Variational Bayesian Inference for Source Separation and Robust Feature Extraction

Kamil Adiloğlu, Emmanuel Vincent

Variational Bayesian Inference for Source Separation and Robust Feature Extraction

Kamil Adiloğlu, Emmanuel Vincent

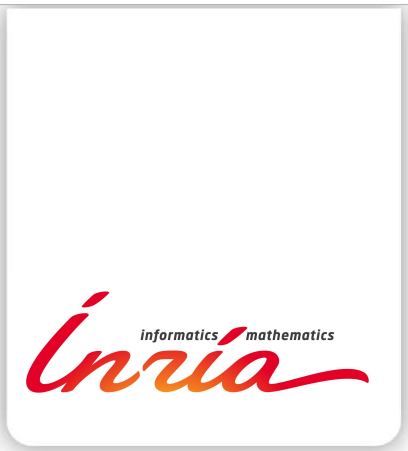
TECHNICAL REPORT

N° 428

August 2012

Project-Team Metiss

ISSN 0249-0803 ISRN INRIA/RT--428--FR+ENG



Variational Bayesian Inference for Source Separation and Robust Feature Extraction

Kamil Adiloğlu, Emmanuel Vincent

Project-Team Metiss

Technical Report n° 428 — August 2012 — 18 pages

Abstract: In this paper, we consider the feature extraction problem from each source individually in a multisource audio recording using a general audio source separation algorithm. The main challenge to tackle with is to estimate the uncertainty of the sources and to propagate it to the features, so as to robustly estimate them despite source separation errors. The state-of-the-art methods estimate the uncertainty in a heuristic manner, whereas we propose to integrate over the parameters of the source separation algorithm. For this purpose, we adapt a variational Bayes method to estimate the posterior probabilities of individual sources and subsequently compute the expectations of the features by propagating the uncertainties using the moment matching method. We evaluated the accuracy of the features in terms of the root mean square error as well as conducted speaker recognition experiments to observe the performance of the features in a real world problem. In both cases, the proposed method yielded the best results.

Key-words: Audio source separation, local Gaussian modeling, non-negative matrix factorization, variational Bayesian inference

This work is a part of the QUAERO project funded by OSEO.

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Inférence variationnelle bayésienne pour la séparation de sources et l'extraction robuste de descripteurs

Résumé : Dans cet article, nous considérons le problème de l'extraction des descripteurs de chaque source dans un enregistrement audio multi-sources à l'aide d'un algorithme général de séparation de sources. La difficulté consiste à estimer l'incertitude sur les sources et à la propager aux descripteurs, afin de les estimer de façon robuste en dépit des erreurs de séparation. Les méthodes de l'état de l'art estiment l'incertitude de façon heuristique, tandis que nous proposons d'intégrer sur les paramètres de l'algorithme de séparation de sources. Nous décrivons dans ce but une méthode d'inférence variationnelle bayésienne pour l'estimation de la distribution a posteriori des sources et nous calculons ensuite l'espérance des descripteurs par propagation de l'incertitude selon la méthode d'identification des moments. Nous évaluons la précision des descripteurs en terme d'erreur quadratique moyenne et conduisons des expériences de reconnaissance du locuteur afin d'observer la performance qui en découle pour un problème réel. Dans les deux cas, la méthode proposée donne les meilleurs résultats.

Mots-clés : Séparation de sources audio, modèle gaussien local, factorisation matricielle positive, inférence variationelle bayésienne

Contents

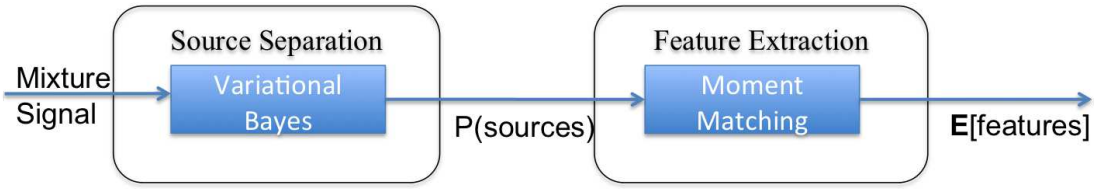


Figure 1: Flow of the proposed Bayesian source separation and feature extraction approach.

1 Introduction

Feature extraction plays an important role in solving many problems in the field of audio information retrieval. Extracting the correct features describing the audio content as well as extracting them properly is essential for the consistency of the application. However, most audio signals consist of a mixture of several sound sources, which have their own characteristics. Applying source separation prior to feature extraction and extracting features from each source individually can increase retrieval accuracy. For increased robustness, the uncertainty over the separated sources must be estimated in the complex-valued time-frequency domain and propagated to the features [13].

A heuristic approach is to assume that the uncertainty is proportional to the squared difference between the separated sources and the mixture [13, 8]. In [4], a more principled approach is taken whereby the separated sources are assumed to follow a Gaussian posterior distribution, whose mean and variance are those of the Wiener filter used for separation. Propagation to the features is then achieved either by moment matching [4] or unscented transform [13]. This approach remains mathematically inaccurate however, since the parameters of the Wiener filter are fixed to a certain value instead of being integrated over in a fully Bayesian approach.

In a preliminary study using a simple local Gaussian source model [9], we proposed a Gibbs sampling algorithm and a variational Bayes (VB) algorithm to address this integration and showed that the latter decreased the RMS error over the resulting Mel frequency cepstral coefficients (MFCC) [1]. In a following paper, we extended this approach to the general modeling framework for source separation recently introduced in [15]. This framework generalizes a wide class of existing source separation algorithms, including nonstationarity-based frequency-domain independent component analysis (FDICA) and single- or multi-channel nonnegative matrix factorization (NMF). We proposed a VB algorithm to estimate the posterior distribution of the source time-frequency coefficients and subsequently derive the expectation of the features by moment matching [2]. However, in this work, in order to obtain closed form update equations, we introduced the so-called source sub-components, which divide each source into several sub-components as many as the number of multilevel NMF parameters. This caused long computation times. The current paper is based upon this work and extends it by using the generalized inverse Gaussian distribution for the multilevel NMF parameters and thereby getting rid of the necessity to artificially introducing the source sub-components. Figure 1 illustrates the workflow of the proposed approach.

This paper is organized as follows. Section 2 introduces the source separation framework. Section 3 presents the proposed VB inference algorithm for the estimation of the posterior distribution of the sources. Section 4 presents the uncertainty propagation method. In Section 5, we evaluate this framework over convolutive mixtures. We conclude in Section 6.

2 General Source Separation Framework

The proposed model operates in the short time Fourier transform (STFT) domain. For J source signals in I channels, the mixing equation is written as

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \boldsymbol{\epsilon}_{fn}. \tag{1}$$

where $\mathbf{x}_{fn} = [x_{1,fn}, \ldots, x_{I,fn}]^T$ represents the $I \times 1$ vector containing the mixture STFT coefficients, $\mathbf{s}_{fn} = [s_{1,fn}, \ldots, s_{J,fn}]^T$ represents the $I \times 1$ vector consisting of the sources, $\mathbf{A}_f = [\mathbf{A}_{1,f}, \ldots, \mathbf{A}_{J,f}]$ represents the $I \times J$ complex valued mixing matrix and $\boldsymbol{\epsilon}_{fn}$ represents the noise. In this formulation, f is the frequency index, n the time frame index, i the channel index and j the source index. Note that this framework also works for diffuse or reverberated sources by modeling each source as a subspace spanned by several point sources [15].

