ESTIMATION OF THE SPATIAL INFORMATION IN GAUSSIAN MODEL BASED AUDIO SOURCE SEPARATION USING WEIGHTED SPECTRAL BASES

Mahmoud Fakhry[1,2]

[1]Doctoral School of ICT
University of Trento
via Sommarive 5, 38123 Trento, Italy
m.fakhry@unitn.it

Piergiorgio Svaizer[2], *and Maurizio Omologo*[2]

[2]Center of Information Technology
Fondazione Bruno Kessler - irst
via Sommarive 18, 38123 Trento, Italy
{svaizer,omologo}@fbk.eu

ABSTRACT

In Gaussian model based audio source separation, source spatial images are modeled by Gaussian distributions. The covariance matrices of the distributions are represented by source variances and spatial covariance matrices. Accordingly, the likelihood of observed mixtures of independent source signals is parametrized by the variances and the covariance matrices. The separation is performed by estimating the parameters and applying multichannel Wiener filtering. Assuming that spectral basis matrices trained on source power spectra are available, this work proposes a method to estimate the parameters by maximizing the likelihood using Expectation-Maximization. In terms of normalization, the variances are estimated applying singular value decomposition. Furthermore, by building weighted matrices from vectors of the trained matrices, semi-supervised nonnegative matrix factorization is applied to estimate the spatial covariance matrices. The experimental results prove the efficiency of the proposed algorithm in reverberant environments.

***Index Terms*—** Spectral bases, nonnegative matrix factorization, spatial covariance matrix, audio source separation.

1. INTRODUCTION

To tackle the problem of blind source separation (BSS)[1], many algorithms have been proposed in the literature. Most of the algorithms work in the time-frequency domain through a short-time Fourier transform (STFT). In frequency-domain independent component analysis [2, 3] and clustering [4, 5], observed mixtures of source signals are modeled as the multiplication of complex spectra of the signals and complex-valued mixing vectors. In the under-determined mixing model, the source signals are obtained by first estimating the mixing vectors, and then applying binary masking [4], soft masking [5] or l_0-norm minimization [6]. Local Gaussian modeling of the mixing process [7, 8, 9] has lately emerged to tackle the source separation problem. Source spatial images in the observed mixtures are locally modeled by multivariate complex Gaussian distributions. The covariance matrices of the distributions are modeled as functions of spatial and spectral parameters. The audio channels from the location of a source to the positions of microphones are represented by a spatial covariance matrix, i.e. the spatial parameter of the model. Furthermore, each time-frequency point of the source power spectrum is represented by a scalar variance, i.e. the spectral parameter of the model. By assuming that the spatial images of the sources are statistically independent, the likelihood of the mixtures is a multivariate complex Gaussian distribution. The source signals are obtained by first estimating the parameters in the sense of maximum likelihood (ML), and then applying multichannel Wiener filtering. Non-negative matrix factorization (NMF) was involved in the model in [10, 11]. Applying NMF, the source variance can be represented as the product of two nonnegative vectors [12, 13], i.e. the source power spectrum is decomposed into two nonnegative matrices: a spectral basis matrix containing constitutive parts of the power spectrum, and a coefficient matrix containing time-varying weights. The factorization is achieved by optimizing a cost function using the widely used multiplicative update rules.

Using a prior knowledge has recently raised as a new trend to increase the performance of BSS. Source separation can benefit from information about the mixing environments [9, 14, 15], or the source signals [16, 17]. Prior information about the source variances can be used to guide the separation system. Assuming that spectral basis vectors trained on source power spectra, are available, in [16, 17] we proposed methods to estimate the spatial covariance matrix of the Gaussian model using the time-varying weights. However, the estimation does not exactly follow the changes in the amplitude values of the covariance matrix from one frequency to another. In this work we propose a method to follow these changes by calculating weighted basis vectors from the trained spectral basis vectors. Then both the weighted and original trained vectors are used to estimate the spatial covariance matrix. The trained spectral basis vectors can be assumed to be directly available as in [16], or it is supposed to obtain them through a

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

redundant library of spectral basis vectors. In the second scenario a detection step is needed to identify the basis vectors that have good match with the source signals in the observed mixtures as in [17]. The rest of the paper is organized as follows. In Section 2, we present the formulation and modelling of the problem. The proposed algorithm is explained in Section 3 and the experimental analysis and evaluation are reported in Section 4. Finally, Section 5 concludes the paper.

