Spatial location priors for Gaussian model based reverberant audio source separation

Ngoc Duong, Emmanuel Vincent, Rémi Gribonval

Spatial location priors for Gaussian model based reverberant audio source separation

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

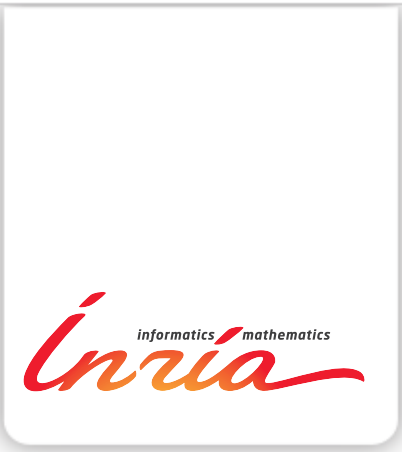
RESEARCH REPORT

N° 8057

September 2012

Project-Team 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-Team Metiss

Research Report n° 8057 — version 2 — initial version September 2012 — revised version April 2013 — 20 pages

Abstract: We consider the Gaussian framework for reverberant audio source separation, where the sources are modeled in the time-frequency domain by their short-term power spectra and their spatial covariance matrices. We propose three alternative probabilistic priors over the spatial covariance matrices which are consistent with the theory of statistical room acoustics and we derive Expectation-Maximization (EM) algorithms for maximum a posteriori (MAP) estimation. We argue that these algorithms provide a statistically principled solution to the permutation problem and to the risk of overfitting resulting from conventional maximum likelihood (ML) estimation. We show experimentally that, in a semi-informed scenario where the source positions and certain room characteristics are known, the algorithms using respectively inverse-Wishart and Gaussian priors outperform their ML counterparts. This opens the way to rigorous statistical treatment of this family of models in other scenarios in the future.

Key-words: audio source separation, spatial covariance, EM algorithm, probabilistic priors, inverse-Wishart, Gaussian

This is the preprint of an article submitted to the EURASIP Journal on Advances in Signal Processing.

* 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 : Nous nous plaçons dans le cadre gaussien pour la séparation de mélanges réverbérants de sources audio, où les sources sont modélisées dans le domaine temps-fréquence par leurs spectres de puissance à court terme et leurs matrices de covariance spatiale. Nous proposons trois distributions *a priori* différentes sur les matrices de covariance spatiale qui sont cohérentes avec la théorie statistique de l'acoustique des salles et nous concevons des algorithmes d'Espérance-Maximisation (EM) pour l'estimation au sens du maximum *a posteriori* (MAP). Nous soutenons que ces algorithmes fournissent une solution statistiquement fondée au problème de permutation et au risque de sur-apprentissage inhérent à l'estimation classique au sens du maximum de vraisemblance (MV). Nous montrons expérimentalement que, dans un scénario semi-informé où les positions des sources et certaines caractéristiques de la pièce sont connues, les algorithmes utilisant respectivement des *a priori* inverse-Wishart et gaussien fournissent une meilleure performance que les algorithmes MV correspondants. Cela ouvre la voie à un traitement statistiquement rigoureux de cette famille de modèles dans d'autres scénarios à l'avenir.

Mots-clés : séparation de sources audio, covariance spatiale, algorithme EM, distributions *a priori*, inverse-Wishart, gaussienne

1 Introduction

We consider the task of reverberant audio source separation, that is to extract individual sound sources from a multichannel microphone array recording. Many approaches have been proposed in the literature, which typically operate in the time-frequency domain via the short-time Fourier transform (STFT) [1, 2, 3]. One category of approaches model the mixture STFT coefficients as the product of the source STFT coefficients and complex-valued *mixing vectors*, which are estimated by frequency-domain independent component analysis (FDICA) [4, 5] or by clustering [6, 7]. In under-determined conditions when the number of sources is greater than the number of channels, the source STFT coefficients are then obtained via binary masking [6], soft masking [7] or ℓ_1-norm minimization [8]. Lately, a Gaussian framework has emerged where the mixture STFT coefficients are modeled as a function of the power spectra and the *spatial covariance matrices* of the sources and separation is achieved by multichannel Wiener filtering [9, 10, 11]. These covariance matrices may equivalently be expressed as the outer product of *subsource mixing matrices*, which reduce to mixing vectors when the spatial covariance matrices have rank 1 [12]. Full-rank matrices have been shown to improve separation performance in reverberant conditions by modeling not only the spatial position of the sources but also their spatial width [11].

While a number of deterministic [13, 14, 12] and probabilistic [15, 16, 17] priors have been proposed over the source spectra, the mixing vectors and the source spatial covariance matrices are usually estimated in an unconstrained manner. The lack of a constraint relating these quantities across frequency causes a permutation problem, which has been coped with by reordering the estimates in each frequency bin while keeping their value [18, 7]. More crucially, the estimated values of the mixing vectors and the source spatial covariance matrices in a given frequency bin are likely to suffer from overfitting when the corresponding sources are little active in that bin.

Building upon the studies for instantaneous mixtures in [19, 20] and the deterministic subspace constraints in [21, 22], a few algorithms have been designed that exploit soft penalties or probabilistic priors over the mixing vectors for increased estimation accuracy. These algorithms typically target semi-informed scenarios such as formal meetings or in-car speech where the spatial locations of the sources are known and they rely on the assumption that the mixing vectors are close to the steering vectors representing the direct path from the sources to the microphones. Squared Euclidean penalties over the blocking vectors are a common choice for FDICA [21, 23]. An inverse-Wishart prior over the outer product of the mixing vectors was also employed in [24]. These penalties and priors were not designed according to the actual statistics of reverberation. Moreover, to the best of our knowledge, no such priors have been designed for full-rank matrices.

In this article, we propose three probabilistic priors over the source spatial covariance matrices or the subsource mixing matrices which are consistent with the theory of statistical room acoustics[1]. We extend the two Gaussian Expectation-Maximization (EM) algorithms in [26, 12] so as to perform maximum a posteriori (MAP) estimation and we compare the resulting separation performance with conventional maximum likelihood (ML) estimation in an under-determined full-rank semi-informed scenario where the source positions and certain room characteristics are known. For clarity, we do not assume any other constraint on the model parameters, which allows us to assess the improvement resulting from these priors alone.

[1] The proposed inverse-Wishart prior was briefly introduced in our preliminary paper [25].

The structure of the article is as follows. In Section 2, we recall the Gaussian framework for audio source separation and we present a result of the theory of statistical room acoustics. We introduce two EM algorithms using inverse-Wishart and Wishart priors in Section 3 and an EM algorithm using a Gaussian prior in Section 4. We evaluate their separation performance in Section 5 and we conclude in Section 6.

