A General Variational Bayesian Framework for Robust Feature Extraction in Multisource Recordings

Kamil Adiloglu, Emmanuel Vincent

A GENERAL VARIATIONAL BAYESIAN FRAMEWORK FOR ROBUST FEATURE EXTRACTION IN MULTISOURCE RECORDINGS

Kamil Adiloğlu *Emmanuel Vincent*

INRIA, Centre de Rennes - Bretagne Atlantique
Campus de Beaulieu, 35042, Rennes cedex, France
email: {kamil.adiloglu, emmanuel.vincent}@inria.fr

ABSTRACT

We consider the problem of extracting features from individual sources in a multisource audio recording using a general source separation algorithm. The main issue is to estimate and propagate the uncertainty over the separated source signals, so as to robustly estimate the features despite source separation errors. While state-of-the-art techniques estimate the uncertainty in a heuristic manner, we propose to integrate over the parameter space of the source separation algorithm. We apply variational Bayes to estimate the posterior probability of the sources and subsequently derive the expectation of the features by moment matching. Experiments over stereo mixtures of three or four sources show that the proposed method provides the best results in terms of the root mean square (RMS) error on the estimated features.

Index Terms— Bayesian source separation, robust feature extraction

1. INTRODUCTION

Many applications in the field of audio information retrieval are typically solved by extracting features describing the audio content and exploiting them for *e.g.* speech recognition, speaker recognition, cover version detection etc. However, most audio signals consist of a mixture of several sound sources, which have their own characteristics. Applying source separation prior to feature extraction 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 [1].

A heuristic approach is to assume that the uncertainty is proportional to the squared difference between the separated sources and the mixture [1, 2]. In [3], 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 [3] or unscented transform [1]. 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, 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) [4]. In this paper, we extend this approach to the general modeling framework for source separation recently introduced in [5]. 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 propose 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. Figure 1 illustrates the workflow of the proposed approach.

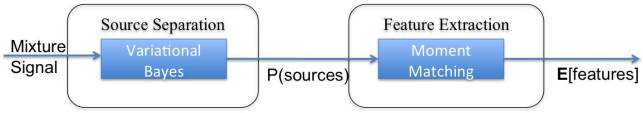

Fig. 1. Flow of the proposed Bayesian source separation and feature extraction approach.

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

Classically, we address the source separation problem in the time-frequency domain by means of the Short-Time Fourier Transform (STFT). For J sources and I channels, the mixing equation writes

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \epsilon_{fn}, \quad (1)$$

where $\mathbf{x}_{fn} = [x_{1,fn}, \ldots, x_{I,fn}]^T$ denotes the mixture STFT coefficients, $\mathbf{s}_{fn} = [s_{1,fn}, \ldots, s_{J,fn}]^T$ the source STFT coefficients. $\mathbf{A}_f = [\mathbf{A}_{1,f}, \ldots, \mathbf{A}_{J,f}]$ is the complex valued mixing matrix, where $\mathbf{A}_{j,f}$ is the vector of mixing coefficients for source j and ϵ_{fn} is 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 [5].

We adopt a local Gaussian model [5] for the source coefficients. We set a zero-mean complex Gaussian prior over the source coefficients $s_{j,fn}$ with variance $v_{j,fn}$:

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

The source variances $v_{j,fn}$, which encode the spectral power are decomposed via an excitation-filter model [5]:

This work was supported by the Quaero Program funded by Oseo.

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

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 [5]. As a result, the complete factorization scheme is as follows:

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

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

3. PROPOSED VARIATIONAL INFERENCE ALGORITHM

For the separation of the sources, we propose a VB algorithm [6].

