An Uncertainty Estimation Approach for the Extraction of Source Features in Multisource Recordings

Kamil Adiloglu, Emmanuel Vincent

AN UNCERTAINTY ESTIMATION APPROACH FOR THE EXTRACTION OF SOURCE FEATURES IN MULTISOURCE RECORDINGS

Kamil Adiloğlu, and Emmanuel Vincent

INRIA, Centre de Rennes - Bretagne Atlantique
Campus de Beaulieu, 35042, Rennes cedex, France
phone: + (33) 2 9984 7227, fax: + (33) 2 9984 7171, email: {kamil.adiloglu, emmanuel.vincent}@inria.fr

ABSTRACT

We consider the extraction of individual source features from a multisource audio recording by combining source separation with feature extraction. The main issue is then 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 were designed for scenarios involving one prominent source plus background noise, we focus on under-determined mixtures involving several sources of interest. We apply either Gibbs sampling or variational Bayes to estimate the posterior probability of the sources and subsequently derive the expectation of the features either by sampling or by moment matching. Experiments over stereo mixtures of three sources show that variational Bayes followed by either feature sampling or moment matching provides the best results for convolutive mixtures, while no improvement is obtained on instantaneous mixtures compared to deterministic feature computation.

1. INTRODUCTION

Audio information retrieval (AIR) covers a wide range of applications, from speech/speaker recognition to music genre classification. All these applications are typically achieved by extracting features describing the audio content and exploiting them for *e.g.* classification. 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 example, the separation of target speech from background noise is known to improve speech recognition performance [9], while the separation of harmonic and percussive sounds improves music genre classification [11].

The problem at hand is then to robustly estimate the features of each source despite source separation errors. This can be achieved in a probabilistic framework by estimating the uncertainty over the separated source signals and propagating it to the features. Because source separation is usually achieved in the time-frequency domain, the uncertainty must be expressed in that domain. The pioneering approach of Cooke et al. [3, 6] relied on a binary uncertainty model, where each time-frequency coefficient of each source was considered either as reliable or not reliable. Deng et al. [4] and Kolossa et al. [9] proposed a more flexible Gaussian model, enabling more precise quantification of the uncertainty in terms of the posterior probability of each time-frequency coefficient. However, their studies relied on specific source separation algorithms designed for mixtures of one prominent source plus background noise.

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

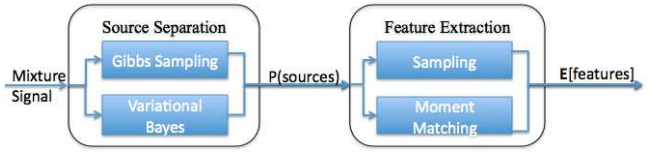

Figure 1: Flow of the proposed probabilistic source separation and feature extraction approach.

In this paper, we address the problem of individual source feature extraction from under-determined mixtures involving several sources of interest, such as *e.g.* polyphonic music recordings. As a preliminary study, we concentrate on the local Gaussian model-based separation [14], and extract the Mel Frequency Cepstral Coefficients (MFCCs) [10] of each source, which are the most popular features for AIR. Figure 1 shows the structure of the proposed approach. We estimate the posterior distribution of the source time-frequency coefficients either by Gibbs sampling or variational Bayes and subsequently derive the expectation of the features either by sampling or by moment matching.

Except for the latter, these techniques have not previously been used in this context to our knowledge. Cemgil et al. [2] applied Gibbs sampling and variational Bayes to source separation, but focused on evaluating performance on instantaneous mixtures and did not consider feature extraction nor convolutive mixtures. Also, somewhat surprisingly, they did not compare performance with Maximum Likelihood (ML) inference. Kolossa et al. [9] used moment matching to propagate uncertainty for MFCC extraction, but more powerful sampling techniques have not been exploited so far.

This paper is organized as follows. Section 2 introduces the source separation framework and the two inference algorithms for the estimation of the posterior distribution of the sources. Section 3 presents the feature extraction and uncertainty propagation algorithms. In Section 4, we evaluate these algorithms in combination with each other over instantaneous or convolutive mixtures. We conclude in Section 5.

