigma}_{j1}(f) &= \mathbb{E}\{\mathbf{a}_{j1}(f)\mathbf{a}_{j1}^H(f)\} - \bar{\mathbf{a}}_{j1}(f)\bar{\mathbf{a}}_{j1}^H(f) \\ &= \sigma_{\text{ee}}^2\boldsymbol{\Omega}(f) \qquad (46)\\ \boldsymbol{\Sigma}_{j2}(f) &= \mathbb{E}\{\mathbf{a}_{j2}(f)\mathbf{a}_{j2}^H(f)\} - \bar{\mathbf{a}}_{j2}(f)\bar{\mathbf{a}}_{j2}^H(f) \\ &= \sigma_{\text{re}}^2\boldsymbol{\Omega}(f) \qquad (47)\end{aligned}$$

where $\boldsymbol{\Omega}(f)$ represents the normalized spatial covariance matrix of a diffuse noise whose entries are given by (15), σ_{ee}^2 and σ_{re}^2 represent the power of early echoes and of the late reverberant part, respectively, and

$$\sigma_{\text{ee}}^2 + \sigma_{\text{re}}^2 = \sigma_{\text{rev}}^2 \qquad (48)$$

where the power of the non-direct part σ_{rev}^2 is computed by (16).

3.4 Learning the hyper-parameter

In order to compute the covariances of the prior distributions, we learn the optimum values of σ_{ee}^2 in (46) from training data and derive σ_{re}^2 via (16). The training data originally consist of the spatial covariance matrices $\mathbf{R}_p(f)$ computed from many source and microphone positions p as for the inverse-Wishart and Wishart priors. Under the assumption that $\mathbf{a}_{p1}(f)$ concentrates most of the power, $\mathbf{a}_{p1}(f)$ can be computed by the first eigenvector corresponding to the largest eigenvalue of $\mathbf{R}_{(f)p}$ while $\mathbf{a}_{p2}(f)$ is the second eigenvector of $\mathbf{R}_p(f)$.

Let us denote by

$$\underline{\mathbf{a}}_p(f) = [\mathbf{a}_{p1}^T(f), \mathbf{a}_{p2}^T(f)]^T \qquad (49)$$

the $I^2 \times 1$ vectorization of $\mathbf{A}_p(f)$ whose mean is

$$\underline{\bar{\mathbf{a}}}_p(f) = [\bar{\mathbf{a}}_{p1}^T(f), \bar{\mathbf{a}}_{p2}^T(f)]^T \qquad (50)$$

and whose covariance is

$$\boldsymbol{\Sigma}_p(f) = \begin{pmatrix} \boldsymbol{\Sigma}_{p1}(f) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{p2}(f) \end{pmatrix} \qquad (51)$$

Since $\underline{\mathbf{a}}_p(f)$ can be measured only up to an arbitrary complex-valued factor $\alpha_p(f)$, σ_{ee} is then estimated in the ML sense by maximizing

$$\mathcal{L}_G = \prod_p \prod_f p\big(\underline{\mathbf{a}}_p(f)|\alpha_p(f), \underline{\bar{\mathbf{a}}}_p(f), \boldsymbol{\Sigma}_p(f)\big) \qquad (52)$$

where $p\big(\underline{\mathbf{a}}_p(f)|\alpha_p(f), \underline{\bar{\mathbf{a}}}_p(f), \boldsymbol{\Sigma}_p(f)\big) = J_{\alpha_p(f)}\mathcal{N}_c\big(\alpha_p(f)\underline{\mathbf{a}}_p(f)|\underline{\bar{\mathbf{a}}}_p(f), \boldsymbol{\Sigma}_p(f)\big)$, $J_{\alpha_p(f)} = |\alpha_p(f)|^{2I^2}$ is the Jacobian of the scaling transform. The optimal value of $\alpha_p(f)$ is chosen such that the likelohood $p\big(\mathbf{a}_p(f)|\alpha_p(f), \underline{\bar{\mathbf{a}}}_p(f), \boldsymbol{\Sigma}_p(f)\big)$, is maximized, that is

$$\alpha_p(f) = \frac{-|b_2| - (|b_2|^2 - 4b_1b_3)^{1/2}}{2b_1}\frac{b_2}{|b_2|} \qquad (53)$$

where

$$b_1 = -\underline{\mathbf{a}}_p^H(f)\boldsymbol{\Sigma}_p^{-1}(f)\underline{\mathbf{a}}_p(f) \qquad (54)$$

$$b_2 = \underline{\mathbf{a}}_p^H(f)\boldsymbol{\Sigma}_p^{-1}(f)\underline{\bar{\mathbf{a}}}_p(f) \qquad (55)$$

$$b_3 = I^2 \qquad (56)$$

Given $\underline{\mathbf{a}}_p(f)$ and $\bar{\underline{\mathbf{a}}}_p(f)$ for all p, f, $\log \mathcal{L}_G$ is then maximized using Matlab's fmincon Newton-based optimizer. As a result, the optimal value of σ^2_{ee} is found and shown in Table 1.

