A fast EM algorithm for Gaussian model-based source separation

Joachim Thiemann, Emmanuel Vincent

A FAST EM ALGORITHM FOR GAUSSIAN MODEL-BASED SOURCE SEPARATION

Joachim Thiemann

CNRS, IRISA – UMR 6074
Campus de Beaulieu
35042 Rennes Cedex, France
joachim.thiemann@irisa.fr

Emmanuel Vincent

Inria, Centre de Nancy - Grand Est
615 rue du Jardin Botanique
54600 Villers-lès-Nancy, France
emmanuel.vincent@inria.fr

ABSTRACT

We consider the FASST framework for audio source separation, which models the sources by full-rank spatial covariance matrices and multilevel nonnegative matrix factorization (NMF) spectra. The computational cost of the expectation-maximization (EM) algorithm in [1] greatly increases with the number of channels. We present alternative EM updates using discrete hidden variables which exhibit a smaller cost. We evaluate the results on mixtures of speech and real-world environmental noise taken from our DEMAND database. The proposed algorithm is several orders of magnitude faster and it provides better separation quality for two-channel mixtures in low input signal-to-noise ratio (iSNR) conditions.

Index Terms— Audio source separation, FASST, EM algorithm, binary masking, DEMAND.

1. INTRODUCTION

We consider the problem of under-determined reverberant source separation, that is to separate the signals of J sources from a mixture recorded by an array of $I < J$ microphones. See [2, 3] for a review. Numerous techniques have been proposed, ranging from early multichannel spatial filtering and single-channel spectral modeling techniques [4, 5] to recent techniques jointly exploiting spatial and spectral cues [1, 6]. The general Gaussian model-based framework in [1], which underlies the FASST toolbox[1], is currently one of the most advanced frameworks which enables the modeling of reverberant sources by means of full-rank spatial covariance matrices and the enforcement of spectral constraints by means of multilevel nonnegative matrix factorization (NMF).

For historical reasons, most of these techniques have been designed or assessed in two-channel scenarios. Additional microphones can improve the separation performance by increasing the spatial resolution of the array, but they imply a greater computational cost. The increase of the computational cost as a function of the number of channels is especially dramatic (cubic) for Gaussian expectation-maximization (EM)-based techniques which consider the source signals as hidden data [1, 7, 8]. Based on the idea of binary masking [4], a range of techniques have been proposed that provide a smaller cost increase by assuming that a single source predominates in each time-frequency bin and by considering the corresponding source indexes as discrete hidden data [6, 9, 10]. In [11], the source separation algorithm by Duong [12] is investigated by an EM algorithm targeting this hidden data.

In this paper, we propose a binary activation EM (BAEM) algorithm for the FASST framework, which extends the algorithm in [11] to multichannel NMF modeling of the spectra instead of unconstrained spectra. This algorithm exhibits a quadratic cost increase as a function of the number of channels while retaining the latter desirable feature of FASST, enabling the enforcement of spectral or temporal constraints on the source short-term spectra. We also test the algorithm for source separation in a noisy environment and we consider a larger number of experimental conditions, such as the level of background noise and the number of channels.

The structure of the rest of the paper is as follows. We recall the conventional EM algorithm for FASST in Section 2 and describe the BAEM algorithm as it fits within the FASST framework in Section 3. We evaluate them in Section 4 and conclude in Section 5.

2. GAUSSIAN MODEL-BASED SOURCE SEPARATION

2.1. Model

Using the Short-Time Fourier Transform (STFT), the $I \times 1$ vector $\mathbf{x}_{fn}$ of mixture STFT coefficients in time frame n and frequency bin f can be expressed as [3]

$$\mathbf{x}_{fn} = \sum_{j=1}^{J} \mathbf{y}_{j,fn} + \mathbf{b}_{fn}. \quad (1)$$

where $\mathbf{y}_{j,fn}$ is the *spatial image* of the jth source and $\mathbf{b}_{fn}$ is a small Gaussian noise with covariance $\mathbf{\Sigma}_{\mathbf{b},fn}$. FASST models