3.1. The Bayesian Approach

Let us denote the set of all model parameters as

$$\boldsymbol{\theta} = \{\mathbf{A}_f, h^{\text{ex}}_{j,mn}, g^{\text{ex}}_{j,km}, u^{\text{ex}}_{j,lk}, w^{\text{ex}}_{j,fl}, h^{\text{ft}}_{j,m'n}, g^{\text{ft}}_{j,k'm'}, u^{\text{ft}}_{j,l'k'}, w^{\text{ft}}_{j,fl'}\}. \quad (6)$$

For the sake of simplicity, we define all prior probabilities to be non-informative, even-though informative priors could be defined as well. Hence, $\mathbf{A}_f$ follows a flat prior and the non-negative factors $\{h^{\text{ex}}_{j,mn}, g^{\text{ex}}_{j,km}, \ldots, u^{\text{ft}}_{j,l'k'}, w^{\text{ft}}_{j,fl'}\}$ follow a Jeffreys prior.

Finally, we assume Gaussian noise ϵ_{fn} with fixed variance $\sigma_b^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

Under these assumptions, we aim to estimate the posterior probability of the source coefficients, which is given by

$$p(\mathbf{s}|\mathbf{x}) \propto \int p(\mathbf{x}|\mathbf{s},\boldsymbol{\theta}) p(\mathbf{s}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (7)$$

This integral is intractable. By contrast with previous approaches [1], which consist of estimating the maximum likelihood (ML) value $\hat{\boldsymbol{\theta}}$ for the parameters and considering $p(\mathbf{x}|\mathbf{s},\hat{\boldsymbol{\theta}})$, we resort to more accurate VB inference [6].

3.2. Algorithm

VB minimizes the Kullback-Leibler (KL) divergence between the true posterior distribution $p(\mathbf{s},\boldsymbol{\theta}|\mathbf{X})$ and some approximation $q(\mathbf{s},\boldsymbol{\theta})$, which is typically specified by assuming some factorization. Unfortunately, direct application of this rule does not yield a closed form solution. Therefore we pursue a decomposition of the sources as in [7] for the above model.

A source $s_{j,fn}$ is sub-divided into $\Lambda_j = \Lambda^{\text{ex}}_j \Lambda^{\text{ft}}_j$ sub-components such that $\Lambda^{\text{ex}}_j = K^{\text{ex}}_j M^{\text{ex}}_j L^{\text{ex}}_j$ and $\Lambda^{\text{ft}}_j = K^{\text{ft}}_j M^{\text{ft}}_j L^{\text{ft}}_j$. Let us define λ to be a joint index of $\{k,m,l\}$ and λ' is a joint index of $\{k',m',l'\}$. We then decompose the sources as

$$s_{j,fn} = \sum_{\lambda} \sum_{\lambda'} c_{j,fn,\lambda,\lambda'}, \quad (8)$$

where each sub-component $c_{j,fn,\lambda,\lambda'}$ follows a complex Gaussian distribution

$$c_{j,fn,\lambda,\lambda'} \sim \mathcal{N}(0, v_{j,fn,\lambda,\lambda'}), \quad (9)$$

with

$$v_{j,fn,\lambda,\lambda'} = h^{\text{ex}}_{j,mn} g^{\text{ex}}_{j,km} u^{\text{ex}}_{j,lk} w^{\text{ex}}_{j,fl} h^{\text{ft}}_{j,m'n} g^{\text{ft}}_{j,k'm'} u^{\text{ft}}_{j,l'k'} w^{\text{ft}}_{j,fl'}. \quad (10)$$

The mixing equation for the source sub-components can be rewritten as

$$\mathbf{x}_{fn} = \mathbf{A}^{+}_f \mathbf{c}_{fn} + \epsilon_f, \quad (11)$$

where $\mathbf{A}^{+}_f$ indicates the extended mixing matrix where the elements $a_{i,j,f}$ of $\mathbf{A}_f$ are repeated Λ_j times.