3.4.1 MAP spatial parameter update

Let us concatenate the column vectors of $\mathbf{A}(f)$ into a $R \times 1$ column vector $\underline{\mathbf{A}}(f)$. The prior distribution (45) is then translated to

$$p\big(\underline{\mathbf{A}}(f)\big) = \mathcal{N}_c\big(\bar{\underline{\mathbf{A}}}(f), \mathbf{\Sigma}_{\underline{\mathbf{A}}}(f)\big) \tag{57}$$

where the entries of the hyper-parameters $\bar{\underline{\mathbf{A}}}(f)$ and $\mathbf{\Sigma}_{\underline{\mathbf{A}}}(f)$ are taken from $\bar{\mathbf{a}}_{jr}(f)$ and $\mathbf{\Sigma}_{jr}(f), \forall j, r$, respectively.

The MAP update of $\mathbf{A}(f)$ in the M-step of the EM algorithm is derived by maximizing the expectation of the log-posterior of the complete data

$$\begin{aligned} Q_{\mathcal{G}}(\theta|\theta^{\text{old}}) = &\sum_{n,f} \log \mathcal{N}_c\big(\mathbf{x}(n,f)|\mathbf{A}(f)\mathbf{s}(n,f), \sigma_b(n,f)\mathbf{I}\big) \\ &+ \gamma \log p\big(\underline{\mathbf{A}}(f)\big) \end{aligned} \tag{58}$$

where γ is a tradeoff hyper-parameter determining the strength of the prior. By computing the gradient of $Q_{\mathcal{G}}(\theta|\theta^{\text{old}})$ with respect to $\underline{\mathbf{A}}(f)$ and equating it to zero, we obtain the MAP update of $\underline{\mathbf{A}}(f)$ as

$$\begin{aligned} \underline{\mathbf{A}}(f) = &\left(\sum_n \frac{1}{\sigma_b^2(n,f)} (\widehat{\mathbf{R}}_{\mathbf{ss}}(n,f) \otimes \mathbf{I}_{R_j})^T + \gamma \mathbf{\Sigma}_{\underline{\mathbf{A}}}^{-1}\right)^{-1} \times \\ &\left(\sum_n \frac{1}{\sigma_b^2(n,f)} \text{vec}(\widehat{\mathbf{R}}_{\mathbf{xs}}(n,f)) + \gamma \mathbf{\Sigma}_{\underline{\mathbf{A}}}^{-1} \bar{\underline{\mathbf{A}}}(f)\right) \end{aligned} \tag{59}$$

where $\otimes$ is the Kronecker product between matrices, $\mathbf{I}_{R_j}$ is the identity matrix of size R_j, and $\text{vec}(\mathbf{X})$ concatenates column vectors of matrix $\mathbf{X}$ to a column vector. The mixing matrix $\mathbf{A}(f)$ is then derived from $\underline{\mathbf{A}}(f)$

4 Experimental evaluation

4.1 Data and evaluation criteria

We use simulated mixtures for both training and testing data since they allow the generation of a wide range of recording configurations while keeping precise control over the configuration parameters, including the geometric setting and the room characteristics. We first generated room impulse responses via the *image method* [24] from three source positions to microphone pair positions using the Roomsim toolbox[1]. The room dimensions were $4.45 \times 3.55 \times 2.5$ m, that are the dimensions used in the SiSEC campaign [7], the microphone spacing and the source-to-microphone distances was fixed to $d = 5$ cm and $r = 50$ cm, respectively. Four different reverberation times were considered: $T_{60} = 50$, 130, 250 and 500 ms. The source images were then computed by convolving 10 s speech signals with the simulated impulse responses.

[1] http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip

Training data: we generated room impulse responses for 20 random source positions for each of 20 random microphone pair positions. This resulted in a total of 400 source image signals indexed by p for each reverberation time.

Testing data: The positions of the sources and the microphones are illustrated in Fig. 1. Two sets of speech signals were considered: male and female speech. This resulted in two stereo mixtures for each T_{60} and 8 mixtures in total.