2 Gaussian modeling and statistical room acoustics

2.1 Gaussian modeling for source separation

Let us consider a mixture signal $\mathbf{x}(t) = [x_1(t), ..., x_I(t)]^T$ recorded by an array of I microphones. Denoting by J the number of sources, the mixing process is expressed as [27]

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

where $\mathbf{c}_j(t) = [c_{1j}(t), ..., c_{Ij}(t)]^T$ is the *spatial image* of the j-th source, that is its contribution to the signals recorded at the microphones. The STFT coefficients $\mathbf{c}_j(n, f)$ of the source spatial images in each time frame n and each frequency bin f are modeled as zero-mean Gaussian random vectors

$$\mathbf{c}_j(n, f) \sim \mathcal{N}(\mathbf{0}, v_j(n, f)\, \mathbf{R}_j(f)) \tag{2}$$

where $v_j(n, f)$ are scalar nonnegative *variances* encoding the short-term power spectra of the sources and $\mathbf{R}_j(f)$ are $I \times I$ spatial covariance matrices encoding their spatial position and their spatial width [9, 11].

Under the assumption that the sources are uncorrelated, the mixture covariance matrix $\mathbf{\Sigma}_\mathbf{x}(n, f)$ is equal to

$$\mathbf{\Sigma}_\mathbf{x}(n, f) = \sum_{j=1}^{J} v_j(n, f)\, \mathbf{R}_j(f). \tag{3}$$

The log-likelihood is then given by [26]

$$\log \mathcal{L} = \sum_{n,f} - \mathrm{tr}(\mathbf{\Sigma}_\mathbf{x}^{-1}(n, f)\widehat{\mathbf{R}}_\mathbf{x}(n, f)) - \log |\pi \mathbf{\Sigma}_\mathbf{x}(n, f)| \tag{4}$$

where $\mathrm{tr}(.)$ and $|.|$ denote the trace and the determinant of a square matrix and $\widehat{\mathbf{R}}_\mathbf{x}(n, f)$ is the *empirical mixture covariance matrix* obtained by local averaging of $\mathbf{x}(n, f)\mathbf{x}^H(n, f)$ over the neighborhood of each time-frequency bin

$$\widehat{\mathbf{R}}_\mathbf{x}(n, f) = \sum_{n', f'} w_{nf}^2(n', f')\mathbf{x}(n', f')\mathbf{x}^H(n', f') \tag{5}$$

where w_{nf} is a bi-dimensional window specifying the shape of the neighborhood [26].

Source separation can then be achieved by estimating the model parameters $\theta = \{v_j(n, f), \mathbf{R}_j(f)\}$ in the ML sense and by deriving the spatial images of all sources in the minimum mean square error (MMSE) sense via multichannel Wiener filtering of the mixture STFT coefficients $\mathbf{x}(n, f)$

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

2.2 A result from the theory of statistical room acoustics

In a scenario such as in [21, 22, 23], the relative positions of the sources and the microphones with respect to each other are assumed to be known but their absolute position in the room is unknown. According to the theory of statistical room acoustics [28, 29], the mean spatial covariance matrix of a source over all possible absolute positions in the room can be expressed as

$$\boldsymbol{\mu}_{\mathbf{R}_j}(f) = \mathbf{d}_j(f)\mathbf{d}_j^H(f) + \sigma^2_{\mathrm{rev}}\boldsymbol{\Omega}(f) \tag{7}$$

where $.^H$ denotes conjugate transposition. The first term of this expression models the contribution of direct sound, where

$$\mathbf{d}_j(f) = \begin{pmatrix} \frac{1}{\sqrt{4\pi}r_{1j}} e^{-2i\pi f \frac{r_{1j}}{c}} \\ \vdots \\ \frac{1}{\sqrt{4\pi}r_{Ij}} e^{-2i\pi f \frac{r_{Ij}}{c}} \end{pmatrix} \tag{8}$$

is the steering vector modeling the direct paths from the source to the microphones, with c the sound velocity and r_{ij} the distance from the j-th source to the i-th microphone. The second term of this expression models the contribution of echoes and reverberation, which are assumed to come from all possible directions on average over all absolute positions: σ^2_{rev} is the power of echoes and reverberation and $\boldsymbol{\Omega}(f)$ is the covariance matrix of a diffuse sound field.

The entries $\Omega_{ii'}(f)$ of $\boldsymbol{\Omega}(f)$ depend on the microphone directivity patterns and on the distance $d_{ii'}$ between the i-th and the i'-th microphone. For omni-directional microphones, this quantity can be shown to be real-valued and equal to [28]

$$\Omega_{ii'}(f) = \frac{\sin(2\pi f d_{ii'}/c)}{2\pi f d_{ii'}/c}. \tag{9}$$

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

$$\sigma^2_{\mathrm{rev}} = \frac{4\beta^2}{\mathcal{A}(1-\beta^2)} \tag{10}$$

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 [29]

$$\beta = \exp\left\{ -\frac{13.82}{(\frac{1}{L_x}+\frac{1}{L_y}+\frac{1}{L_z})cT_{60}} \right\}. \tag{11}$$

In order to match the physics of reverberation, a prior over the source spatial covariance matrices or over the subsource mixing matrices should lead to a mean spatial covariance matrix $\boldsymbol{\mu}_{\mathbf{R}_j}(f)$ satisfying the constraint (7). This is not the case of the prior in [24], whose mean is equal to $\mathbf{d}_j(f)\mathbf{d}_j^H(f) + \epsilon\mathbf{I}_I$ with $\mathbf{I}_I$ the identity matrix of size I and ϵ a small constant. Isotropic Gaussian priors over the subsource mixing matrices would not satisfy this constraint either, due to the interchannel correlation introduced by $\boldsymbol{\Omega}(f)$. Fixed spatial covariance matrices set to the value in (7) were employed for single source localization in [29] and for source separation in [30]. Later work confirmed that the model (7) is valid on average over all absolute positions in the room but that $\mathbf{R}_j(f)$ varies with the absolute position, so that it must be estimated from the observed mixture signal [11].

Algorithm 1 SIEM algorithm [26]

E step:

$$\mathbf{\Sigma}_{\mathbf{c}_j}(n,f) = v_j(n,f)\mathbf{R}_j(f) \tag{12}$$

$$\mathbf{W}_j(n,f) = \mathbf{\Sigma}_{\mathbf{c}_j}(n,f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f) \tag{13}$$

$$\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f) = \mathbf{W}_j(n,f)\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)\mathbf{W}_j^H(n,f) + (\mathbf{I}_I - \mathbf{W}_j(n,f))\mathbf{\Sigma}_{\mathbf{c}_j}(n,f) \tag{14}$$

M step:

$$v_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)) \tag{15}$$

$$\text{Update } \mathbf{R}_j(f). \tag{16}$$

3 Source image-based EM algorithms

3.1 General EM algorithm

Assuming that the spatial covariance matrices are full-rank, ML estimation can be achieved using the source image-based EM (SIEM) algorithm in [26] where the spatial images $\{\mathbf{c}_j(n,f)\}_{n,f}$ of all sources in all time-frequency bins are considered as *hidden data*. Strictly speaking, this algorithm is a generalized form of EM [31] because the M-step increases but does not maximize the expectation of the log-likelihood of the hidden data. Since the priors proposed hereafter pertain to the spatial covariance matrices only, MAP estimation can be achieved via the same algorithm except for the corresponding update in the M-step.