[1] *Flexible Audio Source Separation Toolbox,* a MATLAB toolbox which can be found at http://bass-db.gforge.inria.fr/fasst/

the spatial images of all sources as Gaussian random vectors

$$\mathbf{y}_{j,fn} \sim \mathcal{N}(\mathbf{0}, v_{j,fn}\mathbf{R}_{j,f}) \tag{2}$$

where $v_{j,fn}$ denotes the short-term *power spectrum* of the jth source and $\mathbf{R}_{j,f}$ its *spatial covariance matrix*. The short-term power spectra $\mathbf{V}_j = [v_{j,fn}]_{f,n}$ of each source are further assumed to factor in a multilevel NMF fashion as

$$\mathbf{V}_j = \left(\mathbf{W}_j^{\text{ex}}\,\mathbf{U}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}\,\mathbf{H}_j^{\text{ex}}\right) \odot \left(\mathbf{W}_j^{\text{ft}}\,\mathbf{U}_j^{\text{ft}}\,\mathbf{G}_j^{\text{ft}}\,\mathbf{H}_j^{\text{ft}}\right) \tag{3}$$

where the nonnegative matrices $\mathbf{W}_j^{\text{ex}}$, $\mathbf{U}_j^{\text{ex}}$, $\mathbf{G}_j^{\text{ex}}$ and $\mathbf{H}_j^{\text{ex}}$ encode the fine spectral structure, the spectral envelope, the temporal envelope and the temporal fine structure of the source excitation signal and $\mathbf{W}_j^{\text{ft}}$, $\mathbf{U}_j^{\text{ft}}$, $\mathbf{G}_j^{\text{ft}}$ and $\mathbf{H}_j^{\text{ft}}$ the same quantities for the spectral resonance filter, and $\odot$ denotes entry-wise matrix multiplication. In the following, we assume that the sources are reverberated or diffuse, so that $\mathbf{R}_{j,f}$ is full-rank.

2.2. Subsource EM algorithm

Given this model, the log-likelihood can be expressed in terms of the *empirical mixture covariance matrix* $\widehat{\mathbf{R}}_{\mathbf{x},fn}$ in each time-frequency bin. The classical approach for maximum likelihood (ML) inference in such Gaussian models is to employ the EM algorithm, where the source time-frequency coefficients themselves are considered as hidden data [7, 8].

This approach was applied to FASST by defining vectors of *subsource*[2] coefficients $\mathbf{s}_{j,fn}$ such that $\mathbf{y}_{j,fn} = \mathbf{A}_{j,f}\,\mathbf{s}_{j,fn}$ with $\mathbf{A}_{j,f}\mathbf{A}_{j,f}^H = \mathbf{R}_{j,f}$ and by considering them as hidden data [1]. The resulting subsource EM (SSEM) updates are summarized in Algorithm 1 and Figure 1. The E-step relies on the computation of the Wiener filter $\mathbf{\Omega}_{\mathbf{s},fn}$, while the M-step involves a closed-form update for $\mathbf{A}_f = [\mathbf{A}_{1,f}, \ldots, \mathbf{A}_{J,f}]$ and multiplicative updates for the multilevel NMF parameters. Strictly speaking, this is a generalized EM (GEM) algorithm because the multiplicative updates increase but do not maximize the log-likelihood of the complete data.

Once the parameters $\theta = \{\mathbf{R}_{j,f}, \mathbf{W}_j^{\text{ex}}, \mathbf{U}_j^{\text{ex}}, \mathbf{G}_j^{\text{ex}}, \mathbf{H}_j^{\text{ex}}, \mathbf{W}_j^{\text{ft}}, \mathbf{U}_j^{\text{ft}}, \mathbf{G}_j^{\text{ft}}, \mathbf{H}_j^{\text{ft}}\}_{j,f}$ have been estimated, the source STFT coefficients are obtained via the Wiener filter

$$\hat{\mathbf{y}}_{j,fn} = v_{j,fn}\mathbf{R}_{j,f}\left(\sum\nolimits_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,f}\right)^{-1}\mathbf{x}_{fn}. \tag{4}$$