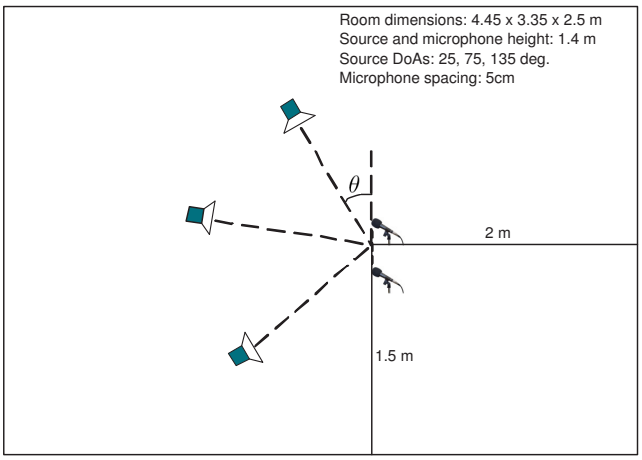


Figure 1: Room geometric setting for testing data.

We evaluate the performance of the algorithm via the widely used energy ratio criteria, which can be applied to any audio mixture and any algorithm and do not require the knowledge of the unmixing parameters or filters. These criteria include the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR) and source image-to-spatial distortion ratio (ISR) criteria expressed in decibels (dB), as defined in [25]. They account respectively for overall distortion of the target source, residual crosstalk from other sources, musical noise and spatial or filtering distortion of the target. These criteria were implemented in Matlab and distributed for public use[2].

4.2 Result for source image-based EM algorithms

We compare the separation performance obtained by the MAP-based spatial covariance update exploiting the inverse-Wishart prior presented in Section 2.2, named MAP inverse-Wishart, and by the update resulting from the Wishart prior presented in Section 2.3, named MAP Wishart, with that achieved by the ML algorithms where $\mathbf{R}_j(f)$ were either blindly initialized via hierarchical clustering as presented in [10], named ML blind init, or initialized from the known geometric setting by (13), named ML geom. init. We used 10 EM iterations for all algorithms. The STFT was computed with a sine window of length 1024 and the empirical mixture covariances $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)$ and $\widehat{\mathbf{\Sigma}}_p(n,f)$ were computed with a bidirectional window w_{nf} of size 3×3. The trade-off parameter γ, which determines the strength of the priors, does not significantly affect the result but we observed that $\gamma = 100$ is globally a good choice. The optimal number of degrees of freedom m of the inverse-Wishart and Wishart priors was learned from training data and is shown in Table 1 together with the mean power σ^2_{rev}

[2] http://bass-db.gforge.inria.fr/bss_eval/

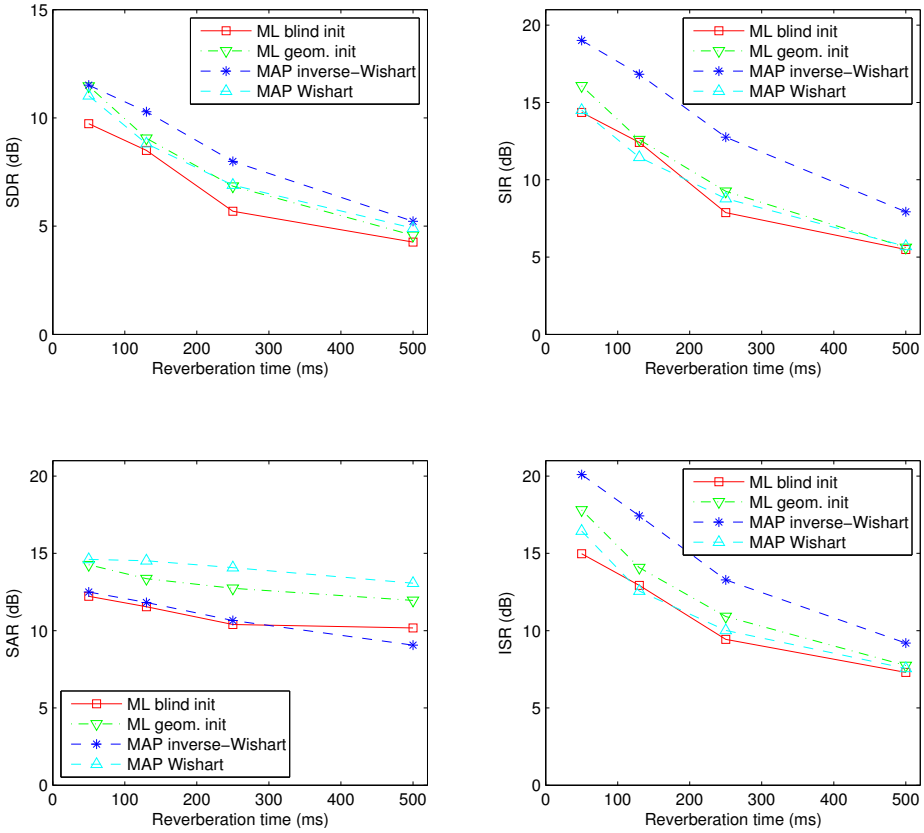


Figure 2: Separation performance of the source image-based algorithms on simulated speech mixtures as a function of the reverberation time.

of echoes and reverberation computed by (16), which both depend on the reverberation time.

