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 and Emmanuel Vincent

***Abstract*—We consider the task of separating and classifying individual sound sources mixed together. The main challenge is to achieve robust classification despite residual distortion of the separated source signals. A promising paradigm is to estimate the uncertainty about the separated source signals and to propagate it through the subsequent feature extraction and classification stages. We argue that variational Bayesian (VB) inference offers a mathematically rigorous way of deriving uncertainty estimators, which contrasts with state-of-the-art estimators based on heuristics or on maximum likelihood (ML) estimation. We propose a general VB source separation algorithm, which makes it possible to jointly exploit spatial and spectral models of the sources. This algorithm achieves 6% and 5% relative error reduction compared to ML uncertainty estimation on the CHiME noise-robust speaker identification and speech recognition benchmarks, respectively, and it opens the way for more complex VB approximations of uncertainty.**

I. Introduction

Audio classification tasks such as speaker or singer identification are addressed by training classifiers on features extracted from the audio. In practical situations, the target signal is often mixed with other sound sources such as environmental noise or musical accompaniment which distort the features and degrade classification accuracy. This problem also arises for automatic speech recognition and more general music information retrieval tasks [1], [2].

A first approach is to attempt to reduce distortion by separating the target from the other sources [3]–[6]. Many source separation algorithms based on generative models have been proposed in the literature, ranging from early algorithms for independent component analysis (ICA), binaural cue clustering and sparse component analysis (SCA) to more recent algorithms for factorial hidden Markov models (HMM) and nonnegative matrix factorization (NMF) [7]–[9]. Nonnegative tensor factorization (NTF) has recently attracted some interest due to its ability to incorporate prior knowledge about the sources to guide the separation [10]. The flexible audio source separation (FASST) framework in [11] merges the concepts of NTF and multichannel Gaussian modeling [12]. It generalizes a number of algorithms [13], [14] and makes it possible to jointly exploit spatial and spectral cues, which often improves separation [11]. Prior knowledge about the spectro-temporal properties of the sources such as harmonicity and continuity can be accounted via deterministic or probabilistic constraints on the NTF parameters, while knowledge about their spatial properties is incorporated by constraints on the spatial covariance matrices. This framework and its variants have been applied for the separation of noisy speech [15], [16], movie soundtracks [17], [18], and music [11]. They have also been used as a preprocessing step for automatic speech recognition [19], musical instrument recognition [5], singing voice detection [20], and audio declipping [21].

Most algorithms rely on maximum likelihood (ML) or maximum a posteriori (MAP) estimation of the model parameters and the source signals, which is sensitive to overfitting [22]. A few algorithms have been designed to improve robustness to overfitting by conducting full Bayesian inference using Markov chain Monte Carlo (MCMC) [23]–[28] or variational Bayes (VB) [29]–[35]. This has been done in particular for SCA [23], local Gaussian modeling [26], binaural cue clustering [33], Markov or Gaussian process continuity models [25], [34], factorial HMM [30], and single-channel NMF [35], but Bayesian inference algorithms for joint spatial and spectral modeling frameworks able to account for advanced spectro-temporal properties for any number of channels such as FASST are lacking.

In practice, source separation is rarely perfect and the target signal remains somewhat distorted. Uncertainty propagation has emerged as a promising complementary approach whereby the separated target signal and its features are not considered as point estimates anymore but their posterior distribution is approximated as a Gaussian with time-varying variance or *uncertainty* that is exploited by the classifier [3], [4]. This approach was introduced for noise-robust automatic speech recognition [36]–[42] and it has also been used for noise-robust speaker identification [43], [44] and singer identification in polyphonic music [45]. While there exist techniques to propagate uncertainty from the separated signal to the features based on moment matching [46], unscented transform [38], or Vector Taylor series (VTS) [47], the estimation of uncertainty on the separated signal remains a difficult problem. A heuristic is to assume that the uncertainty is proportional to the squared difference between the separated target and the mixture in the time-frequency domain [38]. In [40], [41], [48], more principled uncertainty estimators were proposed whose mean and variance are derived from ML estimates of the parameters of the source models. These estimators are fundamentally biased, however, since they do not account for the uncertainty about the parameter estimates themselves, which depends on the test signal.

This paper provides two contributions. Firstly, we argue that full Bayesian inference offers a mathematically rigor-

K. Adiloğlu is with HörTech gGmbH, D-26129 Oldenburg, Germany (e-mail: k.adiloglu@hoertech.de). E. Vincent is with Inria, F-54600 Villers-lès-Nancy, France (emmanuel.vincent@inria.fr). Part of this work was conducted while both authors were with Inria, F-35042 Rennes Cedex, France, and it was supported by OSEO under the Quaero program.

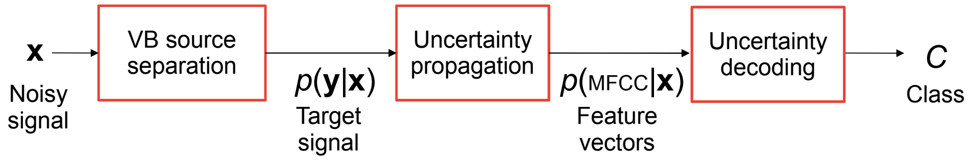

Fig. 1. Flow of the proposed Bayesian source separation and feature extraction approach.

ous way of deriving uncertainty estimators by marginalizing the Wiener uncertainty estimator over the parameters of the source models. Secondly, we propose a general VB inference algorithm for the FASST modeling framework. The proposed algorithm relies on a tighter variational approximation than the one in our preliminary paper [49] and it is to our knowledge the first source separation algorithm able to jointly infer spatial parameters and NMF-based spectral parameters in a full Bayesian sense. We conduct an extensive experimental evaluation in terms of source separation, feature extraction and classification/recognition performance. Fig. 1 illustrates the overall workflow of the proposed approach.

The paper is organized as follows. Section II summarizes the FASST modeling framework. We present the proposed VB inference algorithm in Section III and its usage for uncertainty propagation in Section IV. We describe our experimental evaluation in Section V and we conclude in Section VI.

II. General Source Separation Framework

We consider the FASST modeling framework for source separation in [11]. For simplicity, we focus on the case when all sources are point sources with rank-1 spatial covariance matrix. The extension to diffuse or reverberated sources of any rank is treated in [11] for ML estimation and in the supporting technical report [50] for VB estimation.

In the rest of this article, we denote scalars by plain letters, vectors by bold lowercase letters, and matrices or tensors by bold uppercase letters, respectively. The indices are used as subscripts throughout the paper. Furthermore, we separate the channel (i) or source indices (j) from the other indices by using a comma. The superscripts $\cdot^{\text{ex}}$ and $\cdot^{\text{ft}}$ indicate the excitation and filter NTF coefficients respectively. Finally, we define the operators that we use throughout the paper in the following

- $\mathbf{Diag}(\cdot)$ denotes a diagonal matrix.
- $\mathbf{Diag}(\cdot, \ldots, \cdot)$ denotes a block diagonal matrix .
- $\underline{\cdot}$ reshapes a given matrix into a column vector by concatenating and transposing its rows.
- $[\cdot]$ groups elements into a matrix.
- $\{\cdot\}$ groups elements into a tensor.
- $(\cdot)_j$ denotes the j^{th} element of a vector.
- $(\cdot)_{ij}$ denotes the $(i,j)^{\text{th}}$ element of a matrix.
- $[\cdot]^{\cdot e}$ denotes element-wise exponentiation with exponent e.
- $\odot$ denotes element-wise multiplication.
- $[\cdot]_{ii}$ denotes the i^{th} diagonal block of a matrix.

For the sake of readability, we will recall these definitions when they are used for the first time.

A. Generative Model

In the short time Fourier transform (STFT) domain, for J source signals and I channels, the mixing model is written as

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

where $\mathbf{x}_{fn} = [x_{i,fn}]^T$ is the $I \times 1$ vector of mixture STFT coefficients, $\mathbf{s}_{fn} = [s_{j,fn}]^T$ is the $J \times 1$ vector of source STFT coefficients, $\mathbf{A}_f = [\mathbf{A}_{j,f}]$ denotes the $I \times J$ complex-valued mixing matrix and $\boldsymbol{\epsilon}_{fn}$ represents sensor noise. In this formulation, f is the frequency index, n the time frame index, i the channel index, and j the source index. This mixing model underlies most source separation methods in the literature including, e.g., ICA. But we remind the reader that, unlike ICA, the proposed method can handle underdetermined mixtures ($J > I$). Source separation consists of estimating $\mathbf{A}_f$ and $\mathbf{s}_{fn}$ from $\mathbf{x}_{fn}$.

Each source signal $s_{j,fn}$ is assumed to follow a zero-mean complex-valued Gaussian distribution with time-varying variance $v_{j,fn}$ encoding its short-term power spectrum:

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

In the state of the art, the power spectrogram $\mathbf{V}_j = \{v_{j,fn}\}_{fn}$ of a given source j is decomposed by NMF into an excitation and an activation matrix: $\mathbf{V}_j \approx \mathbf{W}_j \mathbf{H}_j$. In this decomposition, the excitation matrix $\mathbf{W}_j$ represents the spectral shapes and the activation matrix $\mathbf{H}_j$ their weights over time. These two matrices are estimated by minimizing a cost function, typically using multiplicative update rules. These matrices can also be estimated within a probabilistic framework. Instead of this limited one-level decomposition, the short-term power spectra $v_{j,fn}$ are constrained via three-level NTF [11]. At the first level, they are assumed to follow an excitation-filter model

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

At the second level, the excitation spectral power $v^{\text{ex}}_{j,fn}$ is expressed as the sum of basis spectra scaled by time activation coefficients. Finally, at the third level, the basis spectra are defined as the sum of narrowband spectral patterns $w^{\text{ex}}_{j,fl}$ weighted by spectral envelope coefficients $u^{\text{ex}}_{j,lk}$. Similarly, the time activation coefficients are represented as the sum of time-localized patterns $h^{\text{ex}}_{j,mn}$ weighted by temporal envelope coefficients $g^{\text{ex}}_{j,km}$. The same decomposition applies to the filter spectral power $v^{\text{ft}}_{j,fn}$. Overall, the complete factorization scheme is as follows:

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

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

This framework makes it possible to exploit a wide range of prior information about the sources. For instance, harmonicity can be enforced by fixing $w^{\text{ex}}_{j,fl}$ as narrowband harmonic spectra and letting the spectral envelope and the active pitches be inferred from the data via $u^{\text{ex}}_{j,lk}$ and $g^{\text{ex}}_{j,km}$, respectively [11]. For a graphical illustration of the three-level NTF structure as well as more details and examples of possible spectral and temporal constraints, see [11].