The resulting EM updates are listed in Algorithm 1. In the E-step, the Wiener filter $\mathbf{W}_j(n,f)$ and the second order raw moment $\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)$ of the spatial images of all sources are computed. In the M-step $v_j(n,f)$ and $\mathbf{R}_j(f)$ are updated. In the ML case, the update for $\mathbf{R}_j(f)$ in (16) is given by [26]

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)}{v_j(n,f)} \tag{17}$$

where N is the total number of time frames.

Given this algorithm, we now consider the design of suitable priors over $\mathbf{R}_j(f)$. In addition to the physical constraint (7), the priors must satisfy practical engineering constraints: they must be defined over the space of Hermitian positive definite matrices, have a small number of parameters, have a closed-form mean and result in closed-form EM updates. The inverse-Wishart and the Wishart distributions satisfy these constraints.

3.2 MAP estimation using an inverse-Wishart prior

3.2.1 Inverse-Wishart prior

The inverse-Wishart distribution is the *conjugate prior* for the likelihood (4) of our model. This prior is defined as

$$\mathbf{R}_j(f) \sim \mathcal{IW}(\mathbf{\Psi}_j(f), m) \tag{18}$$

where

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

is the inverse-Wishart density over Hermitian positive definite matrices $\mathbf{R}$ with positive definite inverse scale matrix $\mathbf{\Psi}$, m degrees of freedom and mean $\mathbf{\Psi}/(m-I)$ [32], with Γ the gamma function. This density, its mean, and its variance are finite for $m > I-1$, $m > I$, and $m > I+1$ respectively. We fix the inverse scale matrix $\mathbf{\Psi}_j(f)$ as

$$\mathbf{\Psi}_j(f) = (m-I)\boldsymbol{\mu}_{\mathbf{R}_j}(f) \tag{20}$$

so that the mean of $\mathbf{R}_j(f)$ is consistent with (7). The deviation allowed from the mean is controlled by the so-called number of degrees of freedom m, which is not necessarily an integer.

3.2.2 Learning the hyper-parameter

In order to obtain the best fit between this prior and the actual prior distribution of spatial covariance matrices, we learn the number of degrees of freedom m from training data. Given the relative positions of the sources and the microphones, we generate training signals $\mathbf{c}_p(t)$ for a number of absolute positions p in the room by convolving the corresponding room impulse responses with a single-channel signal. We derive the spatial covariance matrix $\mathbf{R}_p(f)$ associated with each position in an *oracle* fashion [30] by alternately applying (15) and (17) to the empirical covariance matrices $\widehat{\mathbf{R}}_{\mathbf{c}_p}(n, f)$ computed as in (5). Such training data can be generated in any practical scenario where the source separation system is to be deployed in fixed, known environment, where the impulse responses can be pre-recorded or simulated via the *image method* [33].

Since $\mathbf{R}_p(f)$ is measured only up to an arbitrary nonnegative scaling factor $\alpha_p(f)$, we jointly estimate the number of degrees of freedom m and the scaling factors in the ML sense by maximizing

$$\begin{aligned}\mathcal{L}_{\mathcal{IW}} &= \prod_{p,f} p(\mathbf{R}_p(f)|\alpha_p(f), \mathbf{\Psi}(f), m) \\ &= \prod_{p,f} J_{\alpha_p(f)} \mathcal{IW}(\alpha_p(f)\mathbf{R}_p(f)|\mathbf{\Psi}(f), m)\end{aligned} \tag{21}$$

where $J_{\alpha_p(f)} = \alpha_p^{I^2}(f)$ is the Jacobian of the scaling transform and $\mathbf{\Psi}(f)$ is the inverse scale matrix in (20) which is fixed for all p. Maximization with respect to m can be achieved using a nonlinear optimization technique [34], where the optimal scaling factors for a given m are given by

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

The values of m learned for the geometrical setting and the reverberation times tested in Section 5 are shown in Table 1.

3.2.3 MAP EM 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 step (16) of Algorithm 1 by maximizing the expectation of the log-posterior of the hidden data

$$Q_{\mathcal{IW}} = \sum_{j,f} \gamma \log \mathcal{IW}(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m) + \sum_{j,n,f} -\operatorname{tr}(\mathbf{\Sigma}_{\mathbf{c}_j}^{-1}(n,f)\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)) - \log|\pi\mathbf{\Sigma}_{\mathbf{c}_j}(n,f)| \quad (23)$$

where γ is a tradeoff hyper-parameter determining the strength of the prior. By computing the partial derivatives of $Q_{\mathcal{IW}}$ with respect to each entry of $\mathbf{R}_j(f)$ and equating them to zero, we obtain the MAP update

$$\mathbf{R}_j(f) = \frac{1}{\gamma(m+I)+N}\left(\gamma\mathbf{\Psi}_j(f) + \sum_{n=1}^{N}\frac{\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)}{v_j(n,f)}\right). \quad (24)$$

When $\gamma = 0$, the contribution of the prior is excluded and (24) becomes equal to the ML update in (17). The setting of $\gamma = 0$ will be discussed in Section 5.3.

3.3 MAP estimation using a Wishart prior

3.3.1 Wishart prior

As an alternative to the above inverse-Wishart prior, we consider a Wishart prior defined as

$$\mathbf{R}_j(f) \sim \mathcal{W}(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m) \quad (25)$$

where

$$\mathcal{W}(\mathbf{R}|\mathbf{\Psi}, m) = \frac{|\mathbf{\Psi}|^{-m}|\mathbf{R}|^{(m-I)}e^{-\operatorname{tr}(\mathbf{\Psi}^{-1}\mathbf{R})}}{\pi^{I(I-1)/2}\prod_{i=1}^{I}\Gamma(m-i+1)} \quad (26)$$

is the Wishart density over Hermitian positive definite matrices $\mathbf{R}$ with positive definite scale matrix $\mathbf{\Psi}$, m degrees of freedom and mean $m\mathbf{\Psi}$ [32]. This density, its mean, and its variance are finite for $m > I-1$, $m > I$, and $m > I+1$ respectively. We fix the scale matrix $\mathbf{\Psi}_j(f)$ as

$$\mathbf{\Psi}_j(f) = \frac{1}{m}\boldsymbol{\mu}_{\mathbf{R}_j}(f) \quad (27)$$

so that the mean of $\mathbf{R}_j(f)$ is consistent with (7).