We assume that each source signal $s_{j,fn}$ follows a zero-mean complex-valued Gaussian distribution with variance $v_{j,fn}$ given as

$$s_{j,fn} \sim \mathcal{N}(0, v_{j,fn}). \tag{2}$$

The source variances $v_{j,fn}$, which encode the spectral power are decomposed via an excitation-filter model [15]

$$v_{j,fn} = v^{\text{ex}}_{j,fn} v^{\text{ft}}_{j,fn}. \tag{3}$$

The excitation spectral power $v^{\text{ex}}_{j,fn}$ is decomposed into characteristic spectral patterns modulated by time activation coefficients. Finally, the characteristic spectral patterns are defined as the sum of narrowband spectral patterns $w^{\text{ex}}_{j,fl}$ with associated weights $u^{\text{ex}}_{j,lk}$. Similarly, the time activation coefficients are represented as a sum of time-localized patterns $h^{\text{ex}}_{j,mn}$ with their weights $g^{\text{ex}}_{j,km}$. The same decomposition applies to the filter spectral power $v^{\text{ft}}_{j,fn}$. This framework makes it possible to incorporate a wide range of constraints about the sources. For instance, harmonicity can be enforced by choosing $w^{\text{ex}}_{j,fl}$ as narrowband harmonic spectra and letting the spectral envelope and the active pitches be inferred from the data via the other parameters. For more details about how to constrain spectral and temporal structures, see [15]. As a result, the complete factorization scheme is as follows

$$v^{\text{ex}}_{j,fn} = \sum_{k=1}^{K^{\text{ex}}_j} \sum_{m=1}^{M^{\text{ex}}_j} \sum_{l=1}^{L^{\text{ex}}_j} h^{\text{ex}}_{j,mn} g^{\text{ex}}_{j,km} u^{\text{ex}}_{j,lk} w^{\text{ex}}_{j,fl}, \tag{4}$$

$$v^{\text{ft}}_{j,fn} = \sum_{k'=1}^{K^{\text{ft}}_j} \sum_{m'=1}^{M^{\text{ft}}_j} \sum_{l'=1}^{L^{\text{ft}}_j} h^{\text{ft}}_{j,m'n} g^{\text{ft}}_{j,k'm'} u^{\text{ft}}_{j,l'k'} w^{\text{ft}}_{j,fl'}. \tag{5}$$

In a fully Bayesian treatment, we need to define the prior distributions of the model parameters. Now, we define the multilevel NMF parameters of the source variances to follow the non-informative Jeffreys prior $\mathcal{J}(x) \sim \frac{1}{x}$.

Finally, for the mixing system, we take the dependencies between the channels and between the source components into account. Therefore, we consider the mixing matrix $\mathbf{A}_f$ as a whole and define the prior distribution accordingly. First, we reshape the mixing matrix $\mathbf{A}_f$ into a vector $\underline{\mathbf{A}}_f$. For this, we concatenate the row vectors of $\mathbf{A}_f$ into the column vector $\underline{\mathbf{A}}_f$. Then, we define the prior distribution of $\underline{\mathbf{A}}_f$ to be a flat prior by defining the covariance matrix as $\boldsymbol{\Sigma}_{\underline{\mathbf{A}},f} \to +\infty$ in a Gaussian distribution.

2.1 Data Likelihood

We assume a Gaussian zero mean noise with a constant noise variance $\boldsymbol{\epsilon}_{fn} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_b)$, where $\boldsymbol{\Sigma}_b = \sigma_b^2 \mathbf{I}$. This gives us the possibility to formulate the likelihood of the mixture coefficients as follows:

$$p(\mathbf{X}|\mathbf{S}, \mathbf{A}) = \prod_{n=1}^{N} \prod_{f=1}^{F} \mathcal{N}(\mathbf{x}_{fn}|\mathbf{A}_f \mathbf{s}_{fn}, \sigma_b^2 \mathbf{I}). \tag{6}$$

In the matrix representation, $\mathbf{X} = \{\mathbf{x}_{fn}\}_{n=1\cdots N, f=1\cdots F}$ and $\mathbf{S} = \{\mathbf{s}_{fn}\}_{n=1\cdots N, f=1\cdots F}$.

2.2 Joint Distribution

Let us define $\mathbf{Z}$ to be the set of model parameters given as

$$\mathbf{Z} = \{\mathbf{S}, \mathbf{A}, \mathbf{W}^{\text{ex}}, \mathbf{U}^{\text{ex}}, \mathbf{G}^{\text{ex}}, \mathbf{H}^{\text{ex}}, \mathbf{W}^{\text{ft}}, \mathbf{U}^{\text{ft}}, \mathbf{G}^{\text{ft}}, \mathbf{H}^{\text{ft}}\}. \tag{7}$$

With the prior information assumed in the previous section, the joint distribution $p(\mathbf{X}, \mathbf{Z})$ is given by

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}) =& p(\mathbf{S}|\mathbf{W}^{\text{ex}}, \mathbf{U}^{\text{ex}}, \mathbf{G}^{\text{ex}}, \mathbf{H}^{\text{ex}}, \mathbf{W}^{\text{ft}}, \mathbf{U}^{\text{ft}}, \mathbf{G}^{\text{ft}}, \mathbf{H}^{\text{ft}}) \\ & p(\mathbf{X}|\mathbf{S}, \mathbf{A}) p(\mathbf{A}) p(\mathbf{W}^{\text{ex}}) p(\mathbf{U}^{\text{ex}}) p(\mathbf{G}^{\text{ex}}) p(\mathbf{H}^{\text{ex}}) \\ & p(\mathbf{W}^{\text{ft}}) p(\mathbf{U}^{\text{ft}}) p(\mathbf{G}^{\text{ft}}) p(\mathbf{H}^{\text{ft}}). \end{aligned} \tag{8}$$

3 Variational Inference

We aim to estimate the posterior distribution of the model parameters $p(\mathbf{Z}|\mathbf{X})$. This estimation is intractable and we resort to a variational Bayesian approximation [5].

3.1 General Approach

Marginalizing out the model parameters from the joint posterior shown in (8) gives us the marginal likelihood or the so-called evidence.

We can also formulate the log marginal likelihood as follows:

$$\log p(\mathbf{X}) = \mathscr{L}(q) + KL(q||p), \tag{9}$$

where $\mathscr{L}(q)$ and $KL(q||p)$ are defined as

$$\mathscr{L}(q) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})}, \tag{10}$$

$$KL(q||p) = -\int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})}. \tag{11}$$

In this formulation, $q(\mathbf{Z})$ is the joint variational distribution of the model parameters, which is used for approximating the real posterior distribution $p(\mathbf{Z}|\mathbf{X})$ of the model parameters. $\mathscr{L}(q)$ is called the free energy and it is a lower bound of the marginal likelihood [5]. Maximizing this lower bound with respect to (wrt.) $q(\mathbf{Z})$ maximizes the marginal likelihood and minimizes the KL-divergence $KL(q||p)$ between $q(\mathbf{Z})$ and the true posterior $p(\mathbf{Z}|\mathbf{X})$. Note that the KL-divergence vanishes, when $q(\mathbf{Z})$ is equal to the true posterior.