2. UNCERTAINTY ESTIMATION IN SOURCE SEPARATION

Classically, we adress the source separation problem in the time-frequency domain by means of the Short-Time Fourier Transform (STFT). The mixing equation is given by

$$\mathbf{x}_{ft} = \mathbf{A}_f \mathbf{s}_{ft} + \boldsymbol{\varepsilon}_{ft}, \tag{1}$$

where $\mathbf{x}_{ft} = [x_{1,ft}, \ldots, x_{I,ft}]^T$ denotes the mixture STFT coefficients, $\mathbf{s}_{ft} = [s_{1,ft}, \ldots, s_{J,ft}]^T$ the source STFT coeffi-

cients, $\mathbf{A}_f = [a_{ijf}]_{ij} \in \mathbb{C}^{I \times J}$ the mixing system and $\boldsymbol{\varepsilon}_{ft}$ the noise. In this formulation, $f = 1,\ldots,F$ is the fequency index, $t = 1,\ldots,T$ the time frame index, $i = 1,\ldots,I$ the channel index and $j = 1,\ldots,J$ the source index. We focus on the under-determined scenario, for which $I < J$.

In this preliminary study, we assume that the mixing system $\mathbf{A}_f$ is known and adopt a local Gaussian model [14] for the source coefficients. As a consequence, the source coefficients can be separately estimated in each time-frequency bin. Therefore, we omit the indices f, t in the rest of this section for the sake of readability.

We set a zero-mean Gaussian prior over the source coefficients $\mathbf{s}$ with variance $\mathbf{v}$:

$$\mathbf{s} \sim \mathcal{N}(0, \mathbf{v}). \tag{2}$$

We assume independence between the sources so that the covariance matrix takes the diagonal form $\mathbf{v} = \mathbf{diag}([v_j]_{j=1\cdots J})$. We then set an inverse-gamma prior over each source variance v_j with shape parameter α_j and scale parameter $\beta_j = 0$:

$$v_j \sim \mathcal{IG}(\alpha_j, 0). \tag{3}$$

This prior is conjugate to the Gaussian likelihood and governs the shape of the variance, where $\alpha_j = 0$ corresponds to the scale-invariant Jeffreys prior. Finally, we assume that the noise $\boldsymbol{\varepsilon}_{ft}$ is Gaussian with fixed covariance matrix $\mathbf{E}$.

2.1 Maximum Likelihood

The state-of-the-art ML approach consists of estimating the source variances by means of the Expectation-Maximization (EM) algorithm in [5] then deriving the source coefficients by multichannel Wiener filtering.

Defining the complete data for each time-frequency bin to be $\{\mathbf{x}_{fn}, \mathbf{s}_{fn}\}$ yields the following updates:

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}, \tag{4}$$
$$\boldsymbol{\Sigma} = (\mathbf{I} - \mathbf{W}\mathbf{A})\mathbf{v}, \tag{5}$$
$$v_j = |\hat{\mathbf{s}}|^2 - \boldsymbol{\Sigma}_{jj}, \tag{6}$$

where

$$\mathbf{W} = \mathbf{v}\mathbf{A}^{\mathbf{H}}(\mathbf{A}\mathbf{v}\mathbf{A}^{\mathbf{H}} + \mathbf{E})^{-\mathbf{1}} \tag{7}$$

is the Wiener filter and $\mathbf{I}$ is the identity matrix. This approach basically gives a point estimate $\hat{\mathbf{s}}$ of the source coefficients.

We aim to estimate the posterior probability of the source coefficients, which is given by

$$p(\mathbf{s}|\mathbf{x}) \propto \int p(\mathbf{x}|\mathbf{s},\mathbf{v})p(\mathbf{s}|\mathbf{v})p(\mathbf{v})d\mathbf{v}. \tag{8}$$

In order to approximate this integral, which is intractable, we now consider two state-of-the-art Bayesian approximation techniques, namely Gibbs sampling and variational Bayes.