3.3.2 Learning the hyper-parameter

Similarly to above, the number of degrees of freedom m is learned from training data in the ML sense by maximizing

$$\mathcal{L}_{\mathcal{W}} = \prod_{p,f} J_{\alpha_p(f)}\mathcal{W}(\alpha_p(f)\mathbf{R}_p(f)|\mathbf{\Psi}(f), m). \quad (28)$$

using a nonlinear optimization technique, where the optimal scaling factors for a given m are given by

$$\alpha_p(f) = \frac{Im}{\operatorname{tr}(\mathbf{\Psi}_p^{-1}(f)\mathbf{R}_p(f))}. \quad (29)$$

The maximization of (21) and (28) generally yields different values of m, except in the two-channel case where it can be shown that the optimal value of m is equal for the two priors. As a result, the learned values of m listed in Table 1 are also valid for the Wishart prior.

3.3.3 MAP EM update

Under this prior, the expectation of the log-posterior of the hidden data can be expressed as

$$Q_{\mathcal{W}} = \sum_{j,f} \gamma \log \mathcal{W}(\mathbf{R}_j(f)|\mathbf{\Psi}_j(f), m) + \sum_{j,n,f} -\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{c}_j}^{-1}(n,f)\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)) - \log|\pi\mathbf{\Sigma}_{\mathbf{c}_j}(n,f)| \quad (30)$$

The maximization of this quantity with respect to $\mathbf{R}_j(f)$ yields the following MAP update derived in the Appendix for the step (16) of Algorithm 1:

$$\mathbf{R}_j(f) = \frac{1}{2}\mathbf{A}^{-1/2}[-\mathbf{B} + (\mathbf{B}^2 - 4\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2})^{1/2}]\mathbf{A}^{-1/2} \quad (31)$$

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

$$\begin{aligned} \mathbf{A} &= \gamma\mathbf{\Psi}_j^{-1}(f) \\ \mathbf{B} &= [-\gamma(m-I) + N]\mathbf{I}_I \\ \mathbf{C} &= \sum_{n=1}^{N} -\frac{\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)}{v_j(n,f)}. \end{aligned} \quad (32)$$

4 Subsource-based EM algorithm

4.1 General EM algorithm

Besides the SIEM algorithm, an alternative subsource-based EM (SSEM) algorithm was proposed for ML estimation in [12] that applies to spatial covariance matrices of any rank R_j. This algorithm relies on the non-unique representation of the source spatial images as $\mathbf{c}_j(n,f) = \mathbf{H}_j(f)\mathbf{s}_j(n,f)$ where the entries $s_{jr}(n,f)$, $r = 1, ..., R_j$, of $\mathbf{s}_j(n,f)$ are uncorrelated complex-valued subsource coefficients distributed as $s_{jr}(n,f) \sim \mathcal{N}(0, v_j(n,f))$ and $\mathbf{H}_j(f)$ is an $I \times R_j$ complex-valued subsource mixing matrix satisfying the constraint [12]

$$\mathbf{R}_j(f) = \mathbf{H}_j(f)\mathbf{H}_j^H(f). \quad (33)$$

This subsource mixing matrix reduces to a mixing vector in the particular case when $\mathbf{R}_j(f)$ has rank 1. Overall the mixture STFT coefficients are written as

$$\mathbf{x}(n,f) = \mathbf{H}(f)\mathbf{s}(n,f) + \mathbf{b}(n,f) \quad (34)$$

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 subsource coefficients with $R = \sum_{j=1}^{J} R_j$, $\mathbf{H}(f) = [\mathbf{H}_1(f), ..., \mathbf{H}_J(f)]$ is an $I \times R$ mixing matrix and $\mathbf{b}(n,f)$ is a small Gaussian noise with covariance matrix

Algorithm 2 SSEM algorithm [12]

E step:

$$\boldsymbol{\Sigma}_{\mathbf{s}}(n,f) = \mathrm{diag}\left([\tilde{v}_r(n,f)]_{r=1}^{R}\right) \tag{35}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}}(n,f) = \mathbf{H}(f)\boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{H}^H(f) + \boldsymbol{\Sigma}_{\mathbf{b}}(n,f) \tag{36}$$

$$\mathbf{W}(n,f) = \boldsymbol{\Sigma}_{\mathbf{s}}(n,f)\mathbf{H}^H(f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f) \tag{37}$$

$$\widehat{\mathbf{R}}_{\mathbf{s}}(n,f) = \mathbf{W}(n,f)\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)\mathbf{W}^H(n,f) + (\mathbf{I}_R - \mathbf{W}(n,f)\mathbf{H}(f))\boldsymbol{\Sigma}_{\mathbf{s}}(n,f) \tag{38}$$

$$\widehat{\mathbf{R}}_{\mathbf{xs}}(n,f) = \widehat{\mathbf{R}}_{\mathbf{x}}(n,f)\mathbf{W}^H(n,f) \tag{39}$$

M step:

$$v_j(n,f) = \frac{1}{R_j}\sum_{r\in\mathcal{R}_j}[\widehat{\mathbf{R}}_{\mathbf{s}}(n,f)]_{rr} \tag{40}$$

$$\text{Update } \mathbf{H}(f). \tag{41}$$

$\boldsymbol{\Sigma}_{\mathbf{b}}(n,f) = \sigma_b^2(f)\mathbf{I}_I$ required by the EM algorithm. The log-likelihood (4) can then be maximized by considering the set $\{\mathbf{x}(n,f), s_j(n,f)\}_{j,n}$ of observed mixture STFT coefficients and hidden subsource STFT coefficients in all time-frequency bins as *complete data*. Once again, it turns out that MAP estimation can be achieved via the same algorithm except for the mixing matrix update in the M-step.