2.3. Complexity

Although it has been successfully employed for the separation of two-channel mixtures [1, 3], the SSEM algorithm has two drawbacks when the number of channels I increases. First, the computational cost of each iteration becomes dominated by that of the matrix multiplications and inversions in (5)–(8), which grows in $\mathcal{O}((J^2 + 6J + 2)I^3FN)$ for N time frames and F frequency bins. Second, we have found in preliminary experiments that the increase in the size of the hidden data space requires more iterations until convergence. Altogether, this results in a major increase of the computational cost.

[2] This concept generalizes that of "source" when $\mathbf{R}_{j,f}$ is full-rank.

[3] The eight matrices in (3) are updated in turn. Denoting by $\mathbf{C}_j$ the matrix to be updated, the factorization (3) can always be rewritten as $\mathbf{V}_j = (\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j) \odot \mathbf{E}_j$, where the nonnegative matrices $\mathbf{B}_j$, $\mathbf{D}_j$ and $\mathbf{E}_j$ are assumed to be fixed while $\mathbf{C}_j$ is updated. For example, if $\mathbf{C}_j = \mathbf{H}_j^{\text{ft}}$, then $\mathbf{B}_j = \mathbf{W}_j^{\text{ft}}\,\mathbf{U}_j^{\text{ft}}\,\mathbf{G}_j^{\text{ft}}$, $\mathbf{D}_j = \mathbf{I}$ and $\mathbf{E}_j = \mathbf{W}_j^{\text{ex}}\,\mathbf{U}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}\,\mathbf{H}_j^{\text{ex}}$.

Algorithm 1 SSEM algorithm for FASST [1].

E-step - Compute the full data statistics

$$\widehat{\mathbf{R}}_{\mathbf{xs},fn} = \widehat{\mathbf{R}}_{\mathbf{x},fn}\mathbf{\Omega}_{\mathbf{s},fn}^H \tag{5}$$

$$\widehat{\mathbf{R}}_{\mathbf{s},fn} = \mathbf{\Omega}_{\mathbf{s},fn}\widehat{\mathbf{R}}_{\mathbf{x},fn}\mathbf{\Omega}_{\mathbf{s},fn}^H + (\mathbf{I} - \mathbf{\Omega}_{\mathbf{s},fn}\mathbf{A}_f)\mathbf{\Sigma}_{\mathbf{s},fn} \tag{6}$$

where

$$\mathbf{\Omega}_{\mathbf{s},fn} = \mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_{fn}^H\mathbf{\Sigma}_{\mathbf{x},fn}^{-1} \tag{7}$$

$$\mathbf{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},fn} \tag{8}$$

$$\mathbf{\Sigma}_{\mathbf{s},fn} = \operatorname{diag}(\underbrace{v_{j,fn} \ldots v_{j,fn}}_{I\text{times}})_{j=1}^{J}. \tag{9}$$

M-step - Update the model parameters (see [1] for notation[3])

$$\mathbf{A}_f = \left(\sum\nolimits_n \widehat{\mathbf{R}}_{\mathbf{xs},fn}\right)\left(\sum\nolimits_n \widehat{\mathbf{R}}_{\mathbf{s},fn}\right)^{-1} \tag{10}$$

$$\mathbf{C}_j = \mathbf{C}_j \odot \frac{\mathbf{B}_j^T[\widehat{\mathbf{\Xi}}_j \odot \mathbf{E}_j^{\cdot -1} \odot (\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j)^{\cdot -2}]\mathbf{D}_j^T}{\mathbf{B}_j^T(\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j)^{\cdot -1}\mathbf{D}_j^T} \tag{11}$$

where $\widehat{\mathbf{\Xi}}_j = [\hat{\xi}_{j,fn}]_{f,n}$ with

$$\hat{\xi}_{j,fn} = \frac{1}{I}\sum\nolimits_{r=(j-1)I+1}^{jI}(\widehat{\mathbf{R}}_{\mathbf{s},fn})_{rr}. \tag{12}$$