2. FORMULATION AND MODELLING

Assume that N sources are observed by an array of M microphones. Applying the discrete short time Fourier transform (STFT), at the frequency bin ω and the time frame l, a $M \times 1$ complex vector $\mathbf{x}(\omega, l)$ of the observed mixtures can be represented as the combination of N source spatial images $\mathbf{c}_n(\omega, l)$ such as

$$\mathbf{x}(\omega, l) = \sum_{n=1}^{N} \mathbf{c}_n(\omega, l). \qquad (1)$$

Over the total number of time frames L and frequency bins Ω, the vectors $\mathbf{c}_n(\omega, l)$ are assumed to be statistically independent, and probabilistically modeled by a zero-mean multivariate Gaussian distribution, with a $M \times M$ covariance matrix $\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)$

$$\mathbf{c}_n(\omega, l) \sim \mathcal{N}_c(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)), \qquad (2)$$

where $\mathbf{0}$ is a $M \times 1$ vector of zeros. Under the assumption that the source signals are statistically independent, the likelihood function of the observed mixtures $\mathbf{x}(\omega, l)$ is also a zero-mean multivariate complex Gaussian distribution with a covariance matrix obtained as

$$\mathbf{\Sigma}_{\mathbf{x}}(\omega, l) = \sum_{n=1}^{N} \mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l). \qquad (3)$$

Over all the time-frequency points, maximum likelihood estimation is shown to be the minimization of the minus log-likelihood as [10]

$$\xi(\theta) = \sum_{\omega, l} tr(\mathbf{\Sigma}_{\mathbf{x}}^{-1}(\omega, l)\tilde{\mathbf{R}}_{\mathbf{x}}(\omega, l)) + \log|\pi \mathbf{\Sigma}_{\mathbf{x}}(\omega, l)|, \qquad (4)$$

where $|.|$ denotes the determinant of a square matrix, $tr(.)$ indicates the trace of a matrix, $\theta = \{\mathbf{\Sigma}_{\mathbf{c}_1}(\omega, l), ..., \mathbf{\Sigma}_{\mathbf{c}_N}(\omega, l)\}_{\omega, l}$ is the set of model parameters, and $\tilde{\mathbf{R}}_{\mathbf{x}}(\omega, l)$ is a covariance matrix of the observed mixtures $\mathbf{x}(\omega, l)$ that can be empirically obtained in linear or quadratic forms as described in [11]. Source separation is performed by first estimating the set θ in the sense of ML. Then, an estimation of the source spatial images $\mathbf{c}_n(\omega, l)$ is obtained in the sense of minimum mean square error (MMSE) by applying multichannel Wiener filtering such as

$$\tilde{\mathbf{c}}_n(\omega, l) = \mathbf{G}_n(\omega, l)\mathbf{x}(\omega, l). \qquad (5)$$

The Wiener filter gain is computed as

$$\mathbf{G}_n(\omega, l) = \mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(\omega, l). \qquad (6)$$

The set θ is estimated by minimizing the criterion in (4) by using a generalized expectation maximization algorithm (GEM) [18] that consists in alternating the following two steps [8]:

1. *E step*, given the current estimate of θ and $\tilde{\mathbf{c}}_n(\omega, l)$, the conditional expectation of so-called natural statistics is computed as follows

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) = \tilde{\mathbf{c}}_n(\omega, l)\tilde{\mathbf{c}}_n^H(\omega, l) + (\mathbf{I} - \mathbf{G}_n(\omega, l))\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l), \qquad (7)$$

   where $\mathbf{I}$ is an $M \times M$ identity matrix.