In order to make a closed form solution possible, the joint distribution of the parameters $q(\mathbf{Z})$ is factorized into Δ number of approximating distributions by assuming independence between the parameters. Subsequently, maximizing the free energy wrt. a certain approximating distribution $q_{\mathbf{Z}_\delta}$ by keeping all the other approximating distributions $\delta' \neq \delta$ constant [5], we obtain a general equation for the solution of the optimal approximatin distribution $q^*_{\mathbf{Z}_\delta}$ maximizing the lower bound as follows:

$$\log q^*_\delta(\mathbf{Z}_\delta) = \mathbb{E}_{\delta' \neq \delta}[\log p(\mathbf{X}, \mathbf{Z})] + \text{const.} \tag{12}$$

where the normalizing constant is such that q^*_δ is a proper probability distribution.

Note that in (12), the optimal approximating distribution $q^*(\mathbf{Z}_\delta)$ depends on the expectation of the joint posterior distribution wrt. all other factors $\delta' \neq \delta$. This leads to perform an expectation-maximization (EM) like iterative optimization algorithm, where in the variational E-step, the expectations are computed and in the variational M-step these expectations are used for computing the parameters of the approximating distributions.

Note that in this equation, the update equation of the optimal approximating distribution $q^*_\delta(\mathbf{Z}_\delta)$ depends on the expectation of the log of the joint distribution wrt. all other variational distributions. Therefore an iterative update procedure is needed. After proper initialization of all the variational distributions, each distribution is updated in an iterative cycle.

In practice, (12) is applicable only when $\mathbb{E}_{\delta' \neq \delta}[\log p(\mathbf{X}, \mathbf{Z})]$ is computable in closed form and corresponds to a known parametric distribution for which the normalizing constant is computable in closed form. When this is not the case, $p(\mathbf{X}, \mathbf{Z})$ must be replaced by a lower bound for which the resulting approximating distribution becomes tractable.

Let us consider a parametric lower bound $f(\mathbf{X}, \mathbf{Z}, \mathbf{\Omega})$ of $p(\mathbf{X}, \mathbf{Z})$ such that

$$p(\mathbf{X}, \mathbf{Z}) \geq f(\mathbf{X}, \mathbf{Z}, \mathbf{\Omega}), \tag{13}$$

where $\mathbf{\Omega}$ is a set of auxiliary variables. Using this definition, we define $\mathscr{B}$, which further lower bounds $\mathscr{L}$ as in the following

$$\mathscr{L}(q) \geq \mathscr{B}(q, \mathbf{\Omega}) = \int q(\mathbf{Z}) \log \frac{f(\mathbf{X}, \mathbf{Z}, \mathbf{\Omega})}{q(\mathbf{Z})} d\mathbf{Z}. \tag{14}$$

By maximizing the lower bound $\mathscr{B}(q, \mathbf{\Omega})$ wrt. $\mathbf{\Omega}$, which tightens the lower bound to the free energy $\mathscr{L}(q)$ and by maximizing it wrt. each of the approximating distributions q_δ in a subsequent step using the same factorization we obtain

$$q^*_\delta(\mathbf{Z}_\delta) = \tilde{f}(\mathbf{X}, \mathbf{Z}_\delta, \mathbf{\Omega}), \tag{15}$$

where

$$\log \tilde{f}(\mathbf{X}, \mathbf{Z}_\delta, \mathbf{\Omega}) = \mathbb{E}_{\delta' \neq \delta}[\log f(\mathbf{X}, \mathbf{Z}, \mathbf{\Omega})]. \tag{16}$$

In the following, we adopt this strategy for the derivation of the approximating distributions of the multilevel NMF parameters and the source components since $\mathbb{E}[\log p(\mathbf{S}|\mathbf{V})]$ is not tractable in closed form.

3.2 Variational Inference of the Local Gaussian Source Separation Model

Pursuing the variational inference, we fully factorize our $q(\mathbf{Z})$, where Z is given in (7), i.e., all variables are independent. As an example, for the source coefficients, the factorization means independence in each time-frequency bin. Similarly, the mixing matrix coefficients are independent

in each frequency bin. Finally, each multilevel NMF parameter, shown in (4, 5) is independently distributed.

Let us define $\eta = \{k, m, l\}$ the joint index of the excitation multilevel NMF parameters and $\eta' = \{k', m', l'\}$ the joint index of the filter multilevel NMF parameters. With these joint indices let us further define $v_{j,fn,\eta,\eta'}$ as the product of the multilevel NMF parameters

$$v_{j,fn,\eta,\eta'} = h^{\text{ex}}_{j,mn} g^{\text{ex}}_{j,km} u^{\text{ex}}_{j,lk} w^{\text{ex}}_{j,fl} h^{\text{ft}}_{j,m'n} g^{\text{ft}}_{j,k'm'} u^{\text{ft}}_{j,l'k'} w^{\text{ft}}_{j,fl'}. \tag{17}$$

Let us further define $v^{\text{ex}}_{j,fn,\eta}$ sd the product of the excitation coefficients of the multilevel NMF components and $v^{\text{ft}}_{j,fn,\eta'}$ as the product of the filter coefficients

$$v^{\text{ex}}_{j,fn,\eta} = h^{\text{ex}}_{j,mn} g^{\text{ex}}_{j,km} u^{\text{ex}}_{j,lk} w^{\text{ex}}_{j,fl}, \tag{18}$$

$$v^{\text{ft}}_{j,fn,\eta'} = h^{\text{ft}}_{j,m'n} g^{\text{ft}}_{j,k'm'} u^{\text{ft}}_{j,l'k'} w^{\text{ft}}_{j,fl'}. \tag{19}$$

Having these variables defined for the sake of readibility, let us have a look at $\mathbb{E}[\log p(\mathbf{S}|\mathbf{V})]$ more closely.

$$\begin{aligned}\mathbb{E}[\log p(\mathbf{S}|\mathbf{V})] = &\sum_{f,n} -J \log \pi - \sum_{j} \mathbb{E}\Big[\log \sum_{\eta} \sum_{\eta'} v_{j,fn,\eta,\eta'}\Big] \\ &- \sum_{j} \mathbb{E}[|s_{j,fn}|^2] \mathbb{E}[\frac{1}{\sum_{\eta} \sum_{\eta'} v_{j,fn,\eta,\eta'}}\Big].\end{aligned} \tag{20}$$

Note that none of the expectations containing $v_{j,fn,\eta,\eta'}$ in (20) is tractable. So, we resort to the alternative method and lower bound $p(\mathbf{S}|\mathbf{V})$ as explained in the following [11]. For the first expectation, we know that $x \to -\log x$ is convex. So, we can lower bound it by its first-order Taylor series expansion around an arbitrary positive point $\omega_{j,fn}$ as follows

$$-\log \sum_{\eta} \sum_{\eta'} v_{j,fn,\eta,\eta'} \geq -\log \omega_{j,fn} + 1 - \frac{1}{\omega_{j,fn}} \sum_{\eta} \sum_{\eta'} v_{j,fn,\eta,\eta'}. \tag{21}$$

For the second expectation, we know that $x \to \frac{-1}{x}$ concave. So, for any positive $\phi_{j,fn,\eta,\eta'}$ such that $\sum_{\eta} \sum_{\eta'} \phi_{j,fn,\eta,\eta'} = 1$ we have

$$-\frac{1}{\sum_{\eta} \sum_{\eta'} v_{j,fn,\eta,\eta'}} \geq -\sum_{\eta} \sum_{\eta'} \phi^2_{j,fn,\eta,\eta'} \frac{1}{v_{j,fn,\eta,\eta'}}. \tag{22}$$

With these two inequalities, we can lower bound $\log p(\mathbf{S}|\mathbf{V})$ using the auxiliary variables $\mathbf{\Omega} = \{\{\omega_{j,fn}\}_{j,fn}, \{\phi_{j,fn,\eta,\eta'}\}_{j,fn,\eta,\eta'}\}$ as follows

$$\begin{aligned}\log p(\mathbf{S}|\mathbf{V}) \geq &-F \cdot N \cdot J \cdot \log \pi \\ &+ \sum_{j,fn} \big(-\log \omega_{j,fn} + 1 - \frac{1}{\omega_{j,fn}} \sum_{\eta} \sum_{\eta'} v_{j,fn,\eta,\eta'} \big) \\ &- \sum_{j,fn} |s_{j,fn}|^2 \sum_{\eta} \sum_{\eta'} \phi^2_{j,fn,\eta,\eta'} \frac{1}{v_{j,fn,\eta,\eta'}}.\end{aligned} \tag{23}$$

Having this lower bound, we have to tighten it wrt. the auxiliary variables ϕ and ω.