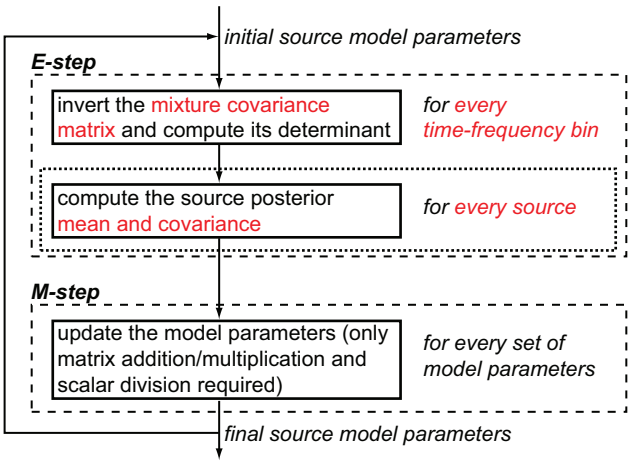

Fig. 1. Processing schema of SSEM.

3. BINARY ACTIVATION EM

In order to address these drawbacks, we propose here the faster BAEM algorithm for FASST under the assumption that a single source predominates in each time-frequency bin [11]. An approximate generative model is proposed for the mixture signal and a GEM algorithm is derived by considering the corresponding source indexes as discrete hidden data.

3.1. Model

Denoting by l_{fn} the index of the predominant source in time-frequency bin (f, n), we replace the generative model (1) by

$$\mathbf{x}_{fn} = \mathbf{y}_{l_{fn},fn} \quad \text{with} \quad l_{fn} \sim \mathrm{Cat}(\pi_{1,fn}, \ldots, \pi_{J,fn}) \quad (13)$$

where the source spatial image coefficients $\mathbf{y}_{j,fn}$ follow the same model as in (2)–(3) and Cat denotes the categorical distribution. The prior probability $\pi_{j,fn}$ that the jth source be predominant in this time-frequency bin may be set to zero in certain time frames for those sources which are known to be inactive and uniformly shared among the other sources.

3.2. Algorithm

We then conduct ML inference by considering l_{fn} as hidden data. The resulting BAEM updates [11] are summarized in Algorithm 2 and Figure 2. The E-step consists of computing the posterior probability $\gamma_{j,fn} = P(l_{fn} = j|\mathbf{x}, \theta)$ of l_{fn} and the M-step of solving the optimization problem $\theta = \arg\max \sum_{j,fn} \gamma_{j,fn}[-\operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{y},j,fn}^{-1}\widehat{\mathbf{R}}_{\mathbf{x},fn}) - \log\det(\pi\boldsymbol{\Sigma}_{\mathbf{y},j,fn})]$. For the source spatial covariance matrices $\mathbf{R}_{j,f}$, the update is obtained in closed form. For the multilevel NMF parameters, this optimization problem is equivalent to the weighted multilevel NMF problem $\arg\min \sum_{fn} \gamma_{j,fn} d_{IS}(\hat{\xi}_{j,fn}|v_{j,fn})$ where $d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1$ is the Itakura-Saito (IS) divergence and $\hat{\xi}_{j,fn}$ is defined in (18), which can be addressed using weighted multiplicative updates [13]. After convergence, the sources are separated by *soft masking*

$$\widehat{\mathbf{y}}_{j,fn} = \gamma_{j,fn}\, \mathbf{x}_{fn}. \quad (14)$$

3.3. Complexity

BAEM does not require matrix multiplications and inversions in each time-frequency bin anymore. By rewriting (15) as

$$\gamma_{j,fn} \propto \pi_{j,fn} \frac{e^{-\operatorname{tr}(\mathbf{R}_{j,f}^{-1}\widehat{\mathbf{R}}_{\mathbf{x},fn})/v_{j,fn}}}{(\pi v_{j,fn})^I \det(\mathbf{R}_{j,f})}, \quad (19)$$

