HAL
open science

Spatial covariance models for under-determined reverberant audio source separation

Ngoc Q.K. Duong, Emmanuel Vincent, Rémi Gribonval

To cite this version:

Ngoc Q.K. Duong, Emmanuel Vincent, Rémi Gribonval. Spatial covariance models for under-determined reverberant audio source separation. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09., Oct 2009, New Paltz, United States. pp. 129 - 132, 10.1109/ASPAA.2009.5346503 . hal-00481529

HAL Id: hal-00481529

https://hal.science/hal-00481529v1

Submitted on 6 May 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SPATIAL COVARIANCE MODELS FOR UNDER-DETERMINED REVERBERANT AUDIO SOURCE SEPARATION

Ngoc Q.K. Duong, Emmanuel Vincent and Rémi Gribonval

METISS project team, IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France
{qduong, emmanuel.vincent, remi.gribonval}@irisa.fr

ABSTRACT

The separation of under-determined convolutive audio mixtures is generally addressed in the time-frequency domain where the sources exhibit little overlap. Most previous approaches rely on the approximation of the mixing process by complex-valued multiplication in each frequency bin. This is equivalent to assuming that the spatial covariance matrix of each source, that is the covariance of its contribution to all mixture channels, has rank 1. In this paper, we propose to represent each source via a full-rank spatial covariance matrix instead, which better approximates reverberation. We also investigate a possible parameterization of this matrix stemming from the theory of statistical room acoustics. We illustrate the potential of the proposed approach over a stereo reverberant speech mixture.

***Index Terms*—** Audio source separation, under-determined mixtures, reverberation, spatial covariance models

1. INTRODUCTION

Most audio signals are mixtures of several sound sources such as speech, music, and background noise. The observed multichannel signal $\mathbf{x}(t)$ can be expressed as

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{s}_j^{\mathrm{img}}(t) \tag{1}$$

where $\mathbf{s}_j^{\mathrm{img}}(t)$ is the spatial image of source j. When the mixture results from the recording of J static point sources via I static microphones, this quantity can be modeled via the convolutive mixing process

$$\mathbf{s}_j^{\mathrm{img}}(t) = \sum_{\tau} \mathbf{h}_j(\tau) s_j(t-\tau) \tag{2}$$

where $s_j(t)$ is the j-th source signal and $\mathbf{h}_j(\tau)$ is the vector of mixing filter coefficients modeling the acoustic path from source j to all microphones. Source separation consists of recovering either the source signals or their spatial images given the mixture signal. This problem is particularly difficult in the under-determined case, *i.e.* when the number of sources J is larger than the number of mixture channels I.

Most existing approaches transform the signals into the time-frequency domain via the short-time Fourier transform (STFT) and approximate the convolutive mixing process by a complex-valued mixing matrix in each frequency bin. Source separation can then be achieved by estimating the mixing matrices in all frequency bins and deriving the source STFT coefficients under some sparse prior distribution. Popular algorithms include binary masking [1] or ℓ_1-norm minimization [2].

In [3], a different framework was proposed whereby the vector $\mathbf{s}_j^{\mathrm{img}}(n,f)$ of STFT coefficients of each spatial source image at time frame n and frequency bin f is modeled as a zero-mean Gaussian variable with covariance matrix

$$\mathbf{R}_{\mathbf{s}_j^{\mathrm{img}}}(n,f) = v_j(n,f)\,\mathbf{R}_j(f) \tag{3}$$

where $v_j(n,f)$ is a scalar time-varying variance and $\mathbf{R}_j(f)$ a time-invariant matrix encoding the spatial properties of the source. Assuming that the sources are uncorrelated, the vector $\mathbf{x}(n,f)$ of STFT coefficients of the mixture signal is also zero-mean Gaussian with covariance matrix

$$\mathbf{R}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(f). \tag{4}$$

The parameters $v_j(n,f)$ and $\mathbf{R}_j(f)$ of the model are estimated in the maximum likelihood (ML) sense. The spatial images of all sources are then obtained in the minimum mean square error (MMSE) sense by Wiener filtering

$$\widehat{\mathbf{s}}_j^{\mathrm{img}}(n,f) = v_j(n,f)\mathbf{R}_j(f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f). \tag{5}$$

This framework was applied to the separation of instantaneous audio mixtures in [4] and shown to provide better separation performance than ℓ_p-norm minimization. Recently, the covariance model (3) was also applied to the separation of convolutive audio mixtures via multichannel nonnegative matrix factorization (NMF) in [5]. However, the mixing process was approximated by complex-valued multiplication in each frequency bin, resulting in a rank-1 approximation of the spatial covariance matrix of each source.