Assuming the following factorization of the variational approximation of the posterior

$$q(\mathbf{c},\boldsymbol{\theta}) = \Big(\prod_{f,n} q(\mathbf{c}_{fn})\Big)\Big(\prod_{f} q(\mathbf{A}_f))\Big) \Big(\prod_{j,m,n} q(h^{\text{ex}}_{j,mn}) \cdots \prod_{j,f,l'} q(w^{\text{ft}}_{j,fl})\Big), \quad (12)$$

we obtain the optimal factors of $\mathbf{c}_{fn}$, $\underline{\mathbf{A}}_f$ and $w^{\text{ex}}_{j,fl}$ as in the following. Due to space limitations, we give the update equations of $w^{\text{ex}}_{j,fl}$ among all of the NMF components. However the rest of the factors follow a similar shape:

$$q^*(\mathbf{c}_{fn}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{c},fn}, \mathbf{R}_{\mathbf{cc},fn}), \quad (13)$$

$$q^*(\underline{\mathbf{A}}_f) \sim \mathcal{N}(\boldsymbol{\mu}_{\underline{A},f}, \mathbf{R}_{\underline{\mathbf{AA}},f}), \quad (14)$$

$$q^*(w^{\text{ex}}_{j,fl}) \sim \text{Inv-Gamma}(\alpha^{\text{ex}}_{w,j,fl}, \beta^{\text{ex}}_{w,j,fl}), \quad (15)$$

where $\underline{\mathbf{A}}_f$ denotes the column vector obtained by concatenating the rows of the mixing matrix $\mathbf{A}_f$ and

$$\mathbf{R}_{\mathbf{cc},fn} = \Big(\mathbf{B}_{c,fn} + (\sigma_b^2\mathbf{I})^{-1} \sum_i ([\boldsymbol{\mu}_{\underline{A},f} \boldsymbol{\mu}^H_{\underline{A},f} + \mathbf{R}_{\underline{\mathbf{AA}},f}]_{ii})^T\Big)^{-1}, \quad (16)$$

$$\boldsymbol{\mu}_{c,fn} = \mathbf{R}_{\mathbf{cc},fn} (\boldsymbol{\mu}^{+}_{\underline{A},f})^H (\sigma_b^2\mathbf{I})^{-1} \mathbf{x}_{fn}, \quad (17)$$

$$\mathbf{R}_{\underline{\mathbf{AA}},f} = \Big(\frac{1}{\sigma_b^2} \sum_n \text{diag} \underbrace{(\mathbf{R}^T_{s,fn}, \ldots, \mathbf{R}^T_{s,fn})}_{\text{I times}}\Big)^{-1}, \quad (18)$$

$$\boldsymbol{\mu}_{\underline{A},f} = \boldsymbol{\Sigma}_{\underline{A},f} (\frac{1}{\sigma_b^2} \sum_n \mathbf{R}_{\mathbf{xs},fn}), \quad (19)$$

$$\alpha^{\text{ex}}_{w,j,fl} = N K^{\text{ex}}_j M^{\text{ex}}_j \Lambda^{\text{ft}}_j, \quad (20)$$

$$\beta^{\text{ex}}_{w,j,fl} = N^5 F^6 (K^{\text{ex}}_j)^5 (M^{\text{ex}}_j)^5 (L^{\text{ex}}_j)^6 (\Lambda^{\text{ft}}_j)^5 \sum_{n,km,\lambda'} \frac{(\boldsymbol{\mu}_{c,fn})^2_{(j,kml,\lambda')} + (\mathbf{R}_{\mathbf{cc},fn})_{(j,kml,\lambda'),(j,kml,\lambda')}}{\beta^{\text{ex}}_{h,j,mn} \beta^{\text{ex}}_{g,j,km} \beta^{\text{ex}}_{u,j,lk} \beta^{\text{ft}}_{h,j,m'n} \beta^{\text{ft}}_{g,j,k'm'} \beta^{\text{ft}}_{u,j,l'k'} \beta^{\text{ft}}_{w,j,fl'}}. \quad (21)$$