2.2 Gibbs Sampling

Gibbs sampling [1] aims to approximate the joint posterior $p(\mathbf{s},\mathbf{v}|\mathbf{x})$ by a collection of R samples $(\mathbf{s}^{(r)},\mathbf{v}^{(r)})$ obtained by alternatingly sampling from the conditional distributions of the variables. It has been shown that these samples converge to the joint distribution, provided that the resulting Markov chain is reversible. The marginal distribution $p(\mathbf{s}|\mathbf{x})$ is then represented by the samples $\mathbf{s}^{(r)}$.

The conditional distribution of the source coefficients boils down to the Wiener filter again

$$p(\mathbf{s}|\mathbf{x},\mathbf{v}) = \mathcal{N}(\mathbf{s}; \mathbf{W}\mathbf{x}, (\mathbf{I} - \mathbf{W}\mathbf{A})\mathbf{v}), \tag{9}$$

where $\mathbf{W}$ is given in (7).

Due to conjugacy rules, the conditional distribution of the variance of each source is again an inverse-Gamma distribution:

$$p(v_j|s_j) = \mathcal{IG}(v_j; \alpha_j + 1, |s_j|^2). \tag{10}$$

Consequently, in order to approximate $p(\mathbf{s},\mathbf{v}|\mathbf{x})$, each iteration r of our Gibbs sampling scheme consists in:

1. Sampling the source variance $v_j^{(r)}$ for $j = 1\ldots J$ from $p(v_j^{(r)}|s_j^{(r-1)})$,
2. Sampling the source coefficients $\mathbf{s}^{(r)}$ from $p(\mathbf{s}^{(r)}|\mathbf{x},\mathbf{v}^{(r)})$.

2.3 Variational Bayes

Variational Bayes [1] is an alternative inference approach, which minimizes the Kullback-Leibler (KL) divergence between the true posterior distribution $p(\mathbf{s},\mathbf{v}|\mathbf{x})$ and some approximation $q(\mathbf{s},\mathbf{v})$, which is typically specified by assuming some factorization. We assume the following factorization:

$$q(\mathbf{s},\mathbf{v}) = q(\mathbf{s})\prod_{j=1}^{J} q(v_j). \tag{11}$$

The factor distributions minimizing the KL divergence can be obtained by taking the expectation of the joint log-distribution of all variables over the other factors [1].

Skipping the necessary maths due to space limitations, we obtain the following optimal factors:

$$q^*(\mathbf{s}) \sim \mathcal{N}_c(\mathbf{s}; \mu, \boldsymbol{\Sigma}), \tag{12}$$
$$q^*(v_j) \sim \mathcal{IG}(v_j; \alpha_j + 1, \beta_j), \tag{13}$$

where

$$\mu = \mathbf{W}\mathbf{x}, \tag{14}$$
$$\boldsymbol{\Sigma} = (\mathbf{I} - \mathbf{W}\mathbf{A})\beta, \tag{15}$$
$$\beta_j = |\mu_j|^2 + \Sigma_{jj}. \tag{16}$$

β is given by $\beta = \mathrm{diag}([\beta_j]_{j=1\ldots J})$, μ_j is the j^{th} element of μ, $\mathbf{I}$ is the identity matrix and the Wiener filter $\mathbf{W}$ writes

$$\mathbf{W} = \beta\mathbf{A}^{\mathbf{H}}(\mathbf{A}\beta\mathbf{A}^{\mathbf{H}} + \mathbf{E})^{-\mathbf{1}}. \tag{17}$$

(14), (15), (16) and (17) depend on each other. After proper initialization of the factors, they are an iterative optimized by cycling through these equations and by replacing the dependent values with their new estimates.

To our surprize, even though the ML approach and our variational Bayes solution optimize different criteria, the update equations for the mean of the source coefficients of both methods are identical up to the change of variables $\mathbf{v} = \beta$ and $\hat{s} = \mu$. However our approach also provides the covariance of the source coefficients.

Note also that, despite the similarity between the update equations of Gibbs sampling and variational Bayes, these two inference methods provide different results. In particular, Gibbs sampling better approximates multimodal posteriors, while variational Bayes is computationally faster.