it appears that the inverse and the determinant of $\mathbf{R}_{j,f}$ must be computed only once for each source in each frequency bin instead. For typical values of N, this has negligible cost and the cost of each iteration becomes dominated by that of the matrix dot product $\operatorname{tr}(\mathbf{R}_{j,f}^{-1}\widehat{\mathbf{R}}_{\mathbf{x},fn})$ in (19), which grows in $\mathcal{O}(JI^2FN)$. The complexity of each iteration is therefore reduced by a factor on the order of $(J+6)I$. Furthermore, we have found by monitoring the log-likelihood in preliminary experiments that the number of iterations needed to reach convergence is one order of magnitude smaller than for SSEM, which is a natural consequence of the drastic reduction of the size of the hidden data space (from $I \times J \times F \times N$ complex-valued variables to $F \times N$ discrete variables).

One limitation of BAEM compared to SSEM is that soft masking induces a smaller oracle upper bound [3] on the separation performance than multichannel Wiener filtering. However, we will see that this does not prevent BAEM from converging to better solutions than SSEM in certain conditions.

Algorithm 2 Proposed BAEM algorithm for FASST.

E-step - Compute the posterior probability of l_{fn}

$$\gamma_{j,fn} \propto \pi_{j,fn} \frac{e^{-\operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{y},j,fn}^{-1}\widehat{\mathbf{R}}_{\mathbf{x},fn})}}{\det(\pi\boldsymbol{\Sigma}_{\mathbf{y},j,fn})} \quad (15)$$

where $\boldsymbol{\Sigma}_{\mathbf{y},j,fn} = v_{j,fn}\mathbf{R}_{j,f}$.

M-step - Update the model parameters (see [1] for notation[3])

$$\mathbf{R}_{j,f} = \frac{\sum_n \gamma_{j,fn}\widehat{\mathbf{R}}_{\mathbf{x},fn}/v_{j,fn}}{\sum_n \gamma_{j,fn}} \quad (16)$$

$$\mathbf{C}_j = \mathbf{C}_j \odot \frac{\mathbf{B}_j^T[\boldsymbol{\Gamma}_j \odot \widehat{\boldsymbol{\Xi}}_j \odot \mathbf{E}_j^{\cdot -1} \odot (\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j)^{\cdot -2}]\mathbf{D}_j^T}{\mathbf{B}_j^T[\boldsymbol{\Gamma}_j \odot (\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j)^{\cdot -1}]\mathbf{D}_j^T} \quad (17)$$

where $\boldsymbol{\Gamma}_j = [\gamma_{j,fn}]_{f,n}$, $\widehat{\boldsymbol{\Xi}}_j = [\hat{\xi}_{j,fn}]_{f,n}$ and

$$\hat{\xi}_{j,fn} = \frac{1}{I}\operatorname{tr}(\mathbf{R}_{j,f}^{-1}\widehat{\mathbf{R}}_{\mathbf{x},fn}). \quad (18)$$

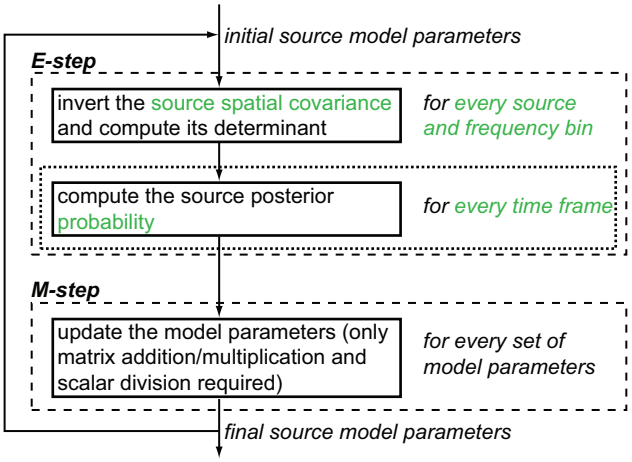

Fig. 2. Processing schema of BAEM.