The details of one iteration are summarized in Algorithm 2, where $\mathcal{R}_j$ denotes the set of subsource indices associated with the j-th source and $\tilde{v}_r(n,f) = v_j(n,f)$ if and only if $r \in \mathcal{R}_j$. In the E-step, the Wiener filter $\mathbf{W}_j(n,f)$ and the second order cross-moments $\widehat{\mathbf{R}}_{\mathbf{s}}(n,f)$ and $\widehat{\mathbf{R}}_{\mathbf{xs}}(n,f)$ are computed. In the M-step $v_j(n,f)$ and $\mathbf{H}(f)$ are updated. In the ML case, the update for $\mathbf{H}(f)$ in (41) is given by [12]

$$\mathbf{H}(f) = \left(\sum_{n=1}^{N}\widehat{\mathbf{R}}_{\mathbf{xs}}(n,f)\right)\left(\sum_{n=1}^{N}\widehat{\mathbf{R}}_{\mathbf{s}}(n,f)\right)^{-1}. \tag{42}$$

4.2 Gaussian prior

The design of a suitable prior over $\mathbf{H}(f)$ is subject to the same practical engineering constraints as above, which lead us to propose a Gaussian prior. We model each column $\mathbf{h}_{jr}(f)$, $r = 1, ..., R_j$, of $\mathbf{H}_j(f)$ as a complex-valued Gaussian random vector

$$\mathbf{h}_{jr}(f) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f), \boldsymbol{\Sigma}_{\mathbf{h}_{jr}}(f)) \tag{43}$$

with mean $\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f)$ and covariance $\boldsymbol{\Sigma}_{\mathbf{h}_{jr}}(f)$. Following the assumption in Section 2.2, echoes and reverberation cancel out on average over all positions in the room so that they appear only in the covariance, while only the part corresponding to direct sound appears in the mean. Without loss of generality, let us select $\mathbf{H}_j(f)$ such that direct sound is concentrated in the first subsource of each source, i.e., the first subsource includes direct sound, echoes and reverberation while the other subsources include

echoes and reverberation only[2]. The mean and the covariance of the prior can then be expressed as

$$\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f) = \begin{cases} \mathbf{d}_j(f) & \text{if } r = 1 \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{44}$$

$$\boldsymbol{\Sigma}_{\mathbf{h}_{jr}}(f) = \sigma_r^2 \boldsymbol{\Omega}(f) \tag{45}$$

where the echo and reverberation power of all subsources sum up to the total power in (10):

$$\sum_{r=1}^{R_j} \sigma_r^2 = \sigma_{\text{rev}}^2. \tag{46}$$

Contrary to the inverse-Wishart and Wishart priors whose variance is governed by a single hyper-parameter m, this prior involves $R_j - 1$ free hyper-parameters σ_r^2, $r = 2, ..., R_j$, which makes it potentially more flexible as soon as $I \geq R_j > 2$. The priors are distinct, however, in the sense that the Gaussian prior does not generalize the other two priors whatever the choice of the hyper-parameters.

4.3 Learning the hyper-parameters

In order to fit the actual distribution of subsource mixing matrices, we learn these free hyper-parameters from training data. The training data consist of the spatial covariance matrices $\mathbf{R}_p(f)$ computed in Section 3.2.2 for different positions p, from which we derive the corresponding subsource mixing matrices $\mathbf{H}_p(f)$ by singular value decomposition $\mathbf{R}_p(f) = \mathbf{H}_p(f)\mathbf{H}_p^H(f)$ such that the columns of $\mathbf{H}_p(f)$ are orthogonal and sorted by decreasing norm.

These columns $\mathbf{h}_{pr}(f)$ are observed only up to an arbitrary scale common to all r and an arbitrary phase rotation specific to each r. Phase rotations do not affect the learned variances σ_r^2 for $r > 1$, since the corresponding means $\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f)$ are zero. Multiplying $\mathbf{H}_p(f)$ by a global complex-valued factor $\alpha_p(f)$ is hence sufficient to address this indeterminacy. Denoting by

$$\underline{\mathbf{h}}_p(f) = \begin{pmatrix} \mathbf{h}_{p1}(f) \\ \vdots \\ \mathbf{h}_{pR_j}(f) \end{pmatrix} \tag{47}$$

the $IR_j \times 1$ vectorization of $\mathbf{H}_p(f)$ with mean

$$\boldsymbol{\mu}_{\underline{\mathbf{h}}_p}(f) = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{h}_{p1}}(f) \\ \vdots \\ \boldsymbol{\mu}_{\mathbf{h}_{p2}}(f) \end{pmatrix} \tag{48}$$

and covariance

$$\boldsymbol{\Sigma}_{\underline{\mathbf{h}}_p}(f) = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{h}_{p1}}(f) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{\Sigma}_{\mathbf{h}_{p2}}(f) \end{pmatrix}, \tag{49}$$

[2]If several $\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f)$ are nonzero multiples of $\mathbf{d}_j(f)$, a unitary transform can be applied to $\mathbf{H}_j(f)$ in (33) such that only the first one remains nonzero.

the hyper-parameters and the multiplication factors are jointly estimated in the ML sense by maximizing

$$\begin{aligned}\mathcal{L}_G &= \prod_{p,f} p(\underline{\mathbf{h}}_p(f)|\alpha_p(f), \boldsymbol{\mu}_{\underline{\mathbf{h}}_p}(f), \boldsymbol{\Sigma}_{\underline{\mathbf{h}}_p}(f)) \\ &= \prod_{p,f} J_{\alpha_p(f)} \mathcal{N}(\alpha_p(f)\underline{\mathbf{h}}_p(f)|\boldsymbol{\mu}_{\underline{\mathbf{h}}_p}(f), \boldsymbol{\Sigma}_{\underline{\mathbf{h}}_p}(f)) \end{aligned} \tag{50}$$

where $J_{\alpha_p(f)} = |\alpha_p(f)|^{2I^2}$ is the Jacobian of the multiplication. Maximization is achieved using a nonlinear optimization technique, where the optimal multiplication factors as a function of the hyper-parameters are found as

$$\alpha_p(f) = \frac{-|b| - (|b|^2 - 4ac)^{1/2}}{2a} \frac{b}{|b|} \tag{51}$$

where

$$\begin{aligned} a &= -\underline{\mathbf{h}}_p^H(f) \boldsymbol{\Sigma}_{\underline{\mathbf{h}}_p}^{-1}(f) \underline{\mathbf{h}}_p(f) \\ b &= \underline{\mathbf{h}}_p^H(f) \boldsymbol{\Sigma}_{\underline{\mathbf{h}}_p}^{-1}(f) \boldsymbol{\mu}_{\underline{\mathbf{h}}_p}(f) \\ c &= I^2 \end{aligned} \tag{52}$$

The values of σ_1^2 and σ_2^2 learned in the setting of Section 5 ($R_j = I = 2$) are displayed in Table 1.

4.4 MAP EM update

Similarly to (47)–(49), let us denote by $\underline{\mathbf{h}}(f)$ the vectorization of $\mathbf{H}(f)$ as an $IR \times 1$ column vector. The prior distribution (43) translates into

$$\underline{\mathbf{h}}(f) \sim \mathcal{N}(\boldsymbol{\mu}_{\underline{\mathbf{h}}}(f), \boldsymbol{\Sigma}_{\underline{\mathbf{h}}}(f)). \tag{53}$$

where $\boldsymbol{\mu}_{\underline{\mathbf{h}}}(f)$ is the $IR \times 1$ vector obtained by concatenating $\boldsymbol{\mu}_{\mathbf{h}_{jr}}(f)$ for all j, r, and $\boldsymbol{\Sigma}_{\underline{\mathbf{h}}}(f)$ is the $IR \times IR$ block-diagonal matrix whose entries are equal to $\boldsymbol{\Sigma}_{\mathbf{h}_{jr}}(f)$ for all j, r.