3. UNCERTAINTY PROPAGATION FOR FEATURE EXTRACTION

Once the posterior distribution of the source coefficients s_{jft} in each time-frequency bin has been computed, we calculate the expectation of the MFCCs separately for each source as

$$\mu_{jt}^{\mathbf{MFCC}} = \int \mathrm{MFCC}(\mathbf{S}_{jt}) P(\mathbf{S}_{jt})\, d\mathbf{S}_{jt} \tag{18}$$

where $\mathbf{S}_{jt} = [s_{jft}]_{f=1\ldots F}$ are the STFT coefficients of source j in time frame t. For ML, this boils down to $\mathrm{MFCC}(\hat{\mathbf{S}}_{jt})$.

The calculation of the MFCCs for a given complex-valued spectrum $\mathbf{S}_{jt}$ consists of taking the magnitude $|\mathbf{S}_{jt}|$ of the spectrum, filtering it by a set of triangular filters spanning the Mel frequency scale and taking the discrete cosine transform (DCT) of the filtered spectrum [10]:

$$\mathrm{MFCC}(\mathbf{S}_{jt}) = 20\,\mathbf{D}\log_{10}(\mathbf{M}|\mathbf{S}_{jt}|). \tag{19}$$

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).

We now present two methods to propagate the uncertainty over the source coefficients to the MFCCs based on sampling and moment matching. Each method is applicable both to the output of Gibbs sampling and variational Bayes.

3.1 MFCC Estimation by Sampling

For each source j and time frame t, we sample L complex-valued source spectra $\mathbf{S}_{jt}^{(l)}$ and calculate the MFCCs using (19). Then, we compute the mean of these MFCCs as

$$\mu_{jt}^{\mathbf{MFCC}} = \frac{1}{L}\sum_{l=1}^{L} \mathrm{MFCC}(\mathbf{S}_{jt}^{(l)}). \tag{20}$$

For the Gibbs sampling, the sample spectra $\mathbf{S}_{jt}^{(l)}$ are obtained by randomly selecting one sample $s_{jft}^{(r)}$ for each frequency bin f, which we have saved during the Gibbs sampling. For the variational Bayes, we sample from the estimated posterior of the source coefficients shown in (12).

3.2 MFCC Estimation by Moment Matching

The moment matching approach propagates the uncertainty expressed by the means and variances of the posterior distributions of the estimated source coefficients through the calculation of the MFCCs. These means and variances are readily given when using variational Bayes or can be estimated from the source samples when using Gibbs sampling.

The absolute value of a complex Gaussian random variable, hence the magnitude spectrum of the estimated source coefficients follows a Rice distribution. For source j in time frame t, the mean and variance of the magnitude spectrum is given by the first and second raw moments of the Rice distribution as: $\mu^{|\mathbf{s}|} = \mathbb{E}[|\mathbf{S}_{j,t}|]$ and $\Sigma^{|\mathbf{s}|} = \mathbb{E}[|\mathbf{S}_{j,t}|^2] - (\mu^{|\mathbf{s}|})^2$ [9].

The mel-filtering of the magnitude spectrum is a linear transformation. So, we can simply match the moments:

$$\mu^{\mathrm{MEL}} = \mathbf{M}\,\mu^{|\mathbf{s}|}, \tag{21}$$

$$\Sigma^{\mathrm{MEL}} = \mathbf{M}\,\Sigma^{|\mathbf{s}|}\,\mathbf{M}^T. \tag{22}$$

The logarithm in the calculation of the MFCCs is not a linear transformation. Here, we assume the log-normality of the MEL features. Incorporating the log-normal transformation proposed by Gales [8], the i-th coefficient is given by:

$$\mu_i^{\log} = 20\log_{10}(\mu_i^{\mathrm{MEL}}) - 10\log_{10}\left(\frac{\Sigma_{ii}^{\mathrm{MEL}}}{(\mu_i^{\mathrm{MEL}})^2} + 1\right). \tag{23}$$

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

$$\mu_{jt}^{\mathbf{MFCC}} = \mathbf{D}\,\mu^{\log}. \tag{24}$$

4. EXPERIMENTAL EVALUATION

We now evaluate the uncertainty estimation and propagation algorithms proposed above in combination with each other.