2. *M step*, given $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, the set θ is updated according to the minimization of

$$\xi(\theta) = \sum_{\omega, l, n} tr(\mathbf{\Sigma}_{\mathbf{c}_n}^{-1}(\omega, l)\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)) + \log|\pi \mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)|. \qquad (8)$$

In the spatial covariance decomposition [7], the covariance matrix of the *n*-th source spatial images is modeled by a scalar variance $v_n(\omega, l)$, encoding the power spectrum of the source at each time-frequency point, and a $M \times M$ time-invariant spatial covariance matrix $\mathbf{R}_n(\omega)$, encoding the spatial information at each frequency bin. The covariance matrix is then represented as follows

$$\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l) = v_n(\omega, l)\mathbf{R}_n(\omega). \qquad (9)$$

For all the probability distributions of the source spatial images in the observed mixtures, the set of model parameters is redefined as follows

$$\theta = \{\{v_1(\omega, l), ..., v_N(\omega, l)\}_{\omega, l}, \{\mathbf{R}_1(\omega), ..., \mathbf{R}_N(\omega)\}_{\omega}\}. \qquad (10)$$

3. PROPOSED ALGORITHM

The computation of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ in (7) can be modified in order to include additional information about the local correlation between propagation channels, which often increases the accuracy of estimation as

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) = \hat{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) + (\mathbf{I} - \mathbf{G}_n(\omega, l))\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l), \qquad (11)$$

where $\hat{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ is the empirical covariance matrix of source spatial images which is obtained such as

$$\hat{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) = \frac{\sum_{\tilde{\omega}, \tilde{l}} \gamma(\tilde{\omega} - \omega, \tilde{l} - l)\tilde{\mathbf{c}}_n(\tilde{\omega}, \tilde{l})\tilde{\mathbf{c}}_n^H(\tilde{\omega}, \tilde{l})}{\sum_{\tilde{\omega}, \tilde{l}} \gamma(\tilde{\omega} - \omega, \tilde{l} - l)}, \qquad (12)$$

where γ is a bi-dimensional window describing the shape of neighbourhood. By substitution, up to a constant, the minimization function in (8) can be expressed in terms of the parameters of the spatial covariance decomposition in (9) as follows

$$\xi(\theta) = \sum_{\omega,l,n} tr({v_n}^{-1}(\omega,l){\mathbf{R}_n}^{-1}(\omega)\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)). \tag{13}$$

The set θ can be estimated in a blind scenario as in [8], or in an informed scenario as in [16]. Following the second scenario, as we will see later, in this work we mitigate a weakness point in [16]. On the other side, $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$ is a matrix with a high condition number (the ratio of the largest singular value to the smallest one). Hence by computing the singular value decomposition (SVD), the matrix can be approximately represented by its largest singular value $\sigma_{n1}(\omega,l)$ such as

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l) \approx \sigma_{n1}(\omega,l)\mathbf{A}_{n1}(\omega,l), \tag{14}$$

where $\mathbf{A}_{n1}(\omega,l)$ is an $M \times M$ unitary matrix associated with $\sigma_{n1}(\omega,l)$. Accordingly, the minimization problem in (13) can be respecified as

$$\xi(\theta) = \sum_{\omega,l,n} tr({v_n}^{-1}(\omega,l){\mathbf{R}_n}^{-1}(\omega)\sigma_{n1}(\omega,l)\mathbf{A}_{n1}(\omega,l)). \tag{15}$$

Considering the minimization problem in (15), in these sense of ML, if the source variance is estimated as [16]

$$v_n(\omega,l) = \sigma_{n1}(\omega,l), \tag{16}$$

the spatial covariance matrix can be obtained as follows

$$\mathbf{R}_n(\omega) = \frac{1}{L}\sum_l \frac{\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)}{v_n(\omega,l)}. \tag{17}$$

The estimation step can be extended by factorizing both absolute information of the numerator and the denominator by applying non-negative matrix factorization in a semi-supervised scenario. The weakness point in [16] is that the factorization is performed using the same pre-trained spectral basis vector. As a result, important absolute information is lost from one frequency to another. To mitigate this weakness, in this work, the estimated source variance $v_n(\omega,l)$ is decomposed using the pre-trained spectral basis vector, and the absolute value of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$ is decomposed using weighted copies of the pre-trained spectral basis vector. Accordingly, as we will see later, the matrix $\mathbf{R}_n(\omega)$ is estimated using both absolute information in the factorization domain and phase information in the time-frequency domain.