3.2.1 Tightening the Bound wrt. the Auxiliary Variables

After updating the variational variables, we need to update $\phi_{j,fn,\eta,\eta'}$ and $\omega_{j,fn}$ to re-tighten the lower bound. For $\omega_{j,fn}$, we simply take the partial derivative of the bound wrt. $\omega_{j,fn}$ and make

it equal to zero, which yields:

$$\omega_{j,fn} = \sum_{\eta} \sum_{\eta'} \mathbb{E}[v_{j,fn,\eta,\eta'}]. \tag{24}$$

For $\phi_{j,fn,\eta,\eta'}$, we use the Lagrange multipliers, because of the constraint. Solving the system of equations for $\phi_{j,fn,\eta,\eta'}$ yields the following

$$\phi_{j,fn,\eta,\eta'} = \frac{1}{C_{j,fn}} \mathbb{E}\Big[\frac{1}{v_{j,fn,\eta,\eta'}}\Big]^{-1}, \tag{25}$$

where $C_{j,fn}$ is the normalization constant given by

$$C_{j,fn} = \sum_{\eta} \sum_{\eta'} \mathbb{E}\Big[\frac{1}{v_{j,fn,\eta,\eta'}}\Big]^{-1}. \tag{26}$$

3.2.2 Variational Updates for the Multilevel NMF Parameters

In this section, we will determine the optimal approximating distributions for the multilevel NMF parameter $w^{\text{ex}}_{j,fl}$. The probability distribution $\tilde{f}(\mathbf{X}, w^{\text{ex}}_{j,fl}, \mathbf{\Omega})$ defined in (16) is given by

$$\begin{aligned}
\log \tilde{f}(\mathbf{X}, w^{\text{ex}}_{j,fl}, \mathbf{\Omega}) = w^{\text{ex}}_{j,fl} \Big(& \sum_{n} - \log \omega_{j,fn} + 1 \\
& - \frac{1}{\omega_{j,fn}} \sum_{k,m} \sum_{\eta'} \mathbb{E}[h^{\text{ex}}_{j,mn} g^{\text{ex}}_{j,km} u^{\text{ex}}_{j,lk} v^{\text{ft}}_{j,fn,\eta'}] \Big) \\
& + \frac{1}{w^{\text{ex}}_{j,fl}} \Big(\sum_{n} -\mathbb{E}[|s_{j,fn}|^2] \\
& \sum_{k,m} \sum_{\eta'} \phi^2_{j,fn,\eta,\eta'} \mathbb{E}\Big[\frac{1}{h^{\text{ex}}_{j,mn} g^{\text{ex}}_{j,km} u^{\text{ex}}_{j,lk} v^{\text{ft}}_{j,fn,\eta'}}\Big] \Big) \\
& - \log w^{\text{ex}}_{j,fl} + \text{const.}
\end{aligned} \tag{27}$$

Observing this distribution, one can see that it involves a linear term in $w^{\text{ex}}_{j,fl}$ and a linear term in $\frac{1}{w^{\text{ex}}_{j,fl}}$. The optimal approximating distribution $q^*(w^{\text{ex}}_{j,fl})$ is hence an instance of the generalized inverse Gaussian (GIG) distribution, whose probability distribution function (PDF) is given by

$$GIG(y; \gamma, \rho, \tau) = \frac{\exp\{(\gamma - 1)\log y - \rho y - \frac{\tau}{y}\} \rho^{\frac{\gamma}{2}}}{2\tau^{\frac{\gamma}{2}} K_{\gamma}(2\sqrt{\rho\tau})}, \tag{28}$$

for $y \geq 0$, $\rho \geq 0$ and $\tau \geq 0$, where $K_{\gamma}(\cdot)$ is the modified Bessel function of the second kind. The gamma distribution is a special case of the GIG distribution [12] when $\tau = 0$ and $\gamma > 0$. Similarly, the inverse gamma distribution is another special case [12] when $\rho = 0$ and $\gamma < 0$.

hence, using an approach similar to "completing the square" [5], we obtain the update equations for $\tau^{\text{ex}}_{w,j,fl}$, $\rho^{\text{ex}}_{w,j,fl}$ and $\gamma^{\text{ex}}_{w,j,fl}$ in matrix form as follows

$$\begin{aligned}
\boldsymbol{\tau}^{\text{ex}}_{w,j} = \mathbb{E}\Big[\frac{1}{\mathbf{W}^{\text{ex}}_{j}}\Big]^{\cdot -2} \odot \Bigg(& \Big(\mathbb{E}[|\mathbf{S}_j|^{\cdot 2}] \odot \mathbf{C}^{\cdot -2}_{\phi,j} \odot \mathbb{E}\Big[\frac{1}{\mathbf{V}^{\text{ft}}_{j}}\Big]^{\cdot -1} \Big) \\
& \Big(\mathbb{E}\Big[\frac{1}{\mathbf{U}^{\text{ex}}_{j}}\Big]^{\cdot -1} \mathbb{E}\Big[\frac{1}{\mathbf{G}^{\text{ex}}_{j}}\Big]^{\cdot -1} \mathbb{E}\Big[\frac{1}{\mathbf{H}^{\text{ex}}_{j}}\Big]^{\cdot -1} \Big)^T \Bigg).
\end{aligned} \tag{29}$$

$$\boldsymbol{\rho}_{w,j}^{\text{ex}} = R_j \mathbb{E}[\mathbf{V}_j^{\text{ex}}]^{\cdot -1} \big(\mathbb{E}[\mathbf{U}_j^{\text{ex}}]\mathbb{E}[\mathbf{G}_j^{\text{ex}}]\mathbb{E}[\mathbf{H}_j^{\text{ex}}]\big)^T. \tag{30}$$

Finally for $\gamma_{w,j,fl}^{\text{ex}}$, we deduce that

$$\gamma_{w,j,fl}^{\text{ex}} = 0. \tag{31}$$

Note that in (29, the power operations like $\mathbf{X}^{\cdot -a}$ are element-wise operations. Furthermore, the symbol $\odot$ means element-wise matrix multiplication. The expectation $\mathbb{E}[|\mathbf{S}_j|^{\cdot 2}]$ is calculated also element-wise as follows

$$\mathbb{E}[|s_{j,fn}|^{\cdot 2}] = |\mu_{s,j,fn}|^2 + (\mathbf{R}_{\mathbf{ss},fn})_{jj}. \tag{32}$$

The expectations related to the generalized inverse Gaussian variables include $\mathbb{E}[y]$ and $\mathbb{E}[\frac{1}{y}]$. These expectations are computed using the following two formulae [12]

$$\mathbb{E}[y] = \frac{\mathcal{K}_{\gamma+1}(2\sqrt{\rho\tau})\sqrt{\tau}}{\mathcal{K}_{\gamma}(2\sqrt{\rho\tau})\sqrt{\rho}}, \tag{33}$$

$$\mathbb{E}[\frac{1}{y}] = \frac{\mathcal{K}_{\gamma-1}(2\sqrt{\rho\tau})\sqrt{\rho}}{\mathcal{K}_{\gamma}(2\sqrt{\rho\tau})\sqrt{\tau}}. \tag{34}$$

For the other multilevel NMF parameters, the derivations are performed by following the same steps. For the update equations of the other multilevel NMF paremeters and their derivations please refer to [3].