4.1 Data

We considered a dataset of 20 stereo mixtures of 3 sources of 10 s duration, including 10 instantaneous mixtures and 10 convolutive mixtures with 250 ms reverberation time. Both types of mixtures were generated from the same 30 source signals, including 15 music and 15 speech signals. The source signals and the mixing filters were taken from the BSS Oracle toolbox [12]. No noise was added. The mixing system $\mathbf{A}_f$ in each frequency bin f was either given by the mixing matrix in the instantaneous case or by the Fourier transform of the mixing filters in the convolutive case.

4.2 Algorithmic Settings

For the variational Bayes method, we used the non-informative Jeffreys prior for the source variances by setting the shape α_j to zero. Furthermore, we assigned the noise covariance $\mathbf{E}$ to zero. These settings are natural given that the mixture does not include any noise and that no prior information is available about the source variances. We iterated the algorithm 50 times at maximum.

These settings cannot be used for Gibbs sampling, otherwise the resulting Markov chain becomes irreversible. If $v_j = 0$, the chain remains stuck in $v_j = 0$ for subsequent iterations. Therefore we resorted to the noisy setting by assigning the noise covariance to $3 \times 10^{-6}\mathbf{I}$. Furthermore, we observed better mixing properties when setting the shape value α of the prior to a negative value. Preliminary experiments with different shape values led us to set the optimal shape value in terms of mean MFCC error to $\alpha = -0.13$. Note that this yields a proper inverse-gamma posterior with a shape value $\alpha = 0.87$.

In the instantaneous case, we generated 100 burn-in samples followed by $R = 400$ samples, which we used for feature extraction. In the convolutive case, we generated 200 burn-in samples followed by $R = 800$ samples.

In the feature extraction step, we calculated the first 20 MFCCs. For MFCC estimation by sampling, we generated $L = 2000$ sample spectra for each time frame t.

4.3 Evaluation Criteria

In order to assess the impact of source separation on feature extraction, we evaluate the proposed algorithms according to both tasks. Source separation quality is evaluated in terms

of the Signal-to-Distortion Ratio (SDR) in [13] between the mean of the estimated and the true source signals.

Feature extraction accuracy is evaluated in terms of the error between the estimated $\mu_{jt}^{\mathbf{MFCC}}$ and the true MFCCs $\mathrm{MFCC}(\mathbf{S}_{jt})$. This evaluation method enables us to assess the performance of the feature extraction without running costly classification experiments with these features. Information retrieval techniques typically do not process the MFCCs on each time frame but average them over longer time intervals [7]. Hence, unless otherwise specified, we average the estimated MFCCs over 1 s intervals. The raw error for source j in time interval t is given by

$$\mathbf{d}_{jt} = \mu_{jt}^{\mathbf{MFCC}} - \mathrm{MFCC}(\mathbf{S}_{jt}). \tag{25}$$

This error function ignores the energy of the corresponding time interval. However, errors in silent time intervals are not crucial for information retrieval and can be ignored. Therefore, we weight the error considering the energy levels. We first calculate the Root Mean Square (RMS) amplitude of the true source in each time interval t as $e_{jt} = \sqrt{\frac{1}{F}\sum_{f=1}^{F}|s_{jft}|^2}$. We then define the weights as

$$\overline{e}_{jt} = \begin{cases} 1 & \text{if } e_{jt} \geq 0.1 \times \max_t e_{jt}, \\ 0 & \text{if } e_{jt} < 0.1 \times \max_t e_{jt} \end{cases} \tag{26}$$

Denoting the error over the n-th MFCC coefficient of source j in time frame t by d_{njt}, the total weighted RMS error over all sources is defined as

$$RMS = \sqrt{\frac{1}{N}\frac{\sum_n \sum_j \sum_t d_{njt}^2 \overline{e}_{jt}}{\sum_j \sum_t \overline{e}_{jt}}}. \tag{27}$$