In (16), $[\cdot]_{ii}$ denotes the diagonal $J \times J$ block corresponding to channel i. In the same equation, the matrix $\mathbf{B}_{c,fn}$ writes

$$\mathbf{B}_{c,fn} = \mathrm{diag}\Big(\frac{N^6 F^6 (\Lambda_j^{\mathrm{ex}})^6}{\beta^{\mathrm{ex}}_{w,j,fl}\beta^{\mathrm{ex}}_{u,j,lk}\beta^{\mathrm{ex}}_{g,j,km}\beta^{\mathrm{ex}}_{h,j,mn}} \frac{(\Lambda_j^{\mathrm{ft}})^6}{\beta^{\mathrm{ft}}_{w,j,fl'}\beta^{\mathrm{ft}}_{u,j,l'k'}\beta^{\mathrm{ft}}_{g,j,k'm'}\beta^{\mathrm{ft}}_{h,j,m'n}}\Big)^{J,\Lambda_j^{\mathrm{ex}},\Lambda_j^{\mathrm{ft}}}_{j,kml,k'm'l'}. \quad (22)$$

In (19), $\mathbf{R}_{\mathbf{xs},fn}$ is given by

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

Finally, the posterior distribution of the source coefficients is calculated by summing up the corresponding elements of the mean and the covariance of the source sub-components as follows:

$$\mu_{\mathbf{s},j,fn} = \sum_{\lambda}\sum_{\lambda'} \mu_{c,j,fn,\lambda,\lambda'}, \quad (24)$$

$$(\mathbf{R}_{\mathbf{ss},fn})_{j,j'} = \sum_{\kappa}\sum_{\lambda'} (\mathbf{R}_{\mathbf{cc},fn})_{(j,\lambda,\lambda'),(j',\lambda,\lambda')}. \quad (25)$$

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

$$\mathbf{R}_{s,fn} = \boldsymbol{\mu}_{s,nf}\boldsymbol{\mu}^H_{s,nf} + \mathbf{R}_{\mathbf{ss},fn}. \quad (26)$$

The VB update equations of the scale parameters of the NMF components $\{\beta^{\mathrm{ex}}_{w,j,fl}, \ldots, \beta^{\mathrm{ft}}_{h,j,m'n}\}$ turn out to be identical to the state-of-the-art expectation maximization (EM) updates of the NMF components $\{\hat{w}^{\mathrm{ex}}_{j,fl}, \ldots, \hat{h}^{\mathrm{ft}}_{j,m'n}\}$ [7] up to a variable change. For $\beta^{\mathrm{ex}}_{w,j,fl}$ this variable change writes

$$\hat{w}^{\mathrm{ex}}_{j,fl} = \frac{\beta^{\mathrm{ex}}_{w,j,fl}}{NK^{\mathrm{ex}}_j M^{\mathrm{ex}}_j K^{\mathrm{ft}}_j M^{\mathrm{ft}}_j L^{\mathrm{ft}}_j}. \quad (27)$$

Therefore, we replaced these EM updates by the multiplicative updates in [5], which converge much faster.

Equations (16) to (27) depend on each other. After proper initialization, we cycle through these equations by replacing the dependent values with their new estimates.

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 [5].

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:

$$\boldsymbol{\mu}_{\mathbf{y},j,fn} = \boldsymbol{\mu}^{\mathrm{post}}_{A,f}\,\boldsymbol{\mu}^{\mathrm{post}}_{s,fn}, \quad (28)$$

$$\boldsymbol{R}_{\mathbf{yy},j,fn} = \Big(\sum_{rr'}((\mathbf{R}_{\mathbf{AA},f})_{(ir,i'r')} + (\mu_{\mathbf{A},f})_{(ir)}(\mu_{\mathbf{A},f})_{(i'r')}) ((\mathbf{R}_{\mathbf{ss},fn})_{(r,r')} + (\boldsymbol{\mu}_{\mathbf{s},fn})_{(r)}(\boldsymbol{\mu}_{\mathbf{s},fn})^H_{(r')})\Big)^{II}_{ii'} - \boldsymbol{\mu}_{\mathrm{y},j,fn}\boldsymbol{\mu}^H_{\mathrm{y},j,fn}. \quad (29)$$