3.2.3 Variational Updates for the Source Components

The distribution $\tilde{f}(\mathbf{X}, \mathbf{s}_{fn}, \boldsymbol{\Omega})$ of the source components $\mathbf{s}_{fn}$ is given by

$$\begin{aligned}\log \tilde{f}(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Omega})_{\mathbf{S}} = {} & \mathbf{s}_{fn}^H \boldsymbol{\mu}_{\mathbf{A},f}^H (\sigma_b^2 \mathbf{I})^{-1} \mathbf{x}_{fn} + \mathbf{x}_{fn}^H (\sigma_b^2 \mathbf{I})^{-1} \boldsymbol{\mu}_{\mathbf{A},f} \mathbf{s}_{fn} \\ & - \frac{1}{\sigma_b^2} (\mathbf{s}_{fn}^H \mathbf{R}_{\mathbf{A},f} \mathbf{s}_{fn}) \\ & - \mathbf{s}_{fn}^H {\mathbf{C}_{fn}}^{-1} \mathbf{s}_{fn} + \text{const},\end{aligned} \tag{35}$$

where ${\mathbf{C}_{fn}}^{-1} = \text{diag}({C_{j,fn}}^{-1})_{r=1}^{R}$ is a diagonal matrix with the main diagonal containing the normalization factor ${C_{j,fn}}^{-1}$ repeated R_j times for each j. $\mathbf{R}_{\mathbf{A},f}$ is the second raw moment of the mixing parameters in matrix form. Recall that the rows of the mixing matrix $\mathbf{A}_f$ are reshaped into a column vector $\underline{\mathbf{A}}_f$. In order to obtain $\mathbf{R}_{\mathbf{A},f}$, we make use of the expectations of the reshaped mixing parameters as follows

$$\mathbf{R}_{\mathbf{A},f} = \sum_i ([\boldsymbol{\mu}_{\underline{\mathbf{A}},f} \boldsymbol{\mu}_{\underline{\mathbf{A}},f}^H + \mathbf{R}_{\underline{\mathbf{A}}\underline{\mathbf{A}},f}]_{ii})^T. \tag{36}$$

In (36), $[\cdot]_{ii}$ denotes the diagonal $J \times J$ block corresponding to channel i. The details of $\mathbf{R}_{\underline{\mathbf{A}}\underline{\mathbf{A}},f}$ is given in the following section in (44). $\boldsymbol{\mu}_{\mathbf{A},f}$ is the mean of the mixing matrix. Due to the reshaping, for obtaining $\boldsymbol{\mu}_{\mathbf{A},f}$, we reshape the mean $\boldsymbol{\mu}_{\underline{\mathbf{A}},f}$ of the mixing parameters, which is given in (45) back into the matrix form.

The distribution given in (35) involves a linear term in $\mathbf{s}_{fn}$, its conjugate, and quadratic terms. The optimal approximating distribution is thus a Gaussian given by

$$\mathbf{s}_{fn} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{s},fn}, \mathbf{R}_{\mathbf{ss},fn}). \tag{37}$$

By "completing the square" wrt. $\boldsymbol{\mu}_{\mathbf{s},fn}$ and $\mathbf{R}_{\mathbf{ss},fn}$ in (35) we derive the update equations for these two parameters. First we derive the update equation of the covariance. For this, we rearrange the quadratic terms make them equal to $-\mathbf{s}_{fn}^H \mathbf{R}_{\mathbf{ss},fn}^{-1} \mathbf{s}_{fn}$. By doing this we obtain the covariance given in the following

$$\mathbf{R}_{\mathbf{ss},fn} = ({\mathbf{C}_{fn}}^{-1} + (\sigma_b^2 \mathbf{I})^{-1} \mathbf{R}_{\mathbf{A},f})^{-1}, \tag{38}$$

Finally, the mean of the optimal factor of the mixing coefficients is given by the following

$$\boldsymbol{\mu}_{\mathbf{s},fn} = \mathbf{R}_{\mathbf{ss},fn} \boldsymbol{\mu}_{\mathbf{A},f}^H (\sigma_b^2 \mathbf{I})^{-1} \mathbf{x}_{fn}, \tag{39}$$

3.2.4 Variational Updates for the Mixing Parameters

The distribution $\tilde{f}(\mathbf{X}, \underline{\mathbf{A}}_f, \boldsymbol{\Omega})$ of the reshaped mixing parameters $\underline{\mathbf{A}}_f$ is given by

$$\begin{aligned} \log \tilde{f}(\mathbf{X}, \underline{\mathbf{A}}_f, \boldsymbol{\Omega}) = & -\underline{\mathbf{A}}_f^H \frac{1}{\sigma_b^2} \sum_n \operatorname{diag}(\underbrace{\mathbf{R}_{\mathbf{s},fn}^T, \ldots, \mathbf{R}_{\mathbf{s},fn}^T}_{\text{I times}}) \underline{\mathbf{A}}_f \\ & + \underline{\mathbf{A}}_f^H \frac{1}{\sigma_b^2} \sum_n \mathbf{R}_{\mathbf{xs},fn} \\ & + \frac{1}{\sigma_b^2} \Big(\sum_n \mathbf{R}_{\mathbf{xs},fn}^H \Big) \underline{\mathbf{A}}_f + \text{const}, \end{aligned} \tag{40}$$

where $\mathbf{R}_{\mathbf{s},fn}$ is the second raw moment of the source coefficients and is given by

$$\mathbf{R}_{\mathbf{s},fn} = \boldsymbol{\mu}_{\mathbf{s},nf} \boldsymbol{\mu}_{\mathbf{s},nf}^H + \mathbf{R}_{\mathbf{ss},fn}. \tag{41}$$

Similarly, $\mathbf{R}_{\mathbf{xs},fn}$ is given by

$$\mathbf{R}_{\mathbf{xs},fn} = [x_{1,fn} \boldsymbol{\mu}_{\mathbf{s},fn}^H, \ldots, x_{I,fn} \boldsymbol{\mu}_{\mathbf{s},fn}^H]^T. \tag{42}$$

The distribution given in (40) involves a linear term in $\underline{\mathbf{A}}_f$, its conjugate, and a quadratic term. Hence, the optimal approximating distribution is a Gaussian given by

$$\underline{\mathbf{A}}_f \sim \mathcal{N}(\boldsymbol{\mu}_{\underline{\mathbf{A}},f}, \mathbf{R}_{\underline{\mathbf{AA}},f}). \tag{43}$$