3.1. Estimation of the spatial covariance matrix

For clean training audio signals of the *n*-th source, the power spectra of several signal utterances are concatenated in one matrix. The spectral basis matrix $\mathbf{U}_n = [\mathbf{u}_n^T(\omega)]_{\Omega\times K}$ is extracted using the multiplicative update rule of minimizing the Kullback-Leibler (KL) divergence [13], where $\mathbf{u}_n(\omega)$ is a spectral basis vector of length K. The *n*-th estimated source power spectrum $\mathbf{V}_n = [v_n(\omega,l)]_{\Omega\times L}$ in (16) can be factorized using $\mathbf{U}_n$ to compute a coefficient matrix $\mathbf{W}_n = [\mathbf{w}_n(l)]_{K\times L}$ that contains time-varying weight vectors $\mathbf{w}_n(l)$ each of length K. Given $\mathbf{u}_n(\omega)$, the estimated source variance $v_n(\omega,l)$ can be represented in the factorization domain as follows

$$v_n(\omega,l) = \mathbf{u}_n^T(\omega)\mathbf{w}_n(l). \tag{18}$$

The (m_1,m_2) coefficient of $\mathbf{R}_n(\omega)$ can be represented in terms of the (m_1,m_2) coefficient of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$ as

$$r_n^{m_1m_2}(\omega) = \frac{1}{L}\sum_l \frac{\tilde{r}_{\mathbf{c}_n}^{m_1m_2}(\omega,l)}{v_n(\omega,l)}. \tag{19}$$

Let us factorize the absolute value of $\tilde{r}_{\mathbf{c}_n}^{m_1m_2}(\omega,l)$ using a time-varying vector $\mathbf{h}_n(\omega,l)$ of length K that is called the weighted spectral vector, as follows

$$r_n^{m_1m_2}(\omega) = \frac{1}{L}\sum_l \frac{\mathbf{h}_n^T(\omega,l)\mathbf{q}_{\mathbf{c}_n}^{m_1m_2}(l)}{v_n(\omega,l)}\angle\tilde{r}_{\mathbf{c}_n}^{m_1m_2}(\omega,l), \tag{20}$$

where $\mathbf{q}_{\mathbf{c}_n}^{m_1m_2}(l)$ is the (m_1,m_2) coefficient vector of the absolute value of $\tilde{r}_{\mathbf{c}_n}^{m_1m_2}(\omega,l)$, and $\angle\tilde{r}_{\mathbf{c}_n}^{m_1m_2}(\omega,l)$ is the phase information of $\tilde{r}_{\mathbf{c}_n}^{m_1m_2}(\omega,l)$. Let us describe the vector $\mathbf{h}_n(\omega,l)$ as follows

$$\mathbf{h}_n^T(\omega,l) = v_n(\omega,l)\mathbf{w}_n^T(l). \tag{21}$$

Accordingly, the (m_1,m_2) coefficient of the matrix $\mathbf{R}_n(\omega)$ in (20) is estimated as

$$r_n^{m_1m_2}(\omega) = \frac{1}{L}\sum_l \mathbf{w}_n^T(l)\mathbf{q}_{\mathbf{c}_n}^{m_1m_2}(l)\angle\tilde{r}_{\mathbf{c}_n}^{m_1m_2}(\omega,l). \tag{22}$$

Then the estimated $r_n^{m_1m_2}(\omega)$ coefficients, m_1, m_2= $1,...,M$ are arranged in the matrix $\mathbf{R}_n(\omega)$ that is normalized using its largest singular value. On the other hand, regarding the equations (18) and (21), the weighted spectral vector $\mathbf{h}_n(\omega,l)$ is represented in terms of the spectral basis vector $\mathbf{u}_n(\omega)$ as

$$\mathbf{h}_n^T(\omega,l) = \mathbf{u}_n^T(\omega)[\mathbf{w}_n(l)\mathbf{w}_n^T(l)], \tag{23}$$

where the weight $[\mathbf{w}_n(l)\mathbf{w}_n^T(l)]$ is the outer-product of the vector $\mathbf{w}_n(l)$ and its transposition.