B. Likelihood and Priors

Let us denote by $\mathbf{X} = \{\mathbf{x}_{fn}\}$, $\mathbf{S} = \{\mathbf{s}_{fn}\}$, $\mathbf{A} = \{\mathbf{A}_f\}$, and $\mathbf{V} = \{v_{j,fn}\}$ the sets of all mixture STFT coefficients, source STFT coefficients, mixing parameters, and short-term power spectra, respectively, and similarly by $\mathbf{W}^{\text{ex}}$, $\mathbf{U}^{\text{ex}}$, $\mathbf{G}^{\text{ex}}$, $\mathbf{H}^{\text{ex}}$, $\mathbf{W}^{\text{ft}}$, $\mathbf{U}^{\text{ft}}$, $\mathbf{G}^{\text{ft}}$, and $\mathbf{H}^{\text{ft}}$ the sets of all NTF coefficients. Assuming zero-mean Gaussian sensor noise with constant diagonal covariance $\boldsymbol{\epsilon}_{fn} \sim \mathcal{N}(0, \sigma_b^2\mathbf{I})$, the likelihood of the model is given by

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

with N and F the number of time frames and frequency bins, respectively.

All parameters are assumed to follow non-informative priors. More precisely, the NTF parameters $\mathbf{W}^{\text{ex}}$, $\mathbf{U}^{\text{ex}}$, $\mathbf{G}^{\text{ex}}$, $\mathbf{H}^{\text{ex}}$, $\mathbf{W}^{\text{ft}}$, $\mathbf{U}^{\text{ft}}$, $\mathbf{G}^{\text{ft}}$, and $\mathbf{H}^{\text{ft}}$ are assumed to follow the Jeffreys prior, e.g., $p(w^{\text{ex}}_{j,fl}) \propto 1/w^{\text{ex}}_{j,fl}$, and the mixing coefficients $\mathbf{A}$ are assumed to follow a flat prior $p(\mathbf{A}) \propto 1$.

C. Joint Distribution

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

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

The joint distribution $p(\mathbf{X}, \mathbf{Z})$ of the observations and the model parameters is given by

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

Fig. 2 shows the dependency graph of the joint distribution of the proposed approach. In this representation, each circle represents one random variable. The rectangles around the circles represent multiple nodes in a more compact way. The labels in those rectangles indicate the number of independent instances of this kind. The arrows show the dependencies. The direction of the arrow indicates the dependent random variable, which depends on the random variable at the other end of the arrow. The observation variable is shaded. The constant noise covariance parameter of the mixture is shown explicitly. The other prior distributions are non-informative and do not possess any parameters.

III. Variational Bayesian Inference

The original FASST algorithm in [11] achieved estimation of the model parameters in the ML sense using an expectation-maximization (EM) algorithm. We now consider the problem of full Bayesian inference, that is to estimate the posterior distribution of the model parameters $p(\mathbf{Z}|\mathbf{X})$. Exact estimation is intractable hence we resort to VB approximation [22], [51], which has been shown to converge more quickly than MCMC in a simple source separation setting [26]. The derivation of the proposed algorithm is detailed in the supporting technical report [50].

A. General Approach

VB inference aims to obtain an approximation $q(\mathbf{Z})$ of the true posterior $p(\mathbf{Z}|\mathbf{X})$ that minimizes the Kullback-Leibler (KL) divergence

$$KL(q||p) = -\int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} d\mathbf{Z}. \tag{9}$$

It can be shown that this quantity satisfies

$$\log p(\mathbf{X}) = \mathscr{L}(q) + KL(q||p), \tag{10}$$

where $p(\mathbf{X})$ is the marginal likelihood or *evidence* and $\mathscr{L}(q)$ is the *free energy* defined as

$$\mathscr{L}(q) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z}. \tag{11}$$

Minimization of the KL divergence is achieved by maximizing the free energy w.r.t. $q(\mathbf{Z})$.

In order to obtain a closed form solution, a factored form is typically assumed for $q(\mathbf{Z})$. As a starting point, we consider the mean field approximation, which consists of partitioning the set of parameters $\mathbf{Z}$ into subsets $\mathbf{Z}_\delta$ and of assuming that $q(\mathbf{Z}) = \prod_\delta q(\mathbf{Z}_\delta)$. Maximizing $\mathscr{L}(q)$ w.r.t. $q(\mathbf{Z}_\delta)$ while fixing the other factors yields the following update [51]:

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

where the normalizing constant is such that $q^*(\mathbf{Z}_\delta)$ integrates to 1. This leads to an iterative EM-like algorithm where all factors of the approximating distribution are alternately updated. In the E-step, sufficient statistics of the model parameters are computed. In the M-step, these statistics are used to update the parameters of the approximating distribution.

B. Auxiliary variables

In practice, (12) is useful only when $\mathbb{E}_{q(\mathbf{Z}_{\delta' \neq \delta})}[\log p(\mathbf{X},\mathbf{Z})]$ has closed form and it corresponds to a known parametric distribution for which the normalizing constant is computable in closed form. As we shall see, the considered model does not satisfy this condition and $\mathscr{L}(q)$ must be further lower bounded so that its maximum can be computed in closed form.

In our preliminary paper [49], we divided the source signals into a number of sub-component signals equal to the number of

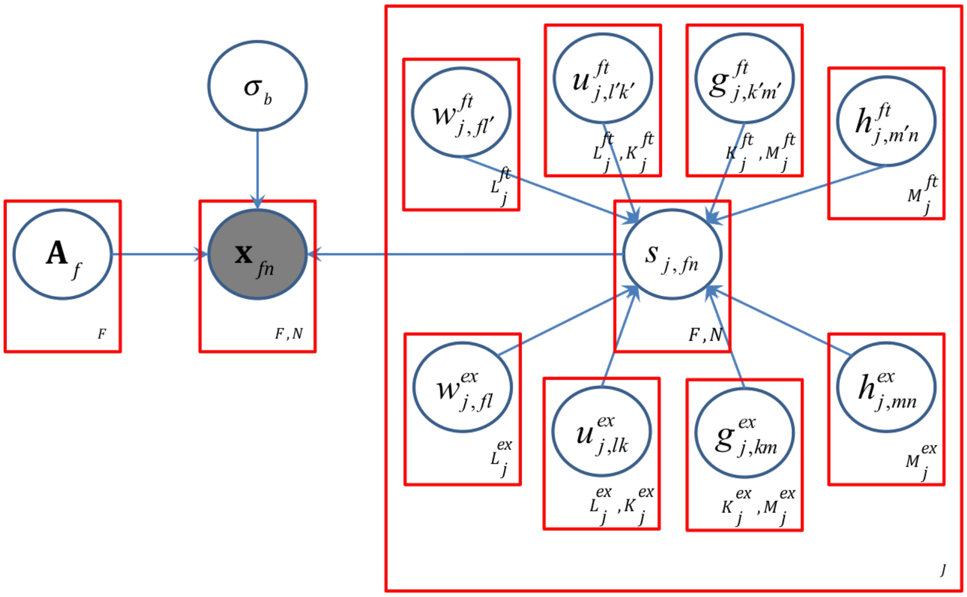

Fig. 2. Graphical model of the proposed Bayesian source separation approach.

NTF parameters and we applied (12) to estimate the posterior distribution of the sub-components. The large number of additional variables induced slow convergence and sensitivity to local maxima. In this paper, we derive a tighter variational approximation by directly lower bounding $\mathscr{L}(q)$.

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

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

where $\boldsymbol{\Omega}$ is a set of auxiliary variables. We define $\mathscr{B}$ which further lower bounds $\mathscr{L}$ as

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

Maximizing $\mathscr{B}(q, \boldsymbol{\Omega})$ w.r.t. $\boldsymbol{\Omega}$ and $q_{(}\mathbf{Z}_\delta)$ while fixing the other factors now yields the following update:

$$\log q^*(\mathbf{Z}_\delta) = \mathbb{E}_{q(\mathbf{Z}_{\delta' \neq \delta})}[\log f(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Omega})] + \text{const.} \tag{15}$$

In the following, we adopt this strategy for the lower bounding of the term $p(\mathbf{S}|\mathbf{V})$ in the expression of $p(\mathbf{X}, \mathbf{Z})$ in (8) whose expectation is not tractable in closed form.