The MAP update for $\mathbf{H}(f)$ is derived by maximizing the expectation of the log-posterior of the complete data that is equal up to a constant to (see [12, eq.18] for the expression of the expectation of the log-likelihood)

$$\begin{aligned} Q_{\mathcal{G}} = \gamma \log \mathcal{N}(\underline{\mathbf{h}}(f)|\boldsymbol{\mu}_{\underline{\mathbf{h}}}(f), \boldsymbol{\Sigma}_{\underline{\mathbf{h}}}(f)) \\ + \sum_{n,f} -\frac{1}{\sigma_b^2(f)} \operatorname{tr}[\widehat{\mathbf{R}}_{\mathbf{x}}(n,f) - \mathbf{H}(f)\widehat{\mathbf{R}}_{\mathbf{xs}}^H(n,f) \\ - \widehat{\mathbf{R}}_{\mathbf{xs}}(n,f)\mathbf{H}^H(f) + \mathbf{H}(f)\widehat{\mathbf{R}}_{\mathbf{s}}(n,f)\mathbf{H}^H(f)] \end{aligned} \tag{54}$$

where γ is a tradeoff hyper-parameter determining the strength of the prior. By rewriting the matrix quadratic form in the log-likelihood term of (54) as a vector quadratic form in terms of $\underline{\mathbf{h}}(f)$ and by computing the gradient of $Q_{\mathcal{G}}$ and equating it to zero, we

obtain

$$\underline{\mathbf{h}}(f) = \left(\gamma \boldsymbol{\Sigma}_{\underline{\mathbf{h}}}^{-1}(f) + \frac{1}{\sigma_b^2(f)} \sum_{n=1}^{N} (\widehat{\mathbf{R}}_{\mathbf{s}}(n,f) \otimes \mathbf{I}_I)^T\right)^{-1} \left(\gamma \boldsymbol{\Sigma}_{\underline{\mathbf{h}}}^{-1}(f) \boldsymbol{\mu}_{\underline{\mathbf{h}}}(f) + \frac{1}{\sigma_b^2(f)} \sum_{n=1}^{N} \mathrm{vec}(\widehat{\mathbf{R}}_{\mathbf{xs}}(n,f))\right) \quad (55)$$

where $.^T$ denotes transposition, $\otimes$ is the Kronecker product and $\mathrm{vec}(.)$ concatenates the columns of a matrix into a single column vector. The mixing matrix $\mathbf{H}(f)$ is then obtained by devectorizing $\underline{\mathbf{h}}(f)$. This update boils down to the ML update (42) when $\gamma = 0$.

5 Experimental evaluation

We assess the performance of the proposed MAP estimation algorithms compared to the conventional ML estimation algorithms for the separation of two-channel convolutive mixtures of three sources. We target a semi-informed scenario where the relative positions of the sources and the microphones are known, but nothing is known about their absolute position in the room nor about the source signals. The reverberant character of the data calls for the use of full-rank spatial covariance matrices and subsource mixing matrices, i.e., $R_j = 2$ for all j. We do not constrain the source variances $v_j(n,f)$, so as to measure the improvement due to the priors alone. The full Matlab code for our experiments can be downloaded from [35].

5.1 Data

The proposed priors can be applied in any scenario where the source separation system is to be deployed in fixed, known environment, where the impulse responses can be pre-recorded or simulated. In the following, we use simulated mixtures so as to test a wide range of room reverberation times. The use of simulated data is widespread in audio source separation and it has been shown to yield a separation performance similar to real-world data in general [36].

The positions of the sources and the microphones in the test data are illustrated in Fig. 1. The room dimensions are $4.45 \times 3.55 \times 2.5$ m as in [36] and the microphone spacing and the source-to-microphone distances are fixed to $d = 5$ cm and $r = 50$ cm, respectively. We generated room impulse responses via the image method [33] using the Roomsimove toolbox[3] for four reverberation times: $T_{60} = 50$, 130, 250 or 500 ms, which we convolved with 10 s speech signals sampled at 16 kHz. Two sets of speech signals were considered: male and female speech, resulting in 2 two-channel mixture signals for each T_{60} and 8 mixture signals in total.

Training data were generated in a similar fashion by simulating room impulse responses for 20 random source directions of arrival for each of 20 random microphone pair positions and orientations for the same d and r as above. This resulted in a total of 400 source image signals indexed by p for each T_{60}.

[3] http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip
This toolbox provides a command-line interface which, in contrast with the original GUI by D. R. Campbell, allows generation of a large amount of data.

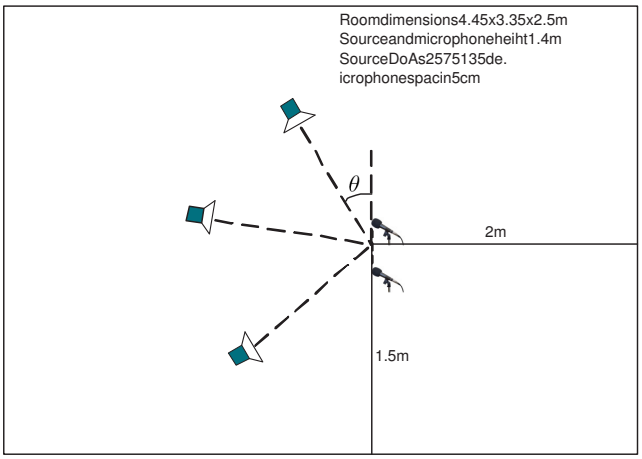


Figure 1: Room geometric setting for testing data.

5.2 Learned hyper-parameter values

Regarding training, preliminary experiments showed that the functions (21), (28) and (50) are concave in practice. Hence, we maximized them using Matlab's fmincon optimizer. The resulting hyper-parameter values are shown in Table 1.

As expected, the total power of echoes and reverberation $\sigma^2_{\text{rev}} = \sigma^2_1 + \sigma^2_2$ strongly increases with T_{60}, such that the direct-to-reverberant 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 [32], decreases with T_{60} while that of the Wishart prior, on the contrary, increases with T_{60}. The ratio $\sigma^2_1/\sigma^2_{\text{rev}}$ decreases with T_{60}, which indicates that the echoic and reverberant part of the impulse responses becomes more and more diffuse.

Table 1: Learned values of the prior hyper-parameters.