4. EXPERIMENTAL RESULTS

A room with size $4.45 \times 3.35 \times 2.5$ meters and an array of 2 omnidirectional microphones spaced of $0.2\ m$ are considered. The microphones are located in the middle of the room and are at the same height (i.e., $1.4\ m$) of three given sources. The distance from each source to the mid point between the two microphones is $1\ m$. The direction of arrivals of the sources in the observed mixtures are 35, 90, and 145 degrees. Synthetic room impulse responses (RIRs) are simulated through ISM

[19] with a sampling frequency of 16 kHz for three reverberation times: $T_{60} = 200, 350$, or 500 ms. Six Italian speakers (3 males and 3 females) are considered as audio sources. Each speaker uttered 20 sentences, of average length 8.75 s. The clean speech signals are divided into 3 speech signals of test data and 17 of training data to train $\mathbf{u}_n(\omega)$, $n = 1, 2, ..., N$. 5 male-female combinations of mixtures of $N = 3$ speech sources are generated, which corresponds to a total of 15 test mixtures for each T_{60}. The discrete time-frequency representation of the mixtures $\mathbf{x}(\omega, l)$ is obtained through STFT with a Hanning analysis window of length 128 ms (or 2048 samples), with a shift of 64 ms ($L = 137$). The window γ for the computation of the empirical covariance matrix of the source images in (12) is a Hanning window of size 3×3. Using the Kullback-Leibler (KL) divergence and applying the multiplicative update rule [13], the training power spectra were factorized with the number of spectral basis K equals 12.

The separation performance was evaluated via the signal-to-distortion ratio (SDR), source image-to-spatial distortion ratio (ISR), signal-to-interference ratio (SIR), and source-to-artifact ratio (SAR) criteria in decibels (dBs) [20], which account for overall distortion, target distortion, residual crosstalk, and musical noise, respectively. To initialize the source spatial images $\tilde{\mathbf{c}}_n(\omega, l)$, the time-frequency points of the observed mixture $\mathbf{x}(\omega, l)$ are assigned to clusters representing each source signal. The time difference of arrival (TDOA) of each source signal is estimated as in [21]. Given the estimated TDOAs of multiple source signals, the time-frequency points of the observed mixture are clustered into multiple clusters each corresponds to a source signal. The clustering is performed by minimizing the error between steering vectors of the estimated TDOAs and the phase differences of time-frequency points of $\mathbf{x}(\omega, l)$.

Table 1 shows comparison results of the separation performance. The proposed algorithm is denoted as weighted spectral bases (WSB). Source separation using NMF that was proposed in [16] is denoted as (NMF). In both algorithms, the separation system is fed by pre-trained source spectral basis matrices. The separation results of blind source separation using a full rank spatial covariance model proposed in [8] (ML), and the blind initialization algorithm using the estimated time difference of arrivals (TDOA) are reported in the table. The results of the ideal binary masking algorithm (BM Ideal) [4] and the ideal l_0-norm minimization algorithm (l_0-norm Ideal) [6] are also reported, in order to verify the upper bound limits of the separation performance.

The results show that the proposed algorithm outperforms the blind algorithms and the one fed by spectral basis matrices. From the reported results, it is obvious that having source-based prior information can improve much the separation performance. Furthermore, by capturing the changes in the amplitude values of the spatial covariance matrices from one frequency to another, the proposed algorithm performs better than the one in [16]. In comparison with the blind algorithms, in environments with low reverberation ($T_{60} = 200$ ms), we could gain about 3 dBs of SDR over the best performing one. The performance gain decreases as the mixing environments become more reverberant. In environments with high reverberation ($T_{60} = 500$ ms), we could gain around 2 dBs. On the other side, comparing to the one that is fed by spectral basis matrices (NMF), we could gain on the average around 1 dB in all the tested mixing environments.