C. Application to FASST

In this section, we summarize the algorithm and the update equations resulting from the application of the general VB approach above to the FASST framework in Section II. For details about the derivation, please refer to the Appendix. For the sake of readability, let us define $\eta = (k, m, l)$ the joint index of the excitation NTF parameters and $\eta' = (k', m', l')$ the joint index of the filter NTF parameters. With these joint indices let us further define $v^{\text{ex}}_{j,fn,\eta}$ as the product of excitation NTF parameters, $v^{\text{ft}}_{j,fn,\eta'}$ as the product of filter NTF parameters and $v_{j,fn,\eta,\eta'}$ as the product of all NTF parameters:

$$v^{\text{ex}}_{j,fn,\eta} = h^{\text{ex}}_{j,mn} g^{\text{ex}}_{j,km} u^{\text{ex}}_{j,lk} w^{\text{ex}}_{j,fl}, \tag{16}$$

$$v^{\text{ft}}_{j,fn,\eta'} = h^{\text{ft}}_{j,m'n} g^{\text{ft}}_{j,k'm'} u^{\text{ft}}_{j,l'k'} w^{\text{ft}}_{j,fl'}, \tag{17}$$

$$v_{j,fn,\eta,\eta'} = v^{\text{ex}}_{j,fn,\eta} v^{\text{ft}}_{j,fn,\eta'}. \tag{18}$$

After proper initialization, each iteration of the algorithm consists of the following steps:

1) compute the statistics of the NTF parameters, the source STFT coefficients and the mixing parameters as in (19)–(21), (22), (23), and (24),

$$\mathbb{E}[|s_{j,fn}|^2] = |(\boldsymbol{\mu}_{\mathbf{s},fn})_j|^2 + (\mathbf{R}_{\mathbf{ss},fn})_{jj} \tag{19}$$

where $\boldsymbol{\mu}_{\mathbf{s},fn}$ and $\mathbf{R}_{\mathbf{ss},fn}$ are the first and second order posterior moments of $\mathbf{s}_{fn}$. $\mathbb{E}[\mathbf{V}^{\text{ex}}_j]$ is obtained as

$$\mathbb{E}[\mathbf{V}^{\text{ex}}_j] = \mathbb{E}[\mathbf{W}^{\text{ex}}_j]\mathbb{E}[\mathbf{U}^{\text{ex}}_j]\mathbb{E}[\mathbf{G}^{\text{ex}}_j]\mathbb{E}[\mathbf{H}^{\text{ex}}_j] \tag{20}$$

and $\mathbb{E}[1/\mathbf{V}^{\text{ft}}_j]^{.-1}$ is a shorthand notation for

$$\mathbb{E}\Big[\frac{1}{\mathbf{W}^{\text{ft}}_j}\Big]^{.-1} \mathbb{E}\Big[\frac{1}{\mathbf{U}^{\text{ft}}_j}\Big]^{.-1} \mathbb{E}\Big[\frac{1}{\mathbf{G}^{\text{ft}}_j}\Big]^{.-1} \mathbb{E}\Big[\frac{1}{\mathbf{H}^{\text{ft}}_j}\Big]^{.-1}. \tag{21}$$

The second order raw moment $\mathbf{R}_{\mathbf{s},fn}$ of $\mathbf{s}, fn$ is given by

$$\mathbf{R}_{\mathbf{s},fn} = \boldsymbol{\mu}_{\mathbf{s},fn} \boldsymbol{\mu}^H_{\mathbf{s},fn} + \mathbf{R}_{\mathbf{ss},fn} \tag{22}$$

and $\mathbf{R}_{\mathbf{xs},fn}$ is the cross-moment given by

$$\mathbf{R}_{\mathbf{xs},fn} = [x_{1,fn}\boldsymbol{\mu}^H_{\mathbf{s},fn}, \ldots, x_{I,fn}\boldsymbol{\mu}^H_{\mathbf{s},fn}]^T. \tag{23}$$

The second order raw moment $\mathbf{R}_{\mathbf{A},f}$ of $\mathbf{A}_f$ is written as follows:

$$\mathbf{R}_{\mathbf{A},f} = \sum_i ([\boldsymbol{\mu}_{\underline{\mathbf{A}},f} \boldsymbol{\mu}^H_{\underline{\mathbf{A}},f} + \mathbf{R}_{\underline{\mathbf{AA}},f}]_{ii})^T. \tag{24}$$

where $\boldsymbol{\mu}_{\underline{\mathbf{A}},f}$ and $\mathbf{R}_{\underline{\mathbf{AA}},f}$ are the first and second moments of $\underline{\mathbf{A}}_f$ obtained by concatenating and transposing the rows of $\mathbf{A}_f$. $[\cdot]_{ii}$ denotes the diagonal $J \times J$ block corresponding to channel i.

2) update the GIG distributions of the NTF parameters according to (25)–(27),

$$\boldsymbol{\rho}^{\mathrm{ex}}_{w,j} = \mathbb{E}[\mathbf{V}^{\mathrm{ex}}_j]^{.-1}\big(\mathbb{E}[\mathbf{U}^{\mathrm{ex}}_j]\mathbb{E}[\mathbf{G}^{\mathrm{ex}}_j]\mathbb{E}[\mathbf{H}^{\mathrm{ex}}_j]\big)^T \tag{25}$$

$$\begin{aligned}\boldsymbol{\tau}^{\mathrm{ex}}_{w,j} &= \mathbb{E}\Big[\frac{1}{\mathbf{W}^{\mathrm{ex}}_j}\Big]^{.-2} \\ &\quad \odot \left(\Big(\mathbb{E}[|\mathbf{S}_j|^{.2}] \odot \mathbf{C}_j^{.-2} \odot \mathbb{E}\Big[\frac{1}{\mathbf{V}^{\mathrm{ft}}_j}\Big]^{.-1}\Big)\right. \\ &\qquad \left.\Big(\mathbb{E}\Big[\frac{1}{\mathbf{U}^{\mathrm{ex}}_j}\Big]^{.-1}\mathbb{E}\Big[\frac{1}{\mathbf{G}^{\mathrm{ex}}_j}\Big]^{.-1}\mathbb{E}\Big[\frac{1}{\mathbf{H}^{\mathrm{ex}}_j}\Big]^{.-1}\Big)^T\right)\end{aligned} \tag{26}$$

$$\gamma^{\mathrm{ex}}_{w,j,fl} = 0 \tag{27}$$

where $[\cdot]^T$ denotes matrix transposition, $\odot$ denotes element-wise multiplication, $[\cdot]^{.\alpha}$ denotes element-wise exponentiation, complex norm and division are also computed element-wise and $\mathbf{C}_j$ is defined in (42).

3) update the complex-valued Gaussian distribution of the source STFT coefficients according to (28) and (29),

$$\mathbf{R}_{\mathbf{ss},fn} = \Big({\mathbf{C}_{fn}}^{-1} + \frac{1}{\sigma_b^2}\mathbf{R}_{\mathbf{A},f}\Big)^{-1}. \tag{28}$$

Similarly, the mean is obtained by identifying the linear term in $\mathbf{s}^H_{fn}$ with $\mathbf{s}^H_{fn}\mathbf{R}^{-1}_{\mathbf{ss},fn}\boldsymbol{\mu}_{\mathbf{s},fn}$:

$$\boldsymbol{\mu}_{\mathbf{s},fn} = \frac{1}{\sigma_b^2}\mathbf{R}_{\mathbf{ss},fn}\boldsymbol{\mu}^H_{\mathbf{A},f}\mathbf{x}_{fn}. \tag{29}$$

4) update the complex-valued Gaussian distribution of the mixing parameters according to (30) and (31),

$$\mathbf{R}_{\underline{\mathbf{A}}\underline{\mathbf{A}},f} = \Big(\frac{1}{\sigma_b^2}\sum_n \mathbf{Diag}\underbrace{(\mathbf{R}^T_{\mathbf{s},fn},\ldots,\mathbf{R}^T_{\mathbf{s},fn})}_{\text{I times}}\Big)^{-1} \tag{30}$$

$$\boldsymbol{\mu}_{\underline{\mathbf{A}},f} = \frac{1}{\sigma_b^2}\mathbf{R}_{\underline{\mathbf{A}}\underline{\mathbf{A}},f}\sum_n \mathbf{R}_{\mathbf{xs},fn}. \tag{31}$$

5) compute the lower bound given in (14) for monitoring the convergence [51]. This bound is the sum of 11 terms: $\mathbb{E}[\log p(\mathbf{X}|\mathbf{S},\mathbf{A})]$, $\mathbb{E}[\log \underline{p}(\mathbf{S}|\mathbf{V}) - \log q(\mathbf{S})]$, $\mathbb{E}[\log p(\mathbf{A}) - \log q(\mathbf{A})]$, and 8 terms of the form $\mathbb{E}[\log p(\mathbf{W}^{\mathrm{ex}}) - \log q(\mathbf{W}^{\mathrm{ex}})]$ for each of the 8 sets of NTF parameters, where $\underline{p}(\mathbf{S}|\mathbf{V})$ is the lower bound to $p(\mathbf{S}|\mathbf{V})$. See Appendix for details.

Similarly to the original ML algorithm, with two or more channels and for typical signal duration, the computational cost is linear in the number of time-frequency bins NF and it is dominated by the multichannel Wiener filter in (28) and (29), which must be computed in each time-frequency bin. Since $\mathbf{R}_{\mathbf{A},f}$ is Hermitian, the expression of this filter simplifies via the Woodbury identity to the inversion of an $I \times I$ matrix. The cost of these updates being similar to the that of the E-step of the original ML algorithm, the overall cost of one iteration of the two algorithms is on the same order. With our current Matlab implementation, the processing of one mixture takes on the order of 50 times real time. We recently developed a real time capable C++ implementation for ML-FASST [15], but did not extend it to VB-FASST yet.

IV. Uncertainty Propagation

After source separation, we wish to estimate the uncertainty about the separated source STFT coefficients and to propagate it to the features considered for classification.

A. ML vs. VB Uncertainty Estimation

The most principled uncertainty estimator proposed so far is perhaps the Wiener estimator in [48]. This estimator considers the posterior distribution of the source STFT coefficients $p(\mathbf{S}|\mathbf{X},\widehat{\mathbf{A}},\widehat{\mathbf{V}})$ given the mixture STFT coefficients $\mathbf{X}$ and ML estimates $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{V}}$ of the mixing parameters and the source short-term power spectra. This distribution has Gaussian form and its mean and variance are given by the Wiener filter.

By considering ML point estimates of $\mathbf{A}$ and $\mathbf{V}$ instead of integrating over them, this estimator intrinsically underestimates uncertainty. We argue that the ideal uncertainty estimator in the Bayesian sense is given by

$$p(\mathbf{S}|\mathbf{X}) = \int p(\mathbf{Z}|\mathbf{X})d\mathbf{Z}' \tag{32}$$

with $\mathbf{Z}' = \mathbf{Z}\backslash\{\mathbf{S}\}$ denoting the set of all model parameters but $\mathbf{S}$. VB inference allows us to approximate this ideal estimator by the mean field estimator as discussed in the previous sections.