In this equation and in the following, we assume that the source signals have been concatenated so that t spans all time frames of all mixtures in the test dataset. We decompose this total error into a bias term and a standard deviation term. The bias over the n-th coefficient is given by $b_{nj} = \frac{\sum_t \overline{e}_{njt} d_{njt}}{\sum_t \overline{e}_{jt}}$. We define the total bias for source j in the RMS sense by

$$b_j = \sqrt{\frac{1}{N}\sum_n b_{nj}^2}. \tag{28}$$

where the summation over n spans a certain subset of coefficients and N is the size of this subset. The standard deviation of the error over MFCC coefficient n of source j is given by $\sigma_{bnj} = \sqrt{\frac{\sum_t \overline{e}_{jt}(d_{njt}-b_{nj})^2}{\sum_t \overline{e}_{jt}}}$. Hence we compute the total standard deviation as

$$\sigma_j = \sqrt{\frac{1}{N}\sum_n \sigma_{bnj}^2}. \tag{29}$$

4.4 Results

Table 1 shows the source separation performance of Gibbs sampling and variational Bayes compared to that of the state-of-the-art ML method. Recall that the mean source coefficients estimated by variational Bayes and by ML provide the same performance. ML and variational Bayes increase SDR by 2.2 dB on average compared to Gibbs sampling for instantaneous mixtures, but decrease it by 0.6 dB for convolutive mixtures.

	s_1		s_2		s_3	
	Sam	ML/Var	Sam	ML/Var	Sam	ML/Var
Inst	12.6	**14.7**	3.6	**5.7**	16.8	**19.2**
Conv	**6.8**	6.3	**7.6**	6.6	**6.2**	5.9

Table 1: SDR in dB achieved by Gibbs sampling (Sam) and by ML or variational Bayes (ML/Var) source separation on instantaneous (Inst) and convolutive (Conv) mixtures.

Source separation	Feature extraction	MFCC 1		MFCCs 2 to 20	
		Inst	Conv	Inst	Conv
ML	Determin.	**19.4**	34.7	**2.4**	4.7
Gibbs Sam.	Determin.	30.2	**29.1**	4.2	4.9
	Sampling	43.0	38.4	5.9	4.7
	Moment	57.1	47.0	6.3	5.0
Var. Bayes	Sampling	27.7	45.7	2.8	**4.6**
	Moment	28.2	46.1	2.8	**4.6**

Table 2: Total RMS error in dB for the first MFCC and for MFCCs 2 to 20 obtained by ML, Gibbs sampling or variational Bayes-based source separation followed by deterministic, sampling-based or moment matching-based feature extraction over instantaneous and convolutive mixtures.

Table 2 shows the total RMS error in dBs of the three above source separation methods in combination with sampling-based and moment-matching based MFCC estimation. In addition, we show the RMS error of deterministic MFCC computation, that uses only the mean estimates of the source STFT coefficients. As one can see, the first MFCC coefficient is significantly misestimated by all methods. This coefficient reflects the energy level of the signal. Source separation methods often fail to estimate this quantity: for instance, ML promotes sparsity and, as a consequence, typically under-estimates the energy of the sources [14]. For the other MFCC coefficients, the error values are on the order of 2 to 5 dB. ML and variational Bayes followed by any of the feature extraction algorithms appear to outperform Gibbs sampling both on instantaneous and on convolutive mixtures. ML and deterministic MFCC computation perform slightly better on instantaneous mixtures, while variational Bayes followed by either sampling-based or moment matching-based MFCC computation performs slightly better on convolutive mixtures. Overall, moment matching does not significantly decrease feature extraction accuracy compared to sampling, while the former is on the order of 1000 times faster.

Interestingly, the fact that Gibbs sampling outperforms ML or variational Bayes in terms of source separation performance over convolutive mixtures does not translate into more accurate feature extraction. Instead, ML and variational Bayes lead to more accurate feature extraction over all mixtures, while they are on the order of 100 times faster. It is possible that Gibbs sampling could yield better results with a larger number of samples, however the computation overhead, which is already high would be infeasible.