4. EXPERIMENTAL EVALUATION

4.1. Acoustic setup

In our experimental setup, one or two target speech signals are convolved by room impulse responses simulated via the source image technique and added to field recordings of environmental noise, as classically done in source separation evaluations [3]. The publicly available database of 16-channel noise recordings (DEMAND) [14] is used. The microphone array setup, the room dimensions and the reverberation time in the simulated environment mirror those in the physical environments in which the DEMAND data was recorded.

Of the original 16 recorded channels, a subset corresponding to a planar array of 8 microphones in three staggered rows is used. The simulated target sources are placed 1 m away from microphone 1 of the array, separated by 45 degrees, in the same plane as the array, 1.5 m off the ground. Proper subsets of the multichannel data are also evaluated. In the two-channel case, only the two microphones of the center row are used, and in the four-channel case, the four microphones forming a diamond in the center of the array are used.

We use 5 s excerpts from 10 of the noise recordings in DEMAND: NFIELD, NPARK, OHALLWAY, OOFFICE, PRESTO, PSTATION, SPSQUARE, STRAFFIC, TCAR, and TMETRO. The simulated sources are mixed with the noise recordings to have an input SNR (iSNR) of -6, 0, 6, 12 and 18 dB for all channels combined. In the case of two targets, the iSNR is measured between the sum of the target sources and the noise recording. All signals are sampled at 16 kHz.

4.2. Algorithm parameters and evaluation criteria

The STFT window size is set to 1024 samples. The spectral variances of the noise source are constrained to NMF with 16 components. The spatial covariance matrices of the speech sources are initialized given knowledge of their position as in [12, eq. 12]. The other parameters are randomly initialized. Based on the convergence rates observed in initial experiments, we use 100 iterations per channel for SSEM and 10 iterations (regardless of the number of channels) for BAEM.

Altogether, BAEM is more than two orders of magnitude faster. In MATLAB, processing a single two-channel mixture takes 25 min with SSEM in contrast to 7.5 s with BAEM.

Separation quality is evaluated in terms of the average Signal-to-Interference Ratio (SIR) and Signal-to-Artifacts Ratio (SAR) [3] over all environments, where the background noise and the second source (if present) are both considered as interfering sources.

4.3. Results

The separation results for a single target in noise and for two targets in noise are shown in Figs. 3 and 4, respectively.

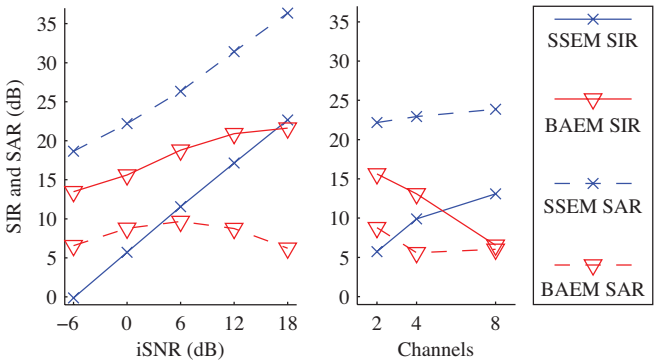

Fig. 3. Average SIR and SAR for the separation of one target in noise, as a function of iSNR with 2 channels (left) and of the number of channels at an iSNR of 0 dB (right).

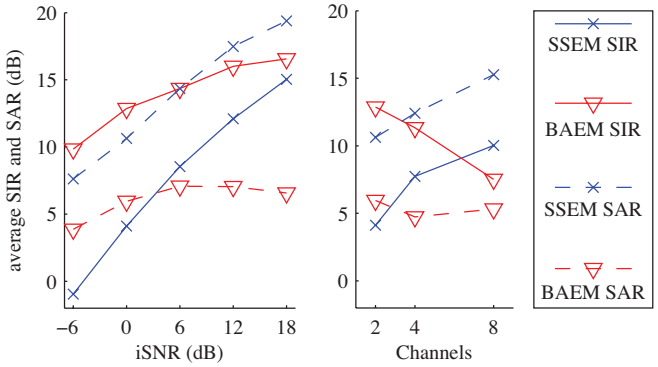

Fig. 4. Average SIR and SAR for the separation of two targets in noise, as a function of iSNR with 2 channels (left) and of the number of channels at an iSNR of 0 dB (right).