B. Uncertainty Propagation to the Spatial Source Images

Due to the phase and scale indeterminacies of source separation, we do not use the source estimates $\mathbf{s}_{j,fn}$ themselves for classification but the *spatial source images* $\mathbf{y}_{j,fn} = \mathbf{A}_{j,f}s_{j,fn}$ instead, which do not suffer from such indeterminacies [8]. The first and second order moments of the sources given in (28), (29) are propagated to the source images as follows:

$$\boldsymbol{\mu}_{\mathbf{y},j,fn} = \boldsymbol{\mu}_{\mathbf{A},j,f}\,\boldsymbol{\mu}_{\mathbf{s},j,fn}, \tag{33}$$

$$\begin{aligned}(\boldsymbol{R}_{\mathbf{yy},j,fn})_{ii'} &= \sum_{jj'}(\mathbf{R}_{\mathbf{A},f})_{(ij,i'j')}(\mathbf{R}_{\mathbf{s},fn})_{(jj')} \\ &\quad - (\boldsymbol{\mu}_{\mathbf{y},j,fn}\boldsymbol{\mu}^H_{\mathbf{y},j,fn})_{ii'}.\end{aligned} \tag{34}$$

C. Uncertainty Propagation to the Features

As an example feature, we then propagate the uncertainty to the Mel frequency cepstral coefficients (MFCC) $\mathrm{MFCC}_{j,n}$ of the first channel $y_{1j,fn}$ of $\mathbf{y}_{j,fn}$. Denoting by $\mathbf{y}_{1j,n} = [y_{1j,fn}]$ the complex-valued spectrum in time frame n, the MFCCs are defined as $\mathrm{MFCC}_{j,n} = \mathbf{D}\log(\mathbf{M}|\mathbf{y}_{1j,n}|)$ where $\mathbf{D}$ is the discrete cosine transform (DCT) matrix, $\mathbf{M}$ is the matrix of Mel filter coefficients, and the complex norm and the logarithm are computed element-wise. We propose two uncertainty propagation techniques for this.

1) Moment Matching: In the moment matching approach, the uncertainty expressed by the means and variances of the posterior distributions of the estimated source images is propagated through the computation of the MFCCs.

The MFCC computation involves two nonlinearities: the computation of the magnitude spectrum and the logarithm of the Mel filter bank output. As the estimates of the source images $\mathbf{y}_{1j,n}$ are complex-valued Gaussian, the magnitude spectrum (i.e., the absolute value of the source images) follows a Rice distribution [38]. For the second non-linearity, we assume the log-normality of the Mel features and use the log-normal transform given in [46]. The Mel transform and the DCT are both linear transforms. Hence linear transformation rules apply for these two steps. For more details about the moment matching approach, please refer to [26].

2) Vector Taylor Series: VTS consists of linearizing the MFCC transform by its first-order Taylor series expansion [43]. The mean of the features are computed as given in the MFCC computation. The Taylor series expansion is then used for propagating the uncertainty through this linear transform from the covariance of the source images to the covariance of the features.

V. Experimental Evaluation

We now evaluate the impact of the proposed VB algorithm for FASST compared to the original ML estimation algorithm in [11] on the three successive steps depicted in Fig. 1: source separation, feature extraction, and classification/recognition. In these experiments, we do not use the NTF filter components $\mathbf{W}^{\mathrm{ft}}$, $\mathbf{U}^{\mathrm{ft}}$, $\mathbf{G}^{\mathrm{ft}}$, and $\mathbf{H}^{\mathrm{ft}}$ by initializing and fixing them to identity matrices.

A. Source Separation

1) Data and Algorithmic Settings: For the evaluation of source separation, we consider 8 alternative configurations of FASST tested in [11], where certain parameters are either *constrained* or estimated from the mixture in an *unconstrained* fashion. These configurations consist of all combinations of the following choices:

- Rank: each source is modeled either as (1) a rank-1 point source or as (2) a full-rank source,
- Spectral structure: the narrowband spectral patterns $\mathbf{W}^{\mathrm{ex}}$ are either (un) estimated from the mixture or (co) fixed to harmonic and noise-like patterns as in [11],
- Temporal structure: the time-localized patterns $\mathbf{H}^{\mathrm{ex}}$ are either (un) estimated from the mixture or (co) fixed to decreasing exponential patterns as in [11].

An STFT window size of 2048 samples is used in all cases. Furthermore, the mixing parameters $\mathbf{A}$ are initialized using the source directions of arrival (DOA) estimated via the algorithm in [52] and estimated from the mixture. Similarly the spectral envelopes $\mathbf{U}^{\mathrm{ex}}$ and the temporal envelopes $\mathbf{G}^{\mathrm{ex}}$ are initialized randomly and estimated from the mixture. The noise variance σ_b^2 is initialized to 10^{-2} and gradually decreased to 10^{-6} using the same annealing values as in [11]. Both algorithms are run for 200 iterations. Preliminary experiments showed that this was enough to achieve convergence with the proposed VB algorithm and that the obtained value of the variational bound was higher than in [49].

The algorithms are evaluated on the same dataset as in [11], namely the development dataset of the 2010 Signal Separation Evaluation Campaign (SiSEC) [53][1]. This dataset contains both synthetic and recorded two-channel mixtures with reverberation times of 130 and 250 ms. There are 32 mixtures of 3 sources and 24 mixtures of 4 sources of 10 s duration each. Separation performance is evaluated in terms of the Signal-to-Distortion Ratio (SDR) [54] in decibels (dB) between the true source images $\mathbf{y}_{j,fn}$ and the estimated source images (in the case of ML) or their posterior mean $\boldsymbol{\mu}_{\mathbf{y},j,fn}$ (in the case of VB) and it is averaged over all sources and all mixtures.

2) Results: The results are shown in Table I. Our VB estimation algorithm provides a modest but consistent SDR improvement on the order of 0.1 dB compared to the state-of-the-art ML algorithm. The best results are achieved in the configuration where both the spectral structure and the temporal structure are constrained (1-co-co). This illustrates the benefit of using flexible algorithms such as FASST which are able to jointly exploit spatial and spectral models for separation. In this configuration, VB achieves a slightly larger improvement of 0.3 dB on average compared to ML. For comparison, the baseline binary masking method [52] yields 0.95 dB SDR.

B. Feature Extraction

1) Data and Algorithmic Settings: Feature extraction is evaluated using the same data and algorithmic settings as above. We compute the 0th to the 19th MFCCs and we contrast the moment matching uncertainty propagation technique in Section IV with deterministic MFCC computation from the separated source images (in the case of ML) or from their posterior mean (in the case of VB). The input uncertainty is estimated using either the proposed VB uncertainty estimator or the ML-based Wiener uncertainty estimator in [48]. Accuracy is measured by the root mean square (RMS) error between the true MFCCs $\mathrm{MFCC}_{j,n}$ and the estimated MFCCs (in the case of deterministic feature extraction) or their posterior mean $\boldsymbol{\mu}_{\mathrm{MFCC},j,n}$ (in the case of uncertainty propagation).

2) Results: As one can see from Table II, VB source separation and uncertainty propagation both reduce the RMS error. The best results are achieved with the proposed VB-based uncertainty propagation technique, which provides a modest improvement over ML-based uncertainty propagation and 10% relative RMS error reduction compared to ML source separation and deterministic feature extraction on average over all configurations. For comparison, the baseline binary masking method in [52] yields a significantly worse RMS error of 1.99.

C. Classification

1) Data and Algorithmic Settings: As an example classification task, we consider the noise-robust speaker identification

[1] http://sisec2010.wiki.irisa.fr/tiki-index.php?page=Underdetermined-+speech+and+music+mixtures

	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 [11]	1.58	1.68	1.87	2.07	1.77	1.75	2.28	2.25
VB	**1.70**	**1.78**	**1.95**	**2.10**	**1.92**	**1.85**	**2.54**	**2.35**

TABLE I

AVERAGE SDR (DB) ACHIEVED BY ML OR VB SOURCE SEPARATION.

	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	up	det	up	det	up	det	up	det	up	det	up	det	up	det	up
ML [11]	1.19	1.18	1.78	**1.47**	1.47	1.45	1.55	1.46	1.53	1.45	1.73	**1.46**	1.56	1.49	1.75	**1.49**
VB	1.16	**1.14**	1.73	**1.47**	1.46	**1.44**	1.52	**1.45**	1.54	**1.44**	1.69	**1.46**	1.58	**1.47**	1.74	**1.49**

TABLE II

AVERAGE RMS ERROR OVER THE MFCCS OBTAINED BY ML OR VB SOURCE SEPARATION FOLLOWED BY (DET) DETERMINISTIC FEATURE EXTRACTION OR (UP) UNCERTAINTY PROPAGATION BY MOMENT MATCHING.

benchmark in [43]. This benchmark consists of noiseless reverberated utterances and real domestic noise backgrounds from the 2nd CHiME Speech Separation and Recognition Challenge [55] which are mixed together at six different signal-to-noise ratios (SNR) from -6 to 9 dB. The training set consists of 680 noiseless reverberated utterances from 34 speakers and the test set consists of 680 other utterances from the same speakers for each SNR condition.

An STFT window size of 1024 samples is used in all cases. The target speech source is modeled by a 256-component speaker-independent rank-1 multichannel NMF model. The posterior distributions of the mixing parameters $\mathbf{A}$ and the NTF parameters $\mathbf{W}^{\text{ex}}$ are estimated by running 100 iterations of the proposed VB algorithm with all the other NTF components are fixed to identity on 15 clean speech samples randomly selected for each speaker from the training set. Similarly, background noise is assumed to be the sum of 4 sources, each of which is modeled by an 8-component rank-1 multichannel NMF, which is randomly initialized and trained by running 30 iterations of the proposed VB algorithm on background noise surrounding the test utterance (10 s before and 10 s after). This approach is known to outperform training on a larger noise dataset [56]. Finally, each test utterance is separated by keeping the NTF parameters $\mathbf{W}^{\text{ex}}$ for the target speaker source fixed and reestimating the other NTF parameters and the mixing parameters by running 50 iterations of the proposed VB algorithm.