T_{60}	50 ms	130 ms	250 ms	500 ms
m	2.1	2.1	3.4	5.3
σ^2_{ee}	0.009	0.033	0.068	0.148
σ^2_{rev}	0.002	0.024	0.063	0.139

Table 1: Learned value of the priors' hyper-parameters and predicted value of σ^2_{rev}

As expected, σ^2_{rev} strongly increases with reverberation, such that the direct-to-reverberant energy ratio is 14 dB lower when $T_{60} = 500$ ms than when $T_{60} = 50$ ms. The variance of the inverse-Wishart prior, which is inversely related to m [22], decreases with reverberation time while that of the Wishart prior, on the contrary, increases with reverberation time.

The separation performance of all algorithms, as evaluated by the energy ratio criteria and averaged for all sources as a function of the reverberation time is shown in Fig. 2. The ML algorithm initialized from known geometry settings results in better performance in terms of all criteria than the blindly initialized ML algorithm at all re-

verberation times. The MAP Wishart algorithm offers the best SAR and very similar SDR to the ML geom. init algorithm. Overall, the MAP inverse-Wishart algorithm outperforms all other algorithms for all considered reverberation times in terms of SDR, SIR, and ISR. For instance, at $T_{60} = 250$ ms, it improves the SDR by 2.3 dB, 1.1 dB and 1.1 dB compared to the ML blind init, the ML geom. init and the MAP Wishart algorithms, respectively. This confirms the benefit of the proposed inverse-Wishart spatial location prior and the associated MAP algorithm.

4.3 Result for source-based EM algorithms

In this experiment, we compare the separation performance obtained by the MAP-based mixing parameter update exploiting the Gaussian prior, named MAP Gaussian, with that achieved by the ML algorithms where the mixing matrix $\mathbf{A}(f)$ were either blindly initialized via hierarchical clustering, named ML blind init, or initialized from the known geometric setting by reshaping $\underline{\bar{\mathbf{A}}}(f)$, named ML geom. init. Parameter settings are the same as for the source image-based algorithms except that the number of EM iteration is increased to 30, since the source-based EM algorithms converge slower than the source image-based EM algorithms, and $\gamma = 10$. The optimum power of the early echoes σ^2_{ee} of the Gaussian prior was learned from training data and shown in Table 1. As expected, the ratio between the power of early echoes and non-direct part $\sigma^2_{\mathrm{ee}}/\sigma^2_{\mathrm{rev}}$ decreases with the reverberation time due to the stronger reverberation.

The separation performance of all algorithms, as evaluated by the energy ratio criteria and averaged for all sources as a function of the reverberation time is shown in Fig. 3. The ML algorithm initialized from known geometry settings results in significantly better performance in terms of all criteria than the blindly initialized ML algorithm at all reverberation times. The MAP Gaussian algorithm offers slightly better separation performance than the ML geom. init. Overall, the MAP Gaussian algorithm outperforms all other algorithms for all considered reverberation times in terms of all criteria. For instance, at $T_{60} = 250$ ms, it improves the SDR by 4.2 dB and 0.3 dB compared to the ML blind init and the ML geom. init, respectively. This confirms the benefit of the proposed Gaussian spatial location prior and the associated MAP algorithm.

5 Conclusion

In this article, we considered two classes of source separation algorithms grounded on an emerging Gaussian modeling framework. While ML parameter estimation algorithms existed, we derived the MAP algorithms exploiting prior knowledge about the source locations and the geometric setting. For that purpose, we investigated three different prior distributions over the spatial parameters namely the inverse-Wishart prior, the Wishart prior and the Gaussian prior, whose means are given by the theory of statistical room acoustics and whose variances are learned from training data. These priors both result in closed-form MAP parameter update and, more interestingly, the MAP algorithms do not suffer from the permutation problem thanks to this prior information about the source location. Experimental results over several reverberation conditions confirm the benefit of the proposed approach compared to other ML algorithms exploiting the same prior knowledge.