T_{60}	50 ms	130 ms	250 ms	500 ms
m	2.1	2.1	3.4	5.3
σ^2_1	0.009	0.033	0.068	0.148
σ^2_2	0.002	0.024	0.063	0.139

5.3 Tested algorithms and evaluation criteria

In addition to the proposed MAP versions of SIEM and SSEM (*MAP inverse-Wishart*, *MAP Wishart*, *MAP Gaussian*), we consider the conventional ML versions of these algorithms where the initial values of $\mathbf{R}_j(f)$ and $\mathbf{H}(f)$ are either set to $\boldsymbol{\mu}_{\mathbf{R}_j}(f)$ and $\boldsymbol{\mu}_{\mathbf{H}}(f)$ given the scene geometry (*ML geom init*) or blindly estimated via hierarchical clustering as in [11] (*ML blind init*).

We computed the STFT with half-overlapping sine windows of length 1024 and the empirical mixture covariance using a window w_{nf} of size 3×3 as in [26]. The trade-off parameter γ does not significantly affect the results but we observed that $\gamma = 100$ and $\gamma = 10$ are good choices for SIEM and SSEM respectively on average. The number of

iterations was fixed to 10 for SIEM and 30 for SSEM, since the convergence of SSEM is typically slower.

The priors did not significantly increase running time. Indeed, the MAP inverse-Wishart update has the same computational complexity as the ML SIEM update. The MAP Wishart update and the MAP Gaussian updates have greater complexity than the ML SIEM and SSEM updates, respectively, but they occur only once per iteration in each frequency bin, in contrast with the updates in the E-step which occur in each time frame. For a typical number of time frames N, the computational complexity is therefore dominated by the E-step, regardless of the priors.

We evaluated the separation quality via 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 in decibels (dB) [36], which account respectively for overall distortion, residual crosstalk, musical noise and target distortion. These criteria were computed using version 3.0 of the BSS Eval toolbox[4] and averaged over all sources and all mixtures for each T_{60}.

5.4 Results for source image-based EM algorithms

The results of the SIEM algorithms are compared in Fig. 2. *ML geom init* results in better performance than *ML blind init* in terms of all criteria for all T_{60}. *MAP Wishart* offers the best SAR and very similar SDR to *ML geom init*. Overall, *MAP inverse-Wishart* outperforms all other algorithms for all considered T_{60} 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 *ML blind init*, *ML geom init* and *MAP Wishart*, respectively. This confirms the benefit of the proposed inverse-Wishart spatial location prior and the associated MAP algorithm.

5.5 Results for subsource-based EM algorithms

The results of the SSEM algorithms are depicted in Fig. 3. Again, *ML geom init* results in significantly better performance than *ML blind init* in terms of all criteria for all T_{60}. But the best performance is achieved by *MAP Gaussian* in terms of all criteria and for all T_{60}, except in terms of ISR at low T_{60}. For instance, at $T_{60} = 250$ ms, *MAP Gaussian* improves the SDR by 4.2 dB and 0.3 dB compared to *ML blind init* and *ML geom init*, respectively. This confirms the benefit of the proposed Gaussian spatial location prior and the associated MAP algorithm.

6 Conclusion

We considered two classes of source separation algorithms grounded on the emerging Gaussian EM framework. In contrast with classical ML estimation of the spatial parameters, we proposed three priors exploiting a result from the theory of statistical room acoustics and we derived closed-form MAP updates. Theoretically speaking, these updates provide a statistically principled solution to the problem of permutation of the source estimates and they help reducing overfitting of the parameter values. In practice, the SIEM algorithm with an inverse-Wishart prior and the SSEM algorithm with a Gaussian prior were shown to outperform their ML counterparts for all room reverberation times in a semi-informed scenario.

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

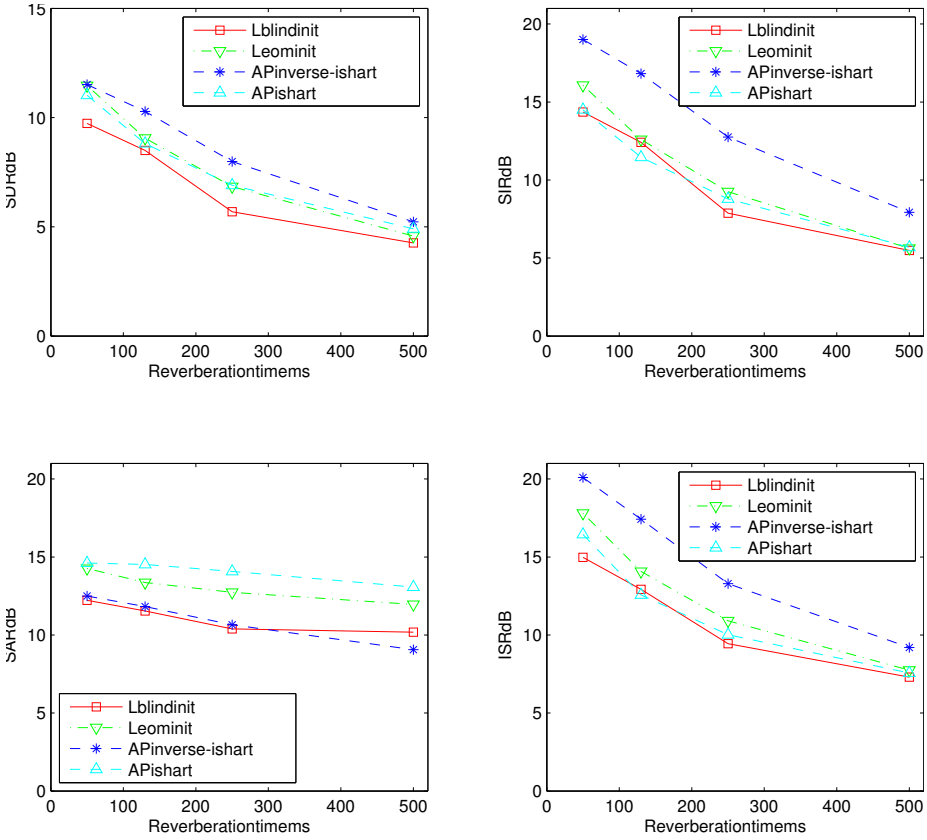


Figure 2: Separation performance of the SIEM algorithms as a function of reverberation time.

The results in this paper can readily be used in certain real-world scenarios where the source positions are known from, e.g., physical constraints or visual input, and the reverberation characteristics can be learned from the environment [21, 22, 23]. Perhaps more importantly, they constitute a first step towards full Bayesian treatment of this family of models in other blind or semi-blind scenarios in the future. In addition to blind estimation of the source positions and possibly of the microphone distance and directivity [37], robustness to erroneous estimation of these hyper-parameters and blind estimation of the hyper-parameters σ^2_{rev}, m and σ^2_r both pose significant challenges, which go beyond the scope of this paper. Future work will concentrate on these challenges by extending blind techniques for room reverberation time estimation [38]. Usage of the proposed Gaussian prior, which is also valid for rank-1 mixing vectors, may also be explored in the context of FDICA, with the difficulty of translating this prior into a prior over the blocking vectors which are usually considered as parameters in this context instead.