The 1st to the 19th MFCCs are extracted from the estimated target speech signal using either deterministic computation or uncertainty propagation using the moment matching technique. Indeed, we found moment matching to outperform VTS in this experiment. Finally, speaker identification is achieved using 32-component Gaussian mixture models (GMM) trained on the noiseless reverberated utterances of the training set. As argued in [43], this benchmark was designed to assess the accuracy of the estimated uncertainties alone without interference from more complex classifiers involved in state-of-the-art speaker identification systems. For more details about the data, the configuration of FASST, and the classifier, please refer to [43].

2) Results: The average speaker identification accuracy is shown in Table III. With deterministic feature extraction, ML and VB source separation both degrade performance compared to the baseline performance obtained without source separation. This is due to the distortion of the target speech signal introduced by the source separation algorithm not being taken into account by the classifier. With uncertainty propagation, however, both ML and VB improve over the baseline. The best results are achieved by the proposed VB-based uncertainty propagation at all SNRs. On average, it achieves a significant relative error reduction of 6% compared to ML uncertainty estimation, 43% compared to deterministic processing, and 27% compared to the baseline. This improvement is meaningful. For instance, [44] reported 7% to 9% relative error rate improvement with their uncertainty propagation method compared to deterministic processing.

D. Speech Recognition

1) Data and Algorithmic Settings: As an example recognition task, we consider noise-robust speech recognition in the framework of the 2nd CHiME Speech Separation and Recognition Challenge [55]. The utterances consist of 6 words of the form <command>, <color>, <preposition>, <letter>, <digit>, <adverb> and are mixed with real domestic backgrounds at six different signal-to-noise ratios (SNR) from -6 to 9 dB. The training set consists of 17000 noiseless reverberated utterances from 34 speakers and the development and test sets consist of 600 other utterances from the same speakers for each SNR condition. The task consists of recognizing the letter and digit keywords.

The target speech source was modeled by a 32-component rank-1 multichannel NMF model individually for each speaker. The posterior distributions of the mixing parameters and the NTF parameters $\mathbf{W}^{\text{ex}}$ are estimated for each speaker by running 100 iterations of the proposed VB algorithm on a subset of the training set comprising 5 randomly selected utterances for each SNR condition (that is, 30 utterances in total), with all the other NTF components fixed to identity. Similarly, background noise is assumed to be the sum of 2 sources, each of which is modeled by a 16-component rank-1 multichannel NMF, which is randomly initialized and trained by running 50 iterations of the proposed VB algorithm on background noise surrounding the test utterance (10 s before and 10 s after). Finally, each test utterance is separated by keeping the NTF parameters $\mathbf{W}^{\text{ex}}$ of the target speech source fixed and reestimating the above NTF parameters and the mixing parameters by running 250 iterations of the proposed VB algorithm.

	-6 dB		-3 dB		0 dB		3 dB		6 dB		9 dB	
	det	up	det	up	det	up	det	up	det	up	det	up
ML [11]	33.53	44.85	34.26	52.35	47.50	70.29	60.74	82.65	72.06	91.32	83.24	95.59
VB	28.68	**46.18**	31.76	**53.24**	42.06	**72.65**	58.97	**84.85**	72.94	**94.26**	80.15	**96.18**
baseline	40.44		41.32		58.09		74.12		84.71		92.35	

TABLE III
Average speaker identification accuracy (%) obtained by ML or VB source separation followed by (det) deterministic feature extraction or (up) uncertainty propagation by moment matching. The baseline performance is the one obtained by computing the deterministic MFCCs on the mixtures without source separation.

Due to phoneme durations, shorter window sizes are preferred in automatic speech recognition experiments. State-of-the-art speech recognizers e.g. [57] use a typical window size of 25 ms. However in the case of source separation, larger window sizes are preferable. In the source separation and speaker identification experiments, we used 2048 sample and 1024 sample STFT windows respectively. Therefore, in the speech recognition experiments, we used a hybrid approach to overcome this inconsistency. We performed two source separation steps: one with a large 1024 sample (64 ms) window and one with a short 384 sample (24 ms) window. During the computation of the MFCCs using the uncertainty propagation, we take the mean estimates of the sources from the source separation performed using the large window, whereas the covariances are taken from the short window source separation experiments. This heuristic approach provides fairly good estimates for both the mean and the covariance of the source coefficients.

The first 13 MFCCs were extracted from the estimated target speech signal using either deterministic computation or VTS. Indeed, we found VTS to outperform moment matching in this experiment. In the next step, these uncertainties were propagated to the first and second derivatives of the MFCCs [58]. In this approach, the estimated covariance matrix is multiplied by a scaling matrix which is estimated using the development data set. For more details about this approach, please refer to [58]. Finally, speech recognition is achieved using the GMM-HMM HTK baseline provided by the challenge organizers [55] trained on the noiseless reverberated utterances of the training set. Uncertainty decoding was performed using the full uncertainty covariance decoding approach proposed in [58]. For more details about this procedure, please refer to [58].

2) Results: The average speech recognition accuracy is shown in Table IV. With deterministic feature extraction, ML and VB source separation both provide an increase in the speech recognition accuracy in low SNR conditions up to 3 dB. This performance gain gradually vanishes with the increasing SNR: the baseline (without source separation) performs better than ML-based source separation at 6 dB SNR and better than both source separation algorithms at 9 dB SNR. As in the case of speaker identification experiments, the reason for this degradation is the distortion of the target speech signal introduced by the source separation algorithms not being taken into account by the recognizer. With uncertainty propagation, however, both ML and VB improve over the baseline. The best results are achieved by the proposed VB-based source separation algorithm at all SNRs except 0 dB. On average, it achieves a significant 5% relative keyword error rate reduction compared to ML uncertainty estimation and 49% relative keyword error rate reduction compared to the baseline.

Overall, these experimental results indicate that the main benefit of VB vs. ML source separation is not to be found in the estimated sources or features, which are marginally better, but in the estimated uncertainties, which result in significantly better classification and recognition. Indeed, as explained in Section IV-A, VB approximates the ideal Bayesian uncertainty estimator by integrating over the model parameters rather than considering a point estimate.

VI. Conclusion

In this paper, we presented a general, fully Bayesian audio source separation algorithm based on VB inference. This algorithm relies on the flexible FASST modeling framework, which jointly accounts for spatial parameters and NTF-like spectral structure of the sources, and on a tight mean field approximation. Experimental results indicate that it provides a modest but consistent improvement of source separation and feature extraction accuracy compared to conventional ML separation. More importantly, the posterior variance of the source STFT coefficients provides a mathematically rigorous estimate of uncertainty which, after propagation to the features, results in significantly better classification and recognition accuracy than the Wiener uncertainty estimator in [48]. This fundamental finding opens the way for further improvements in uncertainty estimation by seeking more complex approximations of the ideal Bayesian uncertainty estimator in (32) based on structured VB [59] for instance.

Beyond the FASST model, the idea of accounting for the uncertainty about the parameter estimates themselves may be applicable to other source separation techniques based on, e.g., deep neural networks (DNNs) [60]–[62]. Although VB does not readily apply to discriminative models such as DNNs, the recent merging of DNNs and spatial covariance models into an EM-like algorithm in [63] opens the way for VB-like extensions. In the longer term, our work is also expected to have some impact on the performance of DNN-based classifiers in noisy conditions. Uncertainty propagation in DNNs has recently started being investigated with promising early results [42], but it still remains an open question.

		-6 dB		-3 dB		0 dB		3 dB		6 dB		9 dB	
		det	up	det	up	det	up	det	up	det	up	det	up
devel	ML [11]	47.00	62.50	55.75	71.58	65.42	**77.83**	72.92	83.58	78.33	87.83	80.75	89.33
	VB	50.75	**64.75**	58.67	**72.33**	67.33	77.50	78.00	**85.50**	81.58	**88.58**	85.50	**91.00**
baseline		38.17		46.25		57.42		69.33		78.67		86.00	
test	ML [11]	48.25	65.42	55.17	71.00	66.08	**81.17**	71.67	85.25	79.50	88.25	82.67	91.33
	VB	51.08	**66.92**	57.25	**72.00**	68.33	80.25	76.83	**87.17**	83.42	**90.00**	87.42	**92.42**
baseline		38.92		46.08		60.50		70.42		81.08		86.50	

TABLE IV

AVERAGE KEYWORD RECOGNITION ACCURACY (%) OBTAINED BY ML OR VB SOURCE SEPARATION FOLLOWED BY (DET) DETERMINISTIC FEATURE EXTRACTION OR (UP) UNCERTAINTY PROPAGATION USING VTS. THE BASELINE PERFORMANCE IS THE ONE OBTAINED BY COMPUTING THE DETERMINISTIC MFCCS ON THE MIXTURES WITHOUT SOURCE SEPARATION.

APPENDIX
DERIVATION OF THE ALGORITHM

In the context of FASST, we consider the following mean field factorization:

$$q(\mathbf{Z}) = \prod_{fn} q(\mathbf{s}_{fn}) \prod_{f} q(\mathbf{A}_f) \prod_{j,fl} q(w^{\text{ex}}_{j,fl}) \prod_{j,lk} q(u^{\text{ex}}_{j,lk}) \prod_{j,km} q(g^{\text{ex}}_{j,km}) \prod_{j,mn} q(h^{\text{ex}}_{j,mn}) \prod_{j,fl'} q(w^{\text{ft}}_{j,fl'}) \prod_{j,l'k'} q(u^{\text{ft}}_{j,l'k'}) \prod_{j,k'm'} q(g^{\text{ft}}_{j,k'm'}) \prod_{j,m'n} q(h^{\text{ft}}_{j,m'n}). \tag{35}$$

We retain the posterior dependencies between the sources in each time-frequency bin and between the channels in each frequency bin and we factor over the other dimensions.