Table 1. Comparison of the separation performance.

dB	BM	$l_0$norm	WSB	NMF	ML	TDOA
	Ideal	Ideal	Inf.	Inf.	Blind	Blind
SDR	10.53	10.12	7.91	6.96	4.62	4.90
ISR	19.44	17.56	13.95	12.32	9.06	11.12
SIR	20.45	15.95	14.46	12.28	7.25	10.48
SAR	11.12	14.20	9.59	10.22	8.30	7.27
Reverberation time = 200 ms						
SDR	10.02	7.80	5.70	4.65	3.56	3.01
ISR	18.70	13.63	11.32	9.38	7.30	8.50
SIR	19.73	13.06	11.44	8.21	5.38	6.68
SAR	10.61	10.03	8.53	9.90	7.93	6.64
Reverberation time = 350 ms						
SDR	9.57	6.30	4.47	3.73	2.48	2.30
ISR	18.08	11.57	10.14	8.38	5.90	7.51
SIR	19.11	11.07	9.56	6.36	3.36	4.87
SAR	10.15	8.59	8.01	9.69	7.41	6.24
Reverberation time = 500 ms						

5. CONCLUSION

This paper presents a method to estimate the model parameters of local Gaussian model based audio source separation. The model is parametrized by source variances and spatial covariance matrices. Spectral basis matrices trained on a set of training power spectra of source signals are assumed to be available. The matrices are obtained by factorizing the spectra applying nonnegative matrix factorization by minimizing the Kullback-Leibler (KL) divergence using multiplicative update rules. The source variances are estimated by applying singular value decomposition, and the spatial covariance matrices are estimated applying semi-supervised nonnegative matrix factorization. By building weighted basis matrices from vectors of the trained spectral matrices, the spatial covariance matrices are estimated using both the trained and weighted matrices. Comparing to blind source separation algorithms, using the proposed algorithm, we can gain between 3 and 2 dBs of SDR in environments with low and high reverberation, respectively. Furthermore, the proposed algorithm outperforms an algorithm informed by the trained basis matrices by around 1 dB in all the mixing environments.

6. REFERENCES

[1] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, New York, NY, USA:Wiley, 2003.

[2] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation:*, Springer, Berlin, 2007.

[3] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York, NY, USA:Wiley, 2001.

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

[5] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[6] Emmanuel Vincent, "Complex nonconvex l_p norm minimization for underdetermined source separation," in *Proceeding of ICA*, 2007, pp. 430–437.

[7] C. Fevotte and J. F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency gaussian models," in *Proc. WASPAA*, 2005, pp. 78–81.

[8] N. Duong, E. Vincent, and R. Gribonval, "Underdetermined 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, 2010.

[9] N Duong, E. Vincent, and R. Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP J. Adv. Sig. Proc.*, pp. 149–162, 2013.

[10] A. Ozerov and C. Fevotte, "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.

[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] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[13] A. Cichocki, R. Zdunek, A. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, 2009.

[14] M. Fakhry and F. Nesta, "Underdetermined source detection and separation using a normalized multichannel spatial dictionary," in *Proceedings of IWAENC 2012*, 2012.

[15] F. Nesta and M. Fakhry, "Unsupervised spatial dictionary learning for sparse underdetermined multichannel source separation," in *Proceedings of ICASSP 2013*, 2013, pp. 86–90.

[16] M. Fakhry, P. Svaizer, and M. Omologo, "Reverberant audio source separation using partially pre-trained nonnegative matrix factorization," in *Proceedings of IWAENC 2014*, 2014.

[17] M. Fakhry, P. Svaizer, and M. Omologo, "Audio source separation using a redundant library of source spectral bases for nonnegative tensor factorization," in *Proceedings of ICASSP 2015*, 2015.

[18] A. Dempster, N. Laird, and D Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[19] J. Allen and D. Berkeley, "Image method for efficiently simulating small-room acoustics," *Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.

[20] Emmanuel Vincent, Shoko Araki, Fabian J. Theis, Guido Nolte, Pau Bofill, Hiroshi Sawada, Alexey Ozerov, Vikrham Gowreesunker, Dominik Lutter, and Ngoc Q. K. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[21] M. Omologo and P. Svaizer, "Use of the cross-power-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 288–292, 1997.