By "completing the square" wrt. $\boldsymbol{\mu}_{\underline{\mathbf{A}},f}$ and $\mathbf{R}_{\underline{\mathbf{AA}},f}$ in (40) we derive the update equations for these two parameters. First we derive the update equation of the covariance. For this, we rearrange the quadratic terms make them equal to $-\underline{\mathbf{A}}_f^H \mathbf{R}_{\underline{\mathbf{AA}},f}^{-1} \underline{\mathbf{A}}_f$. By doing this we obtain the covariance given in the following

$$\mathbf{R}_{\underline{\mathbf{AA}},f} = \Big(\frac{1}{\sigma_b^2} \sum_n \operatorname{diag} \underbrace{(\mathbf{R}_{\mathbf{s},fn}^T, \ldots, \mathbf{R}_{\mathbf{s},fn}^T)}_{\text{I times}} \Big)^{-1}, \tag{44}$$

The update equation of the mean of the optimal factor of the mixing coefficients is written as in the following

$$\boldsymbol{\mu}_{\underline{\mathbf{A}},f} = \mathbf{R}_{\underline{\mathbf{AA}},f} (\frac{1}{\sigma_b^2} \sum_n \mathbf{R}_{\mathbf{xs},fn}), \tag{45}$$

We remind you that we assumed a flat multivariate Gaussian prior for the mixing matrix with an infinite covariance matrix. Due to this assumption, the terms containing the covariance

matrix of the prior distribution do not appear neither in the covariance nor in the mean of the optimal factor.

Note that all update equations depend on each other. After proper initialization, we cycle through these equations by replacing the dependent values with their new estimates. In the variational E-step, we calculate the expectations needed within the update equations of each variational parameters and in the variational M-step, the new values of these parameters are computed.

3.3 Lower Bound

We calculate the new lower bound after each variational M-step. Here we compare the new lower bound with the previous one and set a termination condition according to this change. The lower bound should never decrease. This fact enables us to check the update equations and their implementation for their correctness. In general, if the increase in the lower bound is insignificant compared to previous iterations, we stop. The lower bound has already been defined in (10).

Note that the lower bound consists of the expectation of the logarithm of the prior distribution p minus the expectation of the logarithm of the approximating distribution q pairs except for the likelihood. For one of the multilevel NMF parameters, e.g. for $w^{\text{ex}}_{j,c,fl}$, this pair is written as follows:

$$\begin{aligned}\mathbb{E}[\log p(w^{\text{ex}}_{j,fl})] - \mathbb{E}[\log q(w^{\text{ex}}_{j,fl})] &= \rho^{\text{ex}}_{w,j,fl}\mathbb{E}[w^{\text{ex}}_{j,fl}] \\ &+ \tau^{\text{ex}}_{w,j,fl}\mathbb{E}[\frac{1}{w^{\text{ex}}_{j,fl}}] + \log \mathcal{K}_{\gamma^{\text{ex}}_{w,j,fl}}(2\sqrt{\rho\tau}) + \log 2. \end{aligned} \tag{46}$$

The calculation of the contribution of the other multilevel NMF parameters to the lower bound is calculated similarly using he same formula given above.

We remind you that $\mathbb{E}[p(\mathbf{S}|\mathbf{V})]$ is not tractable. We derived an approximation to that as shown in (23). In calculating the lower bound, we use this approximation. The expectation of the logarithm of the optimal factor of the source components $q(\mathbf{S})$ is given by

$$\begin{aligned}\mathbb{E}[\log q(\mathbf{S})] &= \sum_{fn} \mathbb{E}[\log \mathcal{N}(\mathbf{s}_{fn}|\boldsymbol{\mu}_{s,fn}, \mathbf{R}_{ss,fn})], \\ &= -\sum_{fn} \log\{(\pi e)^J \det(\mathbf{R}_{ss,fn})\}. \end{aligned} \tag{47}$$

About the mixing matrix, we recall you that the prior distribution of the mixing coefficients does not contribute to the lower bound, because it is flat. The expectation of the $q(\cdot)$ of the mixing coefficients is given by

$$\begin{aligned}\mathbb{E}[\log q(\underline{\mathbf{A}})] &= \sum_{f} \mathbb{E}[\log \mathcal{N}(\underline{\mathbf{A}}_{fn}|\boldsymbol{\mu}_{\underline{\mathbf{A}},f}, \mathbf{R}_{\underline{\mathbf{A}}\underline{\mathbf{A}},f})], \\ &= -\sum_{f} \log\{(\pi e)^{IJ} \det(\mathbf{R}_{\underline{\mathbf{A}}\underline{\mathbf{A}},f})\}. \end{aligned} \tag{48}$$

Finally, we need to calculate the expectation of the log-likelihood as follows:

$$\begin{aligned}\mathbb{E}[\log p(\mathbf{X}|\mathbf{S},\mathbf{A})] &= -F\cdot N\cdot I\cdot \log \pi\sigma_b^2 \\ &- \sum_{fn}\frac{1}{\sigma_b^2}\Big(\mathbf{x}^H_{fn}\mathbf{x}_{fn} - \mathbf{x}^H_{fn}\boldsymbol{\mu}_{A,f}\boldsymbol{\mu}_{\mathbf{s},fn} - \boldsymbol{\mu}^H_{\mathbf{s},fn}\boldsymbol{\mu}^H_{\mathbf{A},f}\mathbf{x}_{fn} \\ &+ \operatorname{tr}((\boldsymbol{\mu}_{s,nf}\boldsymbol{\mu}^H_{s,nf} + \mathbf{R}_{\mathbf{ss},fn})\mathbf{R}_{\mathbf{A},f})\Big). \end{aligned} \tag{49}$$

4 Uncertainty Propagation

We now present a moment matching method for propagating the uncertainty over the source components to the source images and in a following step to the MFCC features.

4.1 Uncertainty Propagation for the Source Images

Due to phase and scale indeterminacies in the source estimates $\mathbf{s}_{j,fn}$, we use the spatial source images $\mathbf{y}_{j,fn} = \mathbf{A}_{j,f} s_{j,fn}$ instead for our experiments, which do not suffer from such indeterminacies [15].

Once the posterior distribution of the source coefficients $\mathbf{s}_{fn}$ has been computed, the posterior distribution of the source images is calculated by propagating the first two moments of the sources to the source images as follows:

$$\boldsymbol{\mu}_{\mathbf{y},j,fn} = \boldsymbol{\mu}_{\mathbf{A},j,f}\, \boldsymbol{\mu}_{\mathbf{s},j,fn}, \tag{50}$$

$$\begin{aligned}(\boldsymbol{R}_{\mathbf{yy},j,fn})_{(ii')} =& \Big(\sum_{jj'} (\mathbf{R}_{\mathbf{A},f})_{(ij,i'j')} (\mathbf{R}_{\mathbf{s},fn})_{(jj')} \Big) \\ & - (\boldsymbol{\mu}_{\mathbf{y},j,fn} \boldsymbol{\mu}^H_{\mathbf{y},j,fn})_{(ii')}.\end{aligned} \tag{51}$$

Note that in (51), $(\cdot)_{ii'}$ denotes the $(i,i')^{\text{th}}$ element of an $I \times I$ matrix.

4.2 Uncertainty Propagation for Feature Extraction

We calculate the expectation of the MFCCs for each source as

$$\boldsymbol{\mu}^{\mathbf{MFCC}}_{jn} = \int \mathrm{MFCC}(\mathbf{y}_{j1n}) P(\mathbf{y}_{j1n})\, d\mathbf{y}_{j1n} \tag{52}$$

where $\mathbf{y}_{j1n} = [y_{j,1fn}]_{f=1\ldots F}$ are the STFT coefficients of the first channel of source image j in time frame n. Deterministic calculation without the use of the uncertainty model simply yields $\mathrm{MFCC}(\mathbf{y}_{j1n}) = 20\,\mathbf{D} \log_{10}(\mathbf{M}|\boldsymbol{\mu}_{\mathbf{y},j,1n}|)$. In this formulation, $\mathbf{D}$ is the DCT matrix and $\mathbf{M}$ is the matrix containing the mel filter coefficients. Note that we chose the scaling so that the MFCCs are expressed in decibels (dB).