Let us look at $\mathbb{E}[\log p(\mathbf{S}|\mathbf{V})]$ more closely:

$$\mathbb{E}[\log p(\mathbf{S}|\mathbf{V})] = \sum_{fn} -J \log \pi - \sum_{j} \mathbb{E}\Big[\log \sum_{\eta}\sum_{\eta'} v_{j,fn,\eta,\eta'}\Big] - \sum_{j} \mathbb{E}[|s_{j,fn}|^2]\,\mathbb{E}\Big[\frac{1}{\sum_{\eta}\sum_{\eta'} v_{j,fn,\eta,\eta'}}\Big]. \tag{36}$$

None of the two expectations containing $v_{j,fn,\eta,\eta'}$ in (36) is tractable. Hence we lower bound $\log p(\mathbf{S}|\mathbf{V})$ as explained above by generalizing the bound proposed in [31] for single-channel NMF to the context of FASST. Given that $x \to -\log x$ is convex, the argument of the first expectation is lower bounded by its first-order Taylor series expansion around an arbitrary positive point $\omega_{j,fn}$:

$$-\log \sum_{\eta}\sum_{\eta'} v_{j,fn,\eta,\eta'} \geq -\log \omega_{j,fn} + 1 - \frac{1}{\omega_{j,fn}} \sum_{\eta}\sum_{\eta'} v_{j,fn,\eta,\eta'}. \tag{37}$$

Given that $x \to -1/x$ concave, the argument of the second expectation can be lower bounded using Jensen inequality. For any nonnegative $\phi_{j,fn,\eta,\eta'}$ such that $\sum_{\eta}\sum_{\eta'} \phi_{j,fn,\eta,\eta'} = 1$ we have

$$-\frac{1}{\sum_{\eta}\sum_{\eta'} v_{j,fn,\eta,\eta'}} \geq -\sum_{\eta}\sum_{\eta'} \phi^2_{j,fn,\eta,\eta'} \frac{1}{v_{j,fn,\eta,\eta'}}. \tag{38}$$

With these two inequalities, we can lower bound $\log p(\mathbf{S}|\mathbf{V})$ as follows

$$\log p(\mathbf{S}|\mathbf{V}) \geq -FNJ\log\pi + \sum_{j,fn}\Big(-\log \omega_{j,fn} + 1 - \frac{1}{\omega_{j,fn}} \sum_{\eta}\sum_{\eta'} v_{j,fn,\eta,\eta'}\Big) - \sum_{j,fn} |s_{j,fn}|^2 \sum_{\eta}\sum_{\eta'} \phi^2_{j,fn,\eta,\eta'} \frac{1}{v_{j,fn,\eta,\eta'}}. \tag{39}$$

Having this lower bound and the auxiliary variables $\boldsymbol{\Omega} = \{\{\omega_{j,fn}\},\{\phi_{j,fn,\eta,\eta'}\}\}$, we can now derive the updates for $\boldsymbol{\Omega}$ and for each of the terms in (35).

A. Tightening the Bound w.r.t. the Auxiliary Variables

The updates for $\boldsymbol{\Omega}$ are obtained by maximizing the bound $\mathscr{B}(q,\boldsymbol{\Omega})$. For $\omega_{j,fn}$, we simply compute its partial derivative w.r.t. $\omega_{j,fn}$ and make it equal to zero, which yields

$$\omega_{j,fn} = \sum_{\eta}\sum_{\eta'} \mathbb{E}[v_{j,fn,\eta,\eta'}]. \tag{40}$$

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

$$\phi_{j,fn,\eta,\eta'} = \frac{1}{C_{j,fn}} \mathbb{E}\Big[\frac{1}{v_{j,fn,\eta,\eta'}}\Big]^{-1} \tag{41}$$

where $C_{j,fn}$ is the normalizing factor given by

$$C_{j,fn} = \sum_{\eta}\sum_{\eta'} \mathbb{E}\Big[\frac{1}{v_{j,fn,\eta,\eta'}}\Big]^{-1}. \tag{42}$$

Fast summation over (η,η') is achieved by expressing these updates via the shorthand matrix notation $\boldsymbol{\omega}_j = \mathbb{E}[\mathbf{V}_j]$ and $\mathbf{C}_j = \mathbb{E}[1/\mathbf{V}_j]^{\cdot -1}$ where $\boldsymbol{\omega}_j = [\omega_{j,fn}]$, $\mathbf{C}_j = [C_{j,fn}]$, $\mathbf{V}_j = [v_{j,fn}]$, and $\mathbb{E}[\mathbf{V}_j]$ and $\mathbb{E}[1/\mathbf{V}_j]^{\cdot -1}$ are computed by matrix multiplication as explained at the end of Appendix B.

B. Variational Updates for the NTF Parameters

For the scalar NTF parameter $w^{\text{ex}}_{j,fl}$, the update (15) is expressed as

$$\begin{aligned}
&\log q^*(w^{\text{ex}}_{j,fl}) = \\
&w^{\text{ex}}_{j,fl}\Big(\sum_n -\frac{1}{\omega_{j,fn}}\sum_{km}\sum_{\eta'}\mathbb{E}[h^{\text{ex}}_{j,mn}g^{\text{ex}}_{j,km}u^{\text{ex}}_{j,lk}v^{\text{ft}}_{j,fn,\eta'}]\Big) \\
&\qquad + \frac{1}{w^{\text{ex}}_{j,fl}}\Big(\sum_n -\mathbb{E}[|s_{j,fn}|^2] \\
&\quad \sum_{k,m}\sum_{\eta'}\phi^2_{j,fn,\eta,\eta'}\mathbb{E}\Big[\frac{1}{h^{\text{ex}}_{j,mn}g^{\text{ex}}_{j,km}u^{\text{ex}}_{j,lk}v^{\text{ft}}_{j,fn,\eta'}}\Big]\Big) \\
&\qquad\qquad - \log w^{\text{ex}}_{j,fl} + \text{const}. \qquad (43)
\end{aligned}$$

This distribution involves a linear term in $w^{\text{ex}}_{j,fl}$, a linear term in $1/w^{\text{ex}}_{j,fl}$, and a linear term in $\log w^{\text{ex}}_{j,fl}$. This is an instance of the generalized inverse Gaussian (GIG) distribution [64], whose probability density function (PDF) is given by

$$GIG(y;\gamma,\rho,\tau) = \frac{\exp\{(\gamma-1)\log y - \rho y - \frac{\tau}{y}\}\rho^{\frac{\gamma}{2}}}{2\tau^{\frac{\gamma}{2}}K_\gamma(2\sqrt{\rho\tau})} \qquad (44)$$

for $y \geq 0$, $\rho \geq 0$ and $\tau \geq 0$, where $K_\gamma(\cdot)$ is the modified Bessel function of the second kind. In other words

$$q^*(w^{\text{ex}}_{j,fl}) = GIG(w^{\text{ex}}_{j,fl};\gamma^{\text{ex}}_{w,j,fl},\rho^{\text{ex}}_{w,j,fl},\tau^{\text{ex}}_{w,j,fl}) \qquad (45)$$

where $\rho^{\text{ex}}_{w,j,fl}$, $\tau^{\text{ex}}_{w,j,fl}$, and $\gamma^{\text{ex}}_{w,j,fl}$, are obtained as the coefficients of the terms in $w^{\text{ex}}_{j,fl}$, $1/w^{\text{ex}}_{j,fl}$, and $\log w^{\text{ex}}_{j,fl}$ in (43). After replacing (40) and (41) into (43) and grouping the parameters into matrices $\mathbf{S}_j = [s_{j,fn}]$, $\mathbf{V}^{\text{ex}}_j = [v^{\text{ex}}_{j,fn}]$, $\mathbf{V}^{\text{ft}}_j = [v^{\text{ft}}_{j,fn}]$, $\mathbf{W}^{\text{ex}}_j = [w^{\text{ex}}_{j,fl}]$, $\mathbf{U}^{\text{ex}}_j = [u^{\text{ex}}_{j,lk}]$, $\mathbf{G}^{\text{ex}}_j = [g^{\text{ex}}_{j,km}]$, $\mathbf{H}^{\text{ex}}_j = [h^{\text{ex}}_{j,mn}]$, $\boldsymbol{\rho}^{\text{ex}}_{w,j} = [\rho^{\text{ex}}_{w,j,fl}]$, and $\boldsymbol{\tau}^{\text{ex}}_{w,j} = [\tau^{\text{ex}}_{w,j,fl}]$, we obtain the following updates:

$$\boldsymbol{\rho}^{\text{ex}}_{w,j} = \mathbb{E}[\mathbf{V}^{\text{ex}}_j]^{.-1}\big(\mathbb{E}[\mathbf{U}^{\text{ex}}_j]\mathbb{E}[\mathbf{G}^{\text{ex}}_j]\mathbb{E}[\mathbf{H}^{\text{ex}}_j]\big)^T \qquad (46)$$

$$\begin{aligned}
\boldsymbol{\tau}^{\text{ex}}_{w,j} = \mathbb{E}\Big[\frac{1}{\mathbf{W}^{\text{ex}}_j}\Big]^{.-2} \odot \Bigg(&\Big(\mathbb{E}[|\mathbf{S}_j|^{.2}] \odot \mathbf{C}_j^{.-2} \odot \mathbb{E}\Big[\frac{1}{\mathbf{V}^{\text{ft}}_j}\Big]^{.-1}\Big) \\
&\Big(\mathbb{E}\Big[\frac{1}{\mathbf{U}^{\text{ex}}_j}\Big]^{.-1}\mathbb{E}\Big[\frac{1}{\mathbf{G}^{\text{ex}}_j}\Big]^{.-1}\mathbb{E}\Big[\frac{1}{\mathbf{H}^{\text{ex}}_j}\Big]^{.-1}\Big)^T\Bigg) \qquad (47)
\end{aligned}$$

$$\gamma^{\text{ex}}_{w,j,fl} = 0 \qquad (48)$$

where ^{T} denotes matrix transposition, $\odot$ denotes element-wise multiplication, $^{.\alpha}$ denotes element-wise exponentiation, and complex norm and division are also computed element-wise.

The variational updates for the other NTF parameters $\mathbf{U}^{\text{ex}}$, $\mathbf{G}^{\text{ex}}$, $\mathbf{H}^{\text{ex}}$, $\mathbf{W}^{\text{ft}}$, $\mathbf{U}^{\text{ft}}$, $\mathbf{G}^{\text{ft}}$, and $\mathbf{H}^{\text{ft}}$ are derived by following the same steps as above, yielding similar formulae for the GIG parameters. See [50] for the complete set of formulae.