On two-channel mixtures, BAEM improves the SIR by 2 to 7 dB compared to SSEM when iSNR $\leq$ 0 dB. When the iSNR increases, the SIR increases and remains almost always superior to that of SSEM, while the SAR exhibits a stable, lower value. This is expected for soft masking reconstruction, which often comes close to producing a binary mask on the resulting signal. In such situtations, the SAR can be improved at the expense of a comparable reduction in SIR by temporal smoothing of the mask [15]. This post-processing should allow BAEM to achieve a higher overall quality than SSEM for iSNRs less than about 9 dB. Quality might further be improved by using SSEM with few iterations as a post-processing step for BAEM.

On mixtures of four or more channels, BAEM surprisingly shows a decrease in performance as the number of channels is increased. This might be due to the apparently lower robustness to noise of the updates (16) and (18) in BAEM compared to the updates (10) and (12) in SSEM, which could be addressed by appropriate regularization. Further study of this behavior is needed, however, due to the complex inter-

twining of all updates.

It is essential to note that this has no impact on the low-iSNR real-world applications that motivated this study. Indeed, applying BAEM to 2 channels out of the 8 available provides both the fastest estimation and the best overall quality for these applications.

5. CONCLUSION

We proposed a new EM algorithm for the FASST source separation framework, which greatly reduces its computational cost while retaining its ability of exploiting advanced source models. Even with few channels, this reduction is about one order of magnitude for each iteration whilst also requiring far fewer iterations. In addition, the overall separation quality of the proposed BAEM algorithm on two-channel mixtures is larger than that of the traditional SSEM algorithm in the low-iSNR real-world conditions that motivated this study. Further work will be devoted to the examination of numerical robustness issues, to temporal smoothing of the separation mask and to the investigation of the use of SSEM as a post-processing step for BAEM, so as to extend the observed quality improvement to higher iSNR conditions and to more channels.

6. ACKNOWLEDGMENTS

This work was supported by Canon Research Centre France. We wish to thank G. Kergourlay for helping with the design of the schemas in this paper.

7. REFERENCES

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

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

[3] E. Vincent, S. Araki, F. J. Theis, G. Nolte, et al., "The Signal Separation Evaluation Campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[4] Ö. Yılmaz and S. T. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[5] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems*, 2001, pp. 793–799.

[6] T. Nakatani, S. Araki, T. Yoshioka, and M. Fujimoto, "Joint unsupervised learning of hidden Markov source models and source location models for multichannel source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2011, pp. 237–240.

[7] J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon, "Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2002, pp. 561–564.

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

[9] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[10] A. Deleforge and R. P. Horaud, "A latently constrained mixture model for audio source separation and localization," in *Proc. 10th Int. Conf. on Latent Variable Analysis and Signal Separation*, 2012, pp. 372–379.

[11] R. Sakanashi, S. Miyabe, T. Yamada, and S. Makino, "Comparison of superimposition and sparse models in blind source separation by multichannel Wiener filter," in *Proc. APSIPA Annual Summit and Conf.*, 2012, pp. 1–6.

[12] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[13] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *Proc. ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, 2008, pp. 17–22.

[14] J. Thiemann, N. Ito, and E. Vincent, "The DEMAND database of multichannel environmental noise recordings," in *Proc. Int. Congress on Acoustics*, to appear.

[15] E. Vincent, "An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation," in *Proc. 9th Int. Conf. on Latent Variable Analysis and Signal Separation*, 2010, pp. 157–164.