Table 3 shows the bias and standard deviation of the MFCCs 2 to 20 achieved by variational Bayes followed by moment matching on convolutive mixtures, where the length of the time intervals over which MFCCs are averaged is equal either to 1 s as above or to one single time frame. Time aver-

	s_1		s_2		s_3	
	bias	std	bias	std	bias	std
Framewise	1.5	5.7	2.0	5.9	1.8	6.7
Time-averaged	1.2	3.8	1.5	4.1	2.1	4.9

Table 3: Bias and standard deviation in dB of the MFCCs 2 to 20 calculated framewise or averaged over 1 s using variational Bayes and moment matching on convolutive mixtures.

		ML		Variational Bayes			
		Determin.		Sampling		Moment	
		bias	std	bias	std	bias	std
Inst	s_1	**0.2**	**1.3**	0.4	1.4	0.4	1.6
	s_2	1.0	**3.1**	**0.8**	3.5	**0.8**	3.5
	s_3	**0.7**	**2.3**	0.9	2.7	0.9	2.8
Conv	s_1	1.7	**3.7**	**1.2**	3.8	**1.2**	3.8
	s_2	1.9	**3.9**	**1.5**	4.1	**1.5**	4.1
	s_3	2.9	**4.7**	**2.1**	4.9	**2.1**	5.0

Table 4: Bias and standard deviation in dB over the MFCCs 2 to 20 obtained by the best algorithms in Table 2 over instantaneous and convolutive mixtures.

aging of the MFCCs slightly affects the bias due to energy-based error weighting but, more importantly, it reduces the standard deviation by 1.8 dB. More generally, we could observe this behavior for all feature extraction methods and for all mixtures types.

Finally, Table 4 presents the bias and standard deviation of the MFCCs 2 to 20 for the best algorithms in Table 2. ML leads to smaller bias than variational Bayes on instantaneous mixtures and vice-versa on convolutive mixtures. However, ML followed by deterministic MFCC computation provides smaller standard deviation on all mixtures types. This suggests that the ML approach might provide the best results in terms of information retrieval, since classifiers are typically insensitive to bias. Care must be taken however that the bias over each source depends on the other sources and on the mixing setup. Further experiments are needed to validate this observation by determining the average bias and standard deviation for a given source, *e.g.* a given musical instrument, over a variety of mixtures.

5. CONCLUSION

In this paper, we presented two Bayesian source separation algorithms and two uncertainty propagation algorithms for the computation of the expectation of the MFCCs of individual sources within an under-determined mixture.

Surprisingly, the standard ML source separation algorithm followed by deterministic MFCC computation appears to provide the best feature extraction performance on instantaneous mixtures. However, variational Bayes followed by either sampling-based or moment matching-based MFCC computation provides slightly better results on convolutive mixtures. Gibbs sampling performs worse in both mixing cases despite being 100 times slower for the chosen number of samples.

In the future, we will adapt these methods to more realistic scenarios, where the mixing system is unknown, and to more complex source separation models as in [14]. We will also investigate the classification performance resulting from the estimated features.

REFERENCES

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

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

[3] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data", *Speech Communication*, vol. 34, pp. 267–285, 2001.

[4] 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.

[5] 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.

[6] J. Eggink and G.J. Brown, "Application of missing feature theory to the recognition of musical instruments in polyphonic audio", in *Proc. Int. Symp. on Music Information Retrieval (ISMIR)*, 2003.

[7] D.P.W. Ellis, C.V. Cotton, and M.I. Mandel, "Cross-correlation of beat-synchronous representations for music similarity", in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 57–60, 2008.

[8] M. J. F. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, PhD Thesis, University of Cambridge, UK, September 1995.

[9] 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, 13 pages, 2010.

[10] B. Logan "Mel frequency cepstral coefficients for music modeling", in *Proc. Int. Symp. on Music Information Retrieval (ISMIR)*, 2000.

[11] H. Rump, S. Miyabe, E. Tsunoo, N. Ono, and S. Sagayama "Autoregressive MFCC models for genre classification improved by harmonic-percussion separation " in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, pp.87–92, 2010.

[12] E. Vincent, R. Gribonval, M. D. Plumbley, *BSS Oracle Toolbox*, http://bass-db.gforge.inria.fr/bss_oracle/.

[13] 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, July 2006.

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