Each time-frequency coefficient of the expectation $\mathbb{E}[|\mathbf{S}_j|^{.2}]$ in (47) is calculated as

$$\mathbb{E}[|s_{j,fn}|^2] = |(\boldsymbol{\mu}_{\mathbf{s},fn})_j|^2 + (\mathbf{R}_{\mathbf{ss},fn})_{jj} \qquad (49)$$

where $\boldsymbol{\mu}_{\mathbf{s},fn}$ and $\mathbf{R}_{\mathbf{ss},fn}$ are the first and second order posterior moments of $\mathbf{s}_{fn}$ derived below in (56) and (57). The expectations involving the NTF parameters in (46) and (47) are computed via the following formulae for the GIG distribution [64]:

$$\mathbb{E}[y] = \frac{\mathcal{K}_{\gamma+1}(2\sqrt{\rho\tau})\sqrt{\tau}}{\mathcal{K}_\gamma(2\sqrt{\rho\tau})\sqrt{\rho}}, \qquad (50)$$

$$\mathbb{E}\Big[\frac{1}{y}\Big] = \frac{\mathcal{K}_{\gamma-1}(2\sqrt{\rho\tau})\sqrt{\rho}}{\mathcal{K}_\gamma(2\sqrt{\rho\tau})\sqrt{\tau}}. \qquad (51)$$

$\mathbb{E}[\mathbf{V}^{\text{ex}}_j]$ is obtained as

$$\mathbb{E}[\mathbf{V}^{\text{ex}}_j] = \mathbb{E}[\mathbf{W}^{\text{ex}}_j]\mathbb{E}[\mathbf{U}^{\text{ex}}_j]\mathbb{E}[\mathbf{G}^{\text{ex}}_j]\mathbb{E}[\mathbf{H}^{\text{ex}}_j] \qquad (52)$$

and $\mathbb{E}[1/\mathbf{V}^{\text{ft}}_j]^{.-1}$ is a shorthand notation for

$$\mathbb{E}\Big[\frac{1}{\mathbf{W}^{\text{ft}}_j}\Big]^{.-1}\mathbb{E}\Big[\frac{1}{\mathbf{U}^{\text{ft}}_j}\Big]^{.-1}\mathbb{E}\Big[\frac{1}{\mathbf{G}^{\text{ft}}_j}\Big]^{.-1}\mathbb{E}\Big[\frac{1}{\mathbf{H}^{\text{ft}}_j}\Big]^{.-1}. \qquad (53)$$

$\boldsymbol{\omega}_j = \mathbb{E}[\mathbf{V}_j]$ and $\mathbf{C}_j = \mathbb{E}[1/\mathbf{V}_j]^{.-1}$ are computed similarly to (52) and (53), respectively, by matrix multiplication of the statistics of the 8 sets of NTF parameters.

C. Variational Updates for the Source STFT Coefficients

For the source STFT coefficients $\mathbf{s}_{fn}$, the update (15) is given by

$$\begin{aligned}
\log q^*(\mathbf{s}_{fn}) = &\frac{1}{\sigma_b^2}\mathbf{s}_{fn}^H\boldsymbol{\mu}_{\mathbf{A},f}^H\mathbf{x}_{fn} + \frac{1}{\sigma_b^2}\mathbf{x}_{fn}^H\boldsymbol{\mu}_{\mathbf{A},f}\mathbf{s}_{fn} \\
&- \frac{1}{\sigma_b^2}(\mathbf{s}_{fn}^H\mathbf{R}_{\mathbf{A},f}\mathbf{s}_{fn}) - \mathbf{s}_{fn}^H\mathbf{C}_{fn}{}^{-1}\mathbf{s}_{fn} + \text{const}, \qquad (54)
\end{aligned}$$

where ^{H} denotes conjugate transposition, $\boldsymbol{\mu}_{\mathbf{A},f}$ and $\mathbf{R}_{\mathbf{A},f}$ are the first and second order posterior raw moments of $\mathbf{A}_f$ derived below in (63) and (64), and $\mathbf{C}_{fn} = \mathbf{Diag}([C_{j,fn}]_j)$ is the diagonal matrix with elements $C_{j,fn}$.

This distribution involves a linear term in $\mathbf{s}_{fn}$, its conjugate, and quadratic terms. The optimal variational distribution is thus a complex-valued Gaussian:

$$q^*(\mathbf{s}_{fn}) = \mathcal{N}(\mathbf{s}_{fn};\boldsymbol{\mu}_{\mathbf{s},fn},\mathbf{R}_{\mathbf{ss},fn}). \qquad (55)$$

The covariance is obtained by rearranging the quadratic terms and making them equal to $-\mathbf{s}_{fn}^H\mathbf{R}^{-1}_{\mathbf{ss},fn}\mathbf{s}_{fn}$:

$$\mathbf{R}_{\mathbf{ss},fn} = \Big(\mathbf{C}_{fn}{}^{-1} + \frac{1}{\sigma_b^2}\mathbf{R}_{\mathbf{A},f}\Big)^{-1}. \qquad (56)$$

Similarly, the mean is obtained by identifying the linear term in $\mathbf{s}_{fn}^H$ with $\mathbf{s}_{fn}^H\mathbf{R}^{-1}_{\mathbf{ss},fn}\boldsymbol{\mu}_{\mathbf{s},fn}$:

$$\boldsymbol{\mu}_{\mathbf{s},fn} = \frac{1}{\sigma_b^2}\mathbf{R}_{\mathbf{ss},fn}\boldsymbol{\mu}_{\mathbf{A},f}^H\mathbf{x}_{fn}. \qquad (57)$$

D. Variational Updates for the Mixing Parameters

In order to derive the variational updates for the mixing parameters, we reshape $\mathbf{A}_f$ into a column vector $\underline{\mathbf{A}}_f$ by concatenating and transposing the rows of $\mathbf{A}_f$. The distribution (15) for the reshaped mixing parameters is given by

$$\log q^*(\underline{\mathbf{A}}_f) = \frac{1}{\sigma_b^2}\underline{\mathbf{A}}_f^H \sum_n \mathbf{R}_{\mathbf{xs},fn} + \frac{1}{\sigma_b^2}\sum_n \mathbf{R}_{\mathbf{xs},fn}^H \underline{\mathbf{A}}_f$$

$$- \underline{\mathbf{A}}_f^H \frac{1}{\sigma_b^2}\sum_n \mathbf{Diag}(\underbrace{\mathbf{R}_{\mathbf{s},fn}^T, \dots, \mathbf{R}_{\mathbf{s},fn}^T)}_{\text{I times}} \underline{\mathbf{A}}_f + \text{const} \quad (58)$$

where $\mathbf{Diag}(\mathbf{R}_{\mathbf{s},fn}^T, \dots, \mathbf{R}_{\mathbf{s},fn}^T)$ is the $IJ \times IJ$ block-diagonal matrix with elements $\mathbf{R}_{\mathbf{s},fn}^T$, $\mathbf{R}_{\mathbf{s},fn}$ is the second order raw moment of the source STFT coefficients given by

$$\mathbf{R}_{\mathbf{s},fn} = \boldsymbol{\mu}_{\mathbf{s},fn}\boldsymbol{\mu}_{\mathbf{s},fn}^H + \mathbf{R}_{\mathbf{ss},fn} \quad (59)$$

and $\mathbf{R}_{\mathbf{xs},fn}$ is the cross-moment given by

$$\mathbf{R}_{\mathbf{xs},fn} = [x_{1,fn}\boldsymbol{\mu}_{\mathbf{s},fn}^H, \dots, x_{I,fn}\boldsymbol{\mu}_{\mathbf{s},fn}^H]^T. \quad (60)$$

The distribution in (58) involves a linear term in $\underline{\mathbf{A}}_f$, its conjugate, and a quadratic term. Hence, the optimal variational distribution is again a complex-valued Gaussian:

$$q^*(\underline{\mathbf{A}}_f) = \mathcal{N}(\underline{\mathbf{A}}_f; \boldsymbol{\mu}_{\underline{\mathbf{A}},f}, \mathbf{R}_{\underline{\mathbf{A}\mathbf{A}},f}). \quad (61)$$

The covariance and the mean of this distribution are obtained similarly to above as

$$\mathbf{R}_{\underline{\mathbf{A}\mathbf{A}},f} = \Big(\frac{1}{\sigma_b^2}\sum_n \mathbf{Diag}\,(\underbrace{\mathbf{R}_{\mathbf{s},fn}^T, \dots, \mathbf{R}_{\mathbf{s},fn}^T)}_{\text{I times}}\Big)^{-1} \quad (62)$$

$$\boldsymbol{\mu}_{\underline{\mathbf{A}},f} = \frac{1}{\sigma_b^2}\mathbf{R}_{\underline{\mathbf{A}\mathbf{A}},f}\sum_n \mathbf{R}_{\mathbf{xs},fn}. \quad (63)$$

From these expressions, we can derive the second order raw moment $\mathbf{R}_{\mathbf{A},f}$ of $\mathbf{A}_f$ in (56) as follows:

$$\mathbf{R}_{\mathbf{A},f} = \sum_i ([\boldsymbol{\mu}_{\underline{\mathbf{A}},f}\boldsymbol{\mu}_{\underline{\mathbf{A}},f}^H + \mathbf{R}_{\underline{\mathbf{A}\mathbf{A}},f}]_{ii})^T. \quad (64)$$

where $[\cdot]_{ii}$ denotes the diagonal $J \times J$ block corresponding to channel i. The first order moment $\boldsymbol{\mu}_{\mathbf{A},f}$ is simply obtained by reshaping $\boldsymbol{\mu}_{\underline{\mathbf{A}},f}$ back into matrix form.