In the moment matching approach, the uncertainty expressed by the mean and variance of the posterior distribution of the estimated source coefficients is propagated through the calculation of the MFCCs.

The absolute value of a complex normal random variable is Rice distributed [13]. Hence, the magnitude spectrum of the estimated source coefficients follows a Rice distribution. For the source j in time-frame n, the mean and variance of the magnitude spectrum is given by using the first and second raw moments of the Rice distribution as:

$$\boldsymbol{\mu}_{|\mathbf{s}|,j,n} = \mathbb{E}[|\mathbf{s}_{j,n}|], \tag{53}$$

$$\boldsymbol{\Sigma}_{|\mathbf{s}|,j,n} = \mathbb{E}[|\mathbf{s}_{j,n}|^2] - \boldsymbol{\mu}^2_{\mathbf{s},j,n}. \tag{54}$$

For the details of the moments of the Rice distribution please refer to [13]. The mel-filtering of the magnitude spectrum is a linear transformation. So, we can simply match the moments:

$$\boldsymbol{\mu}_{\mathrm{MEL},j,n} = \mathbf{M} \cdot \boldsymbol{\mu}_{|\mathbf{s}|,j,n}, \tag{55}$$

$$\boldsymbol{\Sigma}_{\mathrm{MEL},j,n} = \mathbf{M} \cdot \boldsymbol{\Sigma}_{|\mathbf{s}|,j,n} \cdot \mathbf{M}^T. \tag{56}$$

The logarithm in the calculation of the MFCCs is not a linear transformation. Here, we assume the log-normality of the MEL features. Incorporating the log-normal transformation proposed by Gales [10], the i^{th} coefficient is given by:

$$\mu^i_{\text{log},j,n} = \log(\mu^i_{\text{MEL},j,n}) - \frac{1}{2}\log(\frac{\Sigma^i_{\text{MEL},j,n}}{{\mu^i_{\text{MEL},j,n}}^2} + 1). \tag{57}$$

$$\Sigma^{ij}_{\text{log},j,n} = \log(\frac{\Sigma^{ij}_{\text{MEL}}}{\mu^i_{\text{MEL}}\mu^j_{\text{MEL}}} + 1). \tag{58}$$

The final step, the DCT, is another linear transform. Hence, we apply moment matching again:

$$\boldsymbol{\mu}_{\text{MFCC},j,n} = \mathbf{D} \cdot \boldsymbol{\mu}_{\text{log},j,n}, \tag{59}$$

$$\boldsymbol{\Sigma}_{\text{MFCC},j,n} = \mathbf{D} \cdot \boldsymbol{\Sigma}_{\text{log},j,n} \cdot \mathbf{D}^T. \tag{60}$$

5 Experimental Evaluation

We now evaluate the uncertainty estimation and propagation algorithm proposed above separately.

In order to assess the impact of source separation on feature extraction, we evaluate the proposed algorithm according to both tasks. Source separation quality is evaluated in terms of the Signal-to-Distortion Ratio (SDR) in [16] between the mean of the estimated source images $\boldsymbol{\mu}_{\mathbf{y},j,fn}$ and the true source images.

Feature extraction accuracy is evaluated in two different scenarios. In the first scenario, the RMS error [1] between the estimated $\boldsymbol{\mu}_{\text{MFCC}_{jn}}$ and the true MFCCs is calculated. In the second scenario, we performed speaker recognition experiments. For both of these tasks, we ignore the first MFCC coefficient and consider the MFCCs 2 to 20 only.

5.1 Data and Algorithmic Settings

5.1.1 Data

For the performance evaluation of the uncertainty estimation and the evaluation of the accuracy of the features in the uncertainty propagation, we considered the development dataset of the 2008 Signal Separation Evaluation Campaign (SiSEC) [1]. This dataset contains synthetic and live recorded convolutive, under-determined, stereo mixtures. There are 32 mixtures of 3 sources and 24 mixtures of 4 sources. Each mixture has a duration of 10 s.

For the speaker recognition task, we used the dataset of the 2011 Computational Hearing in Multisource Environment (CHiME) challenge [7]. However from the training dataset of CHiME, we generated our own dataset by using the clean speech utterances and clean background noise recordings. We mixed them in seven different conditions, namely clean speech (muted background), signal to noise ratio (SNR) at -6, -3, 0, 3, 6, and 9dB. There are 680 audio sound samples in the training and test sets for each of these conditions covering 34 speakers. For more details about the dataset, please refer to [14].

5.1.2 Algorithmic Settings

For the evaluation of the uncertainty estimation and uncertainty propagation tasks, we performed experiments with eight different sets of constraints over the parameters as considered

[1] http://sisec2008.wiki.irisa.fr/tiki-index.php?page=Under-determined+speech+and+music+mixtures

in the experiments section of [15]. These scenarios consist of all combinations of the following possibilities:

- Rank: Each source is either a single point source (1) or a subspace spanned by two point sources (2).
- Spectral Structure: The narrowband spectral patterns $w^{\text{ex}}_{j,fl}$ are either unconstrained (un) or fixed (co) to harmonic and noise-like patterns.
- Temporal Structure: The time-localized patterns $h^{\text{ex}}_{j,mn}$ are either unconstrained (un) or fixed (co) to decreasing exponential patterns.

All other parameters are free. We initialize the mixing matrix $\boldsymbol{A}_j$ using the direction of arrival (DOA) estimation algorithm proposed in [6]. The noise variance σ_b^2 is initialized to 10^{-2}. An annealing mechanism is applied to the noise variance and it is gradually decreased to 10^{-6}. With these settings, we performed 200 iterations for convergence.

For the speaker recognition task, we first trained a general speaker model with the proposed algorithm using 15 clean speech samples randomly selected for each speaker from the training data set. We updated only the narrowband spectral patterns $\mathbf{W}^{\text{ex}}$ and the weights of the time-localized patterns $\mathbf{G}^{\text{ex}}$. We performed at most 100 iterations without annealing and controlled the convergence using the lower bound. After convergence, we saved only the narrowband spectral patterns $\mathbf{W}^{\text{ex}}$ and used them for the initialization of the signal enhancement.

In the signal enhancement step, we performed a two-fold source separation following the same steps as in [14] but using the proposed algorithm. In the first source separation step, we trained the model for separating the background. We performed 30 update iterations with annealing. In the second source separation step, having the background model and the general speaker model, we separated the speaker from the background. In this step, we performed 50 iterations with annealing.

After having the enhanced speech signals, we extracted the MFCCs using the moment matching method as described in Section 4.2. In the final step, we performed speaker recognition experiments using the Gaussian mixture models (GMM) as described in [14], where we trained the GMMs using the clean speech signals and tested them using different SNR conditions described in the previous section.

5.2 Results

	1-un-un	2-un-un	1-co-un	2-co-un	1-un-co	2-un-co	1-co-co	2-co-co
ML	1.58	1.68	1.87	2.07	1.77	1.75	2.28	2.25
VB GIG	1.70	1.78	1.95	2.10	1.92	1.85	**2.54**	2.35

Table 1: SDR in dB achieved by ML or VB source separation over all mixtures.