Acknowledgment

This work was supported by the EUREKA Eurostars i3Dmusic project funded by Oseo.

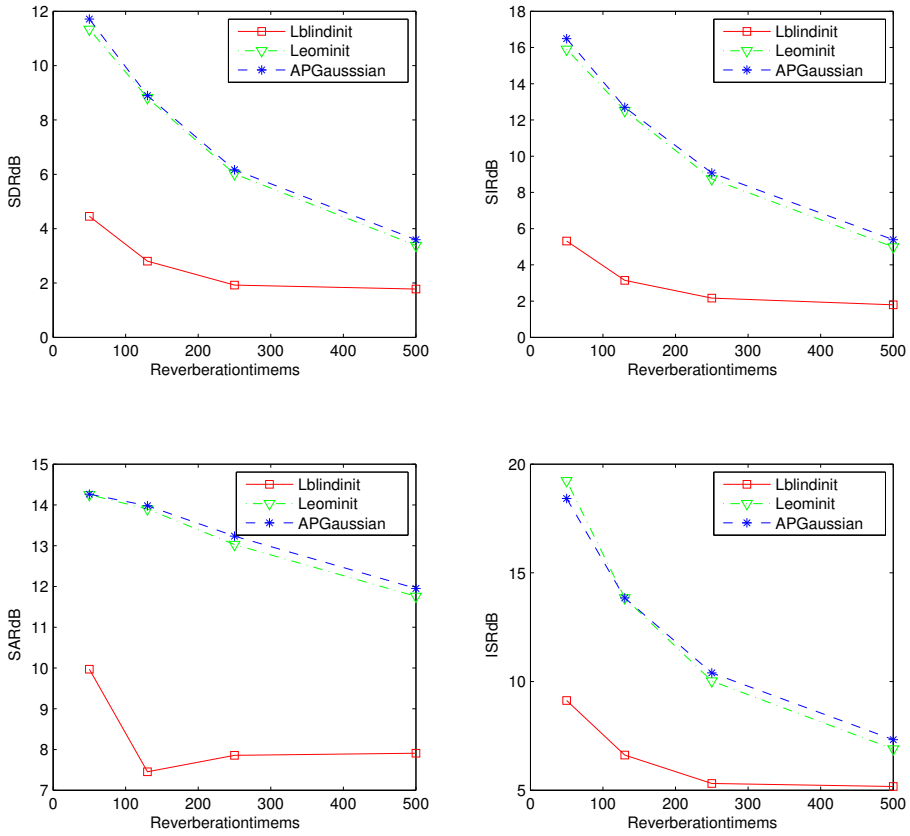


Figure 3: Separation performance of the SSEM algorithms as a function of reverberation time.

A Derivation of the MAP EM update for the Wishart prior

The MAP SIEM update under a Wishart prior is derived as follows. By computing the partial derivatives of $Q_{\mathcal{W}}$ with respect to each entry of $\mathbf{R}_j(f)$ and equating them to zero, we obtain the quadratic matrix equation

$$\mathbf{R}_j(f)\mathbf{A}\mathbf{R}_j(f) + \mathbf{B}\mathbf{R}_j(f) + \mathbf{C} = \mathbf{0} \tag{56}$$

where the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are defined in (32). After the variable change $\mathbf{X} = \mathbf{A}^{1/2}\mathbf{R}_j(f)\mathbf{A}^{1/2}$ when $\gamma \neq 0$, this can be rewritten as

$$\mathbf{X}^2 + \mathbf{B}\mathbf{X} + \mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2} = \mathbf{0}. \tag{57}$$

The first two coefficients of this equation are scalar multiples of the identity matrix and the third one is Hermitian positive definite. Therefore, it has a unique Hermitian positive definite solution given by [39, p.304]

$$\mathbf{X} = \frac{1}{2}[-\mathbf{B} + (\mathbf{B}^2 - 4\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2})^{1/2}]. \tag{58}$$

$\mathbf{R}_j(f)$ is then obtained as $\mathbf{R}_j(f) = \mathbf{A}^{-1/2}\mathbf{X}\mathbf{A}^{-1/2}$.

References

[1] 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.

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

[3] 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.

[4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[5] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation*. Springer, 2007, pp. 47–78.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[14] 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, 2009.

[15] T. Virtanen, A. T. Cemgil, and S. J. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 1825–1828.

[16] O. Dikmen and A. T. Cemgil, "Gamma Markov random fields for audio source modeling," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 589–601, 2010.

[17] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Simultaneous processing of sound source separation and musical instrument identification using Bayesian spectral modeling," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 3816–3819.

[18] 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.

[19] K. H. Knuth, "A Bayesian approach to source separation," in *Proc. Int. Conf. on Independent Component Analysis and Source Separation (ICA)*, 1999, pp. 283–288.

[20] A. T. Cemgil, C. Févotte, and S. J. Godsill, "Variational and stochastic inference for Bayesian source separation," *Digital Signal Processing*, vol. 17, pp. 891–913, 2007.

[21] 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.

[22] 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.

[23] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 726–745, 2013.

[24] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian unification of sound source localization and separation with permutation resolution," in *Proc. 26th AAAI Conf. on Artificial Intelligence*, 2012, pp. 2038–2045.

[25] N. Q. K. Duong, E. Vincent, and R. Gribonval, "An acoustically-motivated spatial prior for under-determined reverberant source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2011, pp. 9–12.

[26] ——, "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.

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

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

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] J. Nocedal and S. J. Wright, *Numerical optimization*. New York, NY: Springer, 1999.

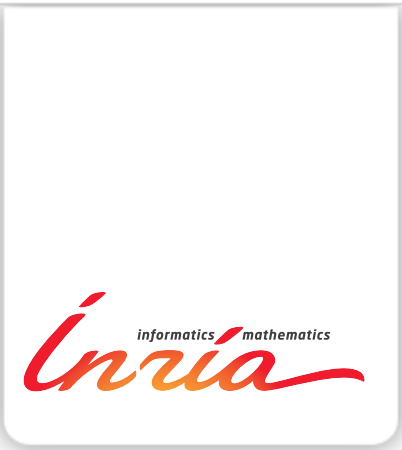
[35] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Matlab code for gaussian model based audio source separation using spatial location priors." [Online]. Available: http://www.loria.fr/~evincent/spatial_priors.zip

[36] 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.

[37] K. Hasegawa, N. Ono, S. Miyabe, and S. Sagayama, "Blind estimation of locations and time offsets for distributed recording devices," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 57–64.

[38] N. D. Gaubitch, H. Löllmann, M. Jeub, T. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[39] N. J. Higham and H. M. Kim, "Solving a quadratic matrix equation by Newton's method with exact line searches," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, pp. 303–316, 2001.



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