4.2. Uncertainty Propagation for Feature Extraction

We calculate the expectation of the MFCCs for each source as

$$\mu^{\mathbf{MFCC}}_{jn} = \int \mathrm{MFCC}(\mathbf{y}_{j1n})P(\mathbf{y}_{j1n})\,d\mathbf{y}_{j1n} \quad (30)$$

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 $\mathrm{MFCC}(\mathbf{y}_{j1n}) = 20\,\mathbf{D}\log_{10}(\mathbf{M}|\mathbf{y}_{j1n}|)$. 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 order to obtain the mean estimate of the MFCCs $\mu^{\mathbf{MFCC}}_{jn}$, we first compute the mean and variance of $|\mathbf{y}_{j1n}|$ using formulae for the Rice distribution in [1] and then propagate them through the logarithm using the moment matching formulae in [8]. Note that this also makes it possible to estimate the variance of the MFCCs and exploit it for *e.g.* classification tasks [1]. In the following, due to space limitations, we simply consider the mean.

5. EXPERIMENTAL EVALUATION

5.1. Data and Algorithmic Settings

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.

We performed experiments with eight different sets of constraints over the parameters as considered in the experiments section of [5]. 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 $w^{\mathrm{ex}}_{j,fl}$ are either unconstrained (un) or fixed (co) to harmonic and noise-like patterns.
- Temporal Structure: The time-localized patterns $h^{\mathrm{ex}}_{j,mn}$ 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 [9]. The noise variance σ^2_b is initialized to 10^{-6}. Finally, we performed 200 iterations for convergence.

5.2. Evaluation Criteria

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 [10] 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 terms of the RMS error [4] between the estimated $\mu^{\mathbf{MFCC}}_{jn}$ and the true MFCCs.We ignore the first MFCC coefficient and consider the MFCCs 2 to 20 only.

[1] http://sisec2008.wiki.irisa.fr/tiki-index.php?page=Under-determined+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	1.58	1.68	1.87	2.07	1.77	1.75	2.28	2.25
VB	1.61	1.72	1.93	2.06	1.84	1.80	**2.49**	2.29

Table 1. SDR in dB achieved by ML or VB source separation over all 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	VB	ML	VB	ML	VB	ML	VB	ML	VB	ML	VB	ML	VB	ML	VB
det	7.55	7.45	8.62	8.55	7.35	7.28	8.01	7.96	7.43	7.32	8.41	8.34	7.67	7.56	9.28	8.74
mm	6.66	6.63	6.85	6.84	6.72	6.69	6.82	6.80	6.59	**6.54**	6.76	6.76	6.61	6.57	7.23	6.92

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.

5.3. Results

Table 1 shows the source separation performance of VB compared to that of the state-of-the-art ML method in [5]. ML and VB perform similarly in almost all of the eight configurations. However VB is 0.2 dB better than the state-of-the-art ML method and yields the best performance for the 1 point source, spectrally and temporally constrained model (1-co-co) with 2.5 dB SDR. The baseline binary masking method [9] 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 both VB and ML algorithms. As one can see, VB based MFCC estimation performs 0.05 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 performs best for the RMS error by 6.54 dB for the 1 point source, spectrally unconstrained, temporally constrained model (1-un-co). The baseline binary masking method [9] performs significantly worse than both VB and ML algorithms and yields 17.25 dB RMS error.

6. CONCLUSION

In this paper, we presented a general, fully Bayesian source separation algorithm and 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 standard ML method, proving that the ML method provides a reasonable approximation, but that it is possible to obtain more accurate estimates.

In the future, we will seek to improve the tightness of the variational bound (and thereby the accuracy of the resulting features) and exploit the variance of the estimated features for uncertainty decoding.

7. REFERENCES

[1] 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, Article ID 651420, 2010.

[2] 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.

[3] 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.

[4] 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)*, 2011, pp. 1663–1667.

[5] 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*, to appear, http://hal.inria.fr/inria-00536917/PDF/RR-7453.pdf.

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

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

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

[9] 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)*, 2011, pp. 2616–2619.

[10] 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.