Table 1 shows the source separation performance of the proposed VB method with GIG optimal factors for the multilevel NMF parameters compared to that of the state-of-the-art ML method in [15]. Both algorithms perform similarly in almost all of the eight configurations. However VB GIG is 0.26 dB better than the state-of-the-art ML method and yields the best performance for the one point source, spectrally and temporally constrained model (1-co-co) with 2.54 dB SDR. The baseline binary masking method [6] yields 0.95 dB SDR.

Table 2 shows the total RMS error in dB over the MFCCs obtained either by deterministic computation or by moment matching for the VB algorithm and for the state-of-the-art ML

	1-un-un		2-un-un		1-co-un		2-co-un		1-un-co		2-un-co		1-co-co		2-co-co	
	det	mm	det	mm	det	mm	det	mm	det	mm	det	mm	det	mm	det	m
ML	7.55	6.66	8.62	6.85	7.35	6.72	8.01	6.82	7.43	6.59	8.41	6.76	7.67	6.61	9.28	7.
VB GIG	7.49	6.63	8.35	6.80	7.24	6.66	7.75	6.75	7.38	6.55	8.16	6.71	7.55	**6.50**	8.96	7.

Table 2: Total RMS error in dB for the MFCCs 2-20 obtained by ML- or VB-based source separation followed by deterministic (det) or moment matching (mm) feature extraction over all mixtures.

algorithm. As one can see, VB based MFCC estimation algorithm perform 0.11 dB better than the ML based estimation. Besides, the moment matching method outperforms deterministic MFCC estimation in all configurations with around 0.9 dB. Again, VB GIG performs best for the RMS error by 6.50 dB for the one point source, spectrally unconstrained, temporally constrained model (1-un-co). The baseline binary masking method [6] performs significantly worse than both VB and ML algorithms and yields 17.25 dB RMS error.

	−6dB		−3dB		0dB		3dB		6dB		9
	det	mm	det	mm	det	mm	det	mm	det	mm	det
ML	33.53%	29.85%	34.26%	32.65%	47.50%	43.09%	60.74%	60.74%	72.06%	72.35%	83.24%
VB GIG	28.68%	46.18%	31.76%	53.24%	42.06%	72.65%	58.97%	84.85%	72.94%	94.26%	80.15%
baseline	40.44%		41.32%		58.09%		74.12%		84.71%		92

Table 3: Classification accuracies of the speaker recognition experiments performed with the MFCCs obtained by using the moment matching for the state-of-the-art ML algorithm and for the proposed VB algorithm.

Finally, Table 3 shows the results of the speaker recognition experiments for three methods. The baseline method simply uses clean speech samples without any source separation and / or enhancement applied for training and evaluates the performance on the noisy test data in a conventional way. The other two methods are the state-of-the-art ML method as described in [14] and the proposed VB GIG method as described in Section 5.1.2. Both perform a source separation step to enhance the quality of the target speech signal and perform speaker recognition on the separated speech signal. In all three methods, the speaker recognition is performed by the GMMs using the MFCC features. The deterministic MFCC extraction has been used in the baseline case. For the other two methods the MFCCs are extracted deterministically or using moment matching. Table 3 shows that VB GIG method together with the moment matching clearly outperforms the other two methods. Surprisingly, the baseline method performed better than the ML method both for the deterministic MFCC extraction and for moment matching. Using moment matching for the MFCC extraction even degrades the performance in the ML case. On the other hand, there is a significant improvement, when the proposed VB GIG method was used together with the moment matching MFCC extraction.

6 Conclusion

In this paper, we presented a general, fully Bayesian source separation algorithm based on the variational inference method. In this algorithm, we proposed the generalized inverse Gaussian distribution for the multilevel NMF parameters of the source variances and thereby obtained closed form update equations for the variational parameters without introducing any other parameters (*e.g.* source sub-components). We also provided an approximation to the variational

lower bound, which can be used for observing the convergence as well as for terminating the iterations. Furthermore, we presented an uncertainty propagation algorithm for the computation of the expectation of the MFCCs of individual sources in multisource recordings.

This algorithm provides a fundamental breakthrough towards mathematically rigorous estimation of uncertainty for robust feature extraction. The resulting MFCC coefficients are slightly more accurate than those obtained via the previous variational inference algorithm proposed by the same authors as well as those obtained via the standard ML method. With this rigorous method, we show that the ML method provides a reasonable approximation, but it is still possible to obtain more accurate estimates.

References

[1] K. Adiloglu and E. Vincent. An uncertainty estimation approach for the extraction of source features in multisource recordings. In *Proceedings of 19th European Signal Processing Conference (EUSIPCO)*, pages 1663–1667, 2011.

[2] K. Adiloglu and E. Vincent. A general variational bayesian framework for robust feature extraction in multisource recordings. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012. to appear.

[3] K. Adiloglu, E. Vincent, and F. Bimbot. A general variational inference algorithm for source separation. Technical report, INRIA Rennes, Bretagne - Atlantique, 2012.

[4] R. F. Astudillo and R. Orglmeister. A MMSE estimator in mel-cepstral domain for robust large vocabulary automatic speech recognition using uncertainty propagation. In *Proceedings of Interspeech*, pages 713–716, 2010.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] C. Blandin, E. Vincent, and A. Ozerov. Multi-source TDOA estimation using SNR-based angular spectra. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2616–2619, 2011.

[7] H. Christensen, J. Barker, N. Ma, and P. Green. The chime corpus: a resource and a challenge for computational hearing in multisource environment (chime). In *Proceedings of Interspeech*, pages 1918–1921, 2008.

[8] M. Delcroix, T. Nakatani, and S. Watanabe. Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing. *IEEE Transactions on Audio, Speech and Language Processing*, 17(2):324–334, 2009.

[9] N.Q.K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech and Language Processing*, 18(7):1830–1840, July 2010.

[10] M. J. F. Gales. *Model-Based Techniques for Noise Robust Speech Recognition*. PhD thesis, Gonville and Caius College, University of Cambridge, 1995.

[11] M. D. Hoffman, D. M. Blei, and P. R. Cook. Bayesian nonparametric matrix factorization for recorded music. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

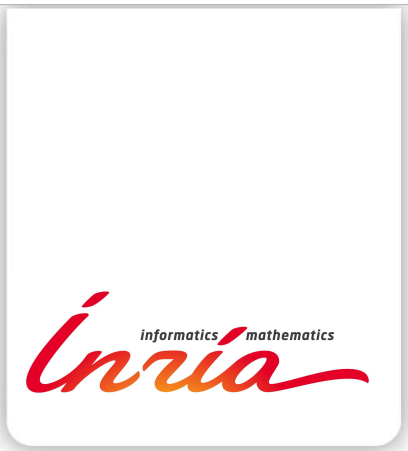
[12] B. Jorgensen. *Statistical Properties of the Generalized Inverse-Gaussian Distribution.* Springer, 1982.

[13] D. Kolossa, R. F. Astudillo, E. Hoffmann, and R. Orglmeister. Independent component analysis and time-frequency masking for speech recognition in multitalker conditions. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, Article ID 651420, 2010.

[14] A. Ozerov, M. Lagrange, and E. Vincent. Uncertainty-based learning of gaussian mixture models from noisy data. Technical report.

[15] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(4):1118–1133, 2012. http://hal.inria.fr/inria-00536917/PDF/RR-7453.pdf.

[16] E. Vincent, R. Gribonval, and C. Févotte. Performance measures in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.



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 0249-0803