E. Lower Bound

The expectation of the log-likelihood is as follows:

$$\mathbb{E}[\log p(\mathbf{X}|\mathbf{S},\mathbf{A})] = -F\,N\,I\log\pi\sigma_b^2$$

$$-\sum_{fn}\frac{1}{\sigma_b^2}\Big(\mathbf{x}_{fn}^H\mathbf{x}_{fn} - \mathbf{x}_{fn}^H\boldsymbol{\mu}_{A,f}\boldsymbol{\mu}_{\mathbf{s},fn} - \boldsymbol{\mu}_{\mathbf{s},fn}^H\boldsymbol{\mu}_{\mathbf{A},f}^H\mathbf{x}_{fn}$$

$$+ \operatorname{tr}((\boldsymbol{\mu}_{s,fn}\boldsymbol{\mu}_{s,fn}^H + \mathbf{R}_{\mathbf{ss},fn})\mathbf{R}_{\mathbf{A},f})\Big). \quad (65)$$

The second term is given by replacing (40) and (41) into (39), taking the expectation and adding the entropy of the Gaussian posterior:

$$\mathbb{E}[\log \underline{p}(\mathbf{S}|\mathbf{V}) - \log q(\mathbf{S})] = F\,N\,J + \sum_{fn}\log\det(\mathbf{R}_{ss,fn})$$

$$-\sum_{j,fn}\Big(\log\omega_{j,fn} + \frac{1}{C_{j,fn}}\mathbb{E}[|s_{j,fn}|^2]\Big). \quad (66)$$

The third term is given by the entropy of the Gaussian posterior:

$$\mathbb{E}[\log p(\mathbf{A}) - \log q(\mathbf{A})] = F\,I\,J(\log\pi + 1)$$

$$+ \sum_f \log\det(\mathbf{R}_{\underline{\mathbf{A}\mathbf{A}},f}). \quad (67)$$

The term associated with $\mathbf{W}^{\text{ex}}$ is written as follows:

$$\mathbb{E}[\log p(\mathbf{W}^{\text{ex}}) - \log q(\mathbf{W}^{\text{ex}})] = \sum_{fl}\Big(\rho_{w,j,fl}^{\text{ex}}\mathbb{E}[w_{j,fl}^{\text{ex}}]$$

$$+ \tau_{w,j,fl}^{\text{ex}}\mathbb{E}\Big[\frac{1}{w_{j,fl}^{\text{ex}}}\Big] + \log\mathcal{K}_{\gamma_{w,j,fl}^{\text{ex}}}(2\sqrt{\rho\tau}) + \log 2\Big). \quad (68)$$

The contribution of the other NTF parameters to the lower bound is calculated using a similar formula.

References

[1] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding, part 1," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 75–80, May 2009.

[2] J. S. Downie, D. Byrd, and T. Crawford, "Ten years of ISMIR: Reflections on challenges and opportunities," in *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*, 2009, pp. 13–18.

[3] L. Deng, "Front-end, back-end, and hybrid techniques for noise-robust speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*. Springer, 2011, pp. 67–99.

[4] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[5] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals," in *Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 559–564.

[6] J. Zapata and E. Gómez, "Improving beat tracking in the presence of highly predominant vocals using source separation techniques: Preliminary study," in *Proceedings of 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, 2012, pp. 583–590.

[7] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010.

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

[9] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 2015.

[10] U. Şimşekli, T. Virtanen, and A. T. Cemgil, "Non-negative tensor factorization models for Bayesian audio processing," *Digital Signal Processing*, vol. 47, pp. 178–191, 2015.

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

[12] 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*, vol. 18, no. 7, pp. 1830–1840, July 2010.

[13] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, vol. 2008, 2008, article ID 872425.

[14] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[15] Y. Salaün, E. Vincent, N. Bertin, N. Souviraà-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot, "The flexible audio source separation toolbox version 2.0," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Show & Tell*, 2014.

[16] M. Fakhry, P. Svaizer, and M. Omologo, “Audio source separation using a redundant library of source spectral bases for non-negative tensor factorization,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 251–255.

[17] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, “Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization,” *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, 2015.

[18] N. Souviraà-Labastie, A. Olivero, E. Vincent, and F. Bimbot, “Multichannel audio source separation using multiple deformed references,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1775–1787, 2015.

[19] D. T. Tran, E. Vincent, D. Jouvet, and K. Adiloğlu, “Using full-rank spatial covariance models for noise-robust ASR,” in *Proceedings of 2nd International Workshop on Machine Listening in Multisource Environments (CHiME)*, 2013, pp. 31–32.

[20] B. Lehner and G. Widmer, “Monaural blind source separation in the context of vocal detection,” in *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*, 2015, pp. 309–315.

[21] A. Ozerov, Ç. Bilen, and P. Pérez, “Multichannel audio declipping,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 31–32.

[22] H. Attias, “A variational Bayesian framework for graphical models,” in *Advances in Neural Information Processing Systems (NIPS)*, 1999.

[23] C. Févotte and S. J. Godsill, “A Bayesian approach for blind separation of sparse sources,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2174–2188, 2006.

[24] R.-B. Chen and Y. N. Wu, “A null space method for over-complete blind source separation,” *Computational Statistics & Data Analysis*, vol. 51, pp. 5519–5536, 2007.

[25] C. Févotte, B. Torrésani, L. Daudet, and S. J. Godsill, “Sparse linear regression with structured priors and application to denoising of musical audio,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 174–185, 2008.

[26] K. Adiloğlu 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)*, 2011, pp. 1663–1667.

[27] M. Kim, P. Smaragdis, G. G. Ko, and R. A. Rutenbar, “Stereophonic spectrogram segmentation using Markov random fields,” in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012, pp. 1–6.

[28] J.-T. Chien and H.-L. Hsieh, “Bayesian group sparse learning for music source separation,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, p. 18, 2013.

[29] 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, April 2007.

[30] S. J. Rennie, J. R. Hershey, and P. A. Olsen, “Single-channel multitalker speech recognition,” *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 66–80, 2010.

[31] 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.

[32] G. Mysore and M. Sahani, “Variational inference in non-negative factorial hidden Markov models for efficient audio source separation,” in *Proceedings of 29th International Conference on Machine Learning (ICML)*, 2012, pp. 1887–1894.

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

[34] J.-T. Chien and H.-L. Hsieh, “Nonstationary source separation using sequential and variational Bayesian learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 681–694, 2013.

[35] J.-T. Chien and P.-K. Yang, “Bayesian factorization and learning for monaural source separation,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 1, pp. 185–195, 2016.

[36] L. Deng, J. Droppo, and A. Acero, “Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, May 2005.

[37] 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*, vol. 17, no. 2, pp. 324–334, 2009.

[38] 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*, vol. 2010, 2010, article ID 651420.

[39] R. Astudillo, “Integration of short-time Fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition,” Ph.D. dissertation, TU Berlin, 2010.

[40] F. Nesta, M. Matassoni, and R. F. Astudillo, “A flexible spatial blind source extraction framework for robust speech recognition in noisy environments,” in *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME)*, 2013, pp. 33–40.

[41] D. T. Tran, E. Vincent, and D. Jouvet, “Nonparametric uncertainty estimation and propagation for noise robust ASR,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 11, pp. 1835–1846, 2015.

[42] A. H. Abdelaziz, S. Watanabe, J. R. Hershey, E. Vincent, and D. Kolossa, “Uncertainty propagation through deep neural networks,” in *Proceedings of Interspeech*, 2015, pp. 3561–3565.

[43] A. Ozerov, M. Lagrange, and E. Vincent, “Uncertainty-based learning of acoustic models from noisy data,” *Computer Speech and Language*, vol. 27, no. 3, pp. 874–894, 2013.

[44] C. Yu, G. Liu, S. Hahm, and J. H. L. Hansen, “Uncertainty propagation in front end factor analysis for noise robust speaker recognition,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4017–4021.

[45] M. Lagrange, A. Ozerov, and E. Vincent, “Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning,” in *Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 595–600.

[46] M. J. F. Gales, “Model-based techniques for noise robust speech recognition,” Ph.D. dissertation, Gonville and Caius College, University of Cambridge, 1995.

[47] P. J. Moreno, B. Raj, and R. M. Stern, “A vector Taylor series approach for environment-independent speech recognition,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 1996, pp. 733–736.

[48] 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*, 2010, pp. 713–716.

[49] K. Adiloğlu 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, pp. 273–276.

[50] ——, “Supporting technical report for the article “Variational Bayesian inference for source separation and robust feature extraction”,” Inria, Tech. Rep. RT-0423, 2012. [Online]. Available: http://hal.inria.fr/hal-00687162/PDF/RT-423.pdf

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

[52] C. Blandin, A. Ozerov, and E. Vincent, “Multi-source TDOA estimation in reverberant audio using angular spectra and clustering,” *Signal Processing*, vol. 92, pp. 1950–1960, 2012.

[53] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, “The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges,” *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[54] E. Vincent, R. Gribonval, and C. Févotte, “Performance measures in blind audio source separation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[55] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, “The second ‘CHiME’ speech separation and recognition challenge: : Datasets, tasks and baselines,” in *Proceedings of IEEE Conference on Audio, Speech and SIgnal Processing (ICASSP)*, 2013, pp. 162–167.

[56] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, “Modelling non-stationary noise with spectral factorisation in automatic speech recognition,” *Computer Speech and Language*, vol. 27, no. 3, pp. 763–779, May 2013.

[57] N. Moritz, M. R. Schädler, K. Adiloğlu, B. T. Meyer, T. Jürgens, T. Gerkmann, and S. Goetze, “Noise robust distant automatic speech recognition utilizing NMF based source separation and auditory feature extraction,” in *Proceedings of 2nd CHiME challenge workshop*, 2013.

[58] D. T. Tran, E. Vincent, and D. Jouvet, “Extension of uncertainty propagation to dynamic MFCCs for noise robust ASR,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 31–32.

[59] L. K. Saul and M. I. Jordan, “Exploiting tractable substructures in intractable networks,” in *Advances in Neural Information Processing Systems (NIPS)*, 1995, pp. 486–492.

[60] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, Sep. 2014.

[61] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[62] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in *Proceedings of 12th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2015, pp. 91–99.

[63] S. Sivasankaran, A. A. Nugraha, E. Vincent, J. A. Morales-Cordovilla, S. Dalmia, and I. Illina, “Robust ASR using neural network based speech enhancement and feature simulation,” in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 482–489.

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