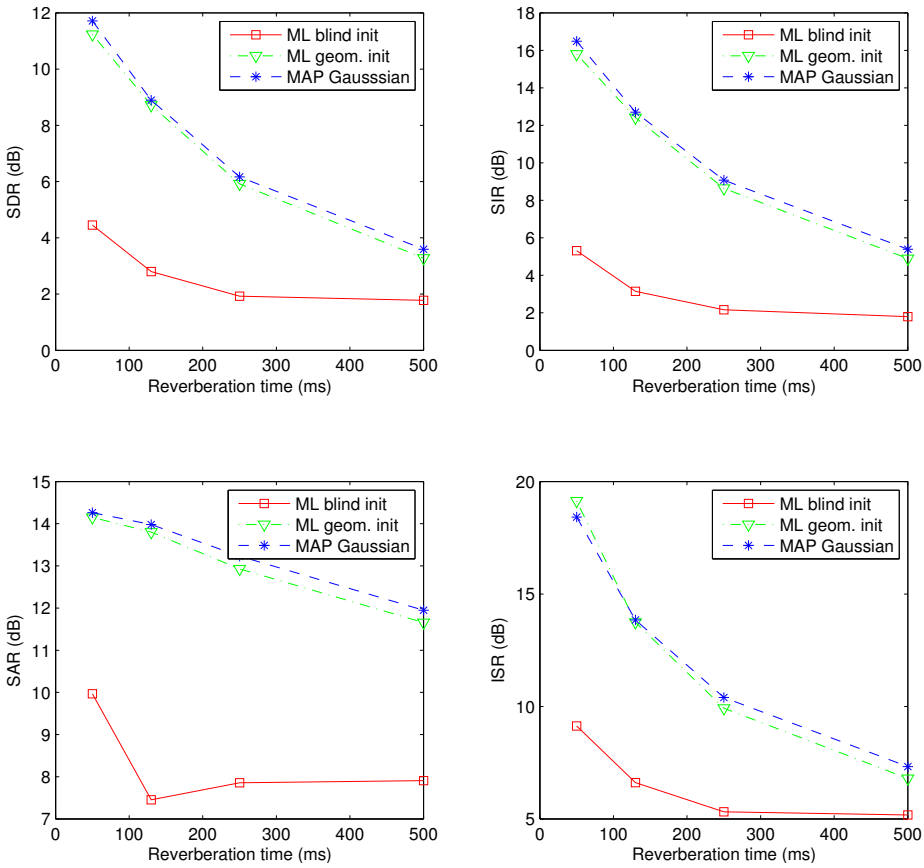


Figure 3: Separation performance of the source-based algorithms on simulated speech mixtures as a function of the reverberation time.

References

[1] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998, pp. 1941–1944.

[2] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007.

[3] P. O'Grady, B. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, pp. 18–33, 2005.

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

[5] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, article ID 24717, 2007.

[6] E. Vincent, S. Araki, and P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 734–741.

[7] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[8] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 78–81.

[9] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI Global, 2010, pp. 162–185.

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

[11] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. on Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[12] 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 Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[13] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 715–726, 2007.

[14] L. Parra and C. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[15] Q. Pan and T. Aboulnasr, "Combined spatial/beamforming and time/frequency processing for blind source separation," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2008.

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

[17] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Sep. 2010, pp. 73–80.

[18] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. New York, NY: Wiley, 1997.

[19] H. Kuttruff, *Room Acoustics*, 4th ed. New York: Spon Press, 2000.

[20] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 791–803, 2003.

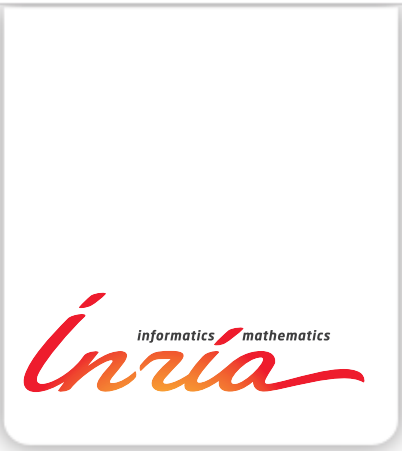
[21] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial covariance models for under-determined reverberant audio source separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 129–132.

[22] D. Maiwald and D. Kraus, "Calculation of moments of complex Wishart and complex inverse-Wishart distributed matrices," *IEE Proceedings on Radar, Sonar and Navigation*, vol. 147, pp. 162–168, 2000.

[23] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[24] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[25] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, algorithms and results," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 552–559.



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399