In the following, we investigate the modeling of each source by a full-rank spatial covariance matrix instead. This generalization was shown to improve flexibility of the model for astronomical data in [6]. We argue that it provides a better approximation of reverberation within audio recordings and study a possible full-rank parameterization stemming from the theory of statistical room acoustics [7]. We demonstrate the potential of the proposed approach by considering the separation of a reverberant speech mixture in a semi-blind context, where the source spatial covariance matrices $\mathbf{R}_j(f)$ are known. We also compute a performance upper-bound for each model.

The structure of the rest of the paper is as follows. We present rank-1 and full-rank spatial covariance models in Section 2 and explain how to estimate the source variances $v_j(n,f)$ in Section 3. We evaluate the separation performance achieved by all models on speech data in Section 4 and conclude in Section 5.

2. SPATIAL COVARIANCE MODELS

We investigate four spatial source models with different degrees of flexibility. For simplicity, we focus our presentation on stereo ($I = 2$) signals, although the models can also be defined for $I > 2$ channels.

2.1. Rank-1 convolutive model

Most existing approaches to audio source separation approximate the convolutive mixing process (2) by the complex-valued multiplication $\mathbf{s}_j^{\text{img}}(n, f) = \mathbf{h}_j(f)s_j(n, f)$ where $\mathbf{h}_j(f)$ is the Fourier transform of the mixing filters $\mathbf{h}_j(\tau)$ and $s_j(n, f)$ is the STFT of $s_j(t)$. The covariance matrix of $\mathbf{s}_j^{\text{img}}(n, f)$ is then given by (3) where $v_j(n, f)$ is the variance of $s_j(t)$ and $\mathbf{R}_j(f)$ is equal to the rank-1 matrix

$$\mathbf{R}_j(f) = \mathbf{h}_j(f)\mathbf{h}_j^H(f) \tag{6}$$

with ^{H} denoting matrix conjugate transposition. In the following, we assume that the vectors $\mathbf{h}_j(f)$ associated with different sources j are not collinear.

2.2. Rank-1 anechoic model

In the particular case where the recording environment is anechoic, each mixing filter $h_{ij}(\tau)$ is the combination of a delay and a gain. The FFT of the mixing filters $\mathbf{h}_j(f)$ is then denoted as $\mathbf{a}_j(f)$ and specified by the distance r_{ij} between each source j and each microphone i. More precisely,

$$\mathbf{a}_j(f) = \begin{pmatrix} \kappa(r_{1j})e^{-2i\pi f\tau_{1j}} \\ \kappa(r_{2j})e^{-2i\pi f\tau_{2j}} \end{pmatrix} \tag{7}$$

where

$$\kappa(r_{ij}) = \frac{1}{\sqrt{4\pi}r_{ij}} \quad \text{and} \quad \tau_{ij} = \frac{r_{ij}}{c} \tag{8}$$

are respectively the mixing gains and delays for source j with sound velocity c [7].

2.3. Full-rank direct+diffuse model

The above rank-1 models rely on an approximation of the actual mixing process, whereby the sound of source j as recorded on the microphones comes from a single spatial position at each frequency f, as specified by $\mathbf{h}_j(f)$. In practice, reverberation increases the spatial spread of each source, due to echoes at many different positions on the walls of the recording room.

The theory of statistical room acoustics assumes that the spatial image of each source is composed of two uncorrelated parts: a direct part modeled by $\mathbf{a}_j(f)$ and a reverberant part. The spatial covariance $\mathbf{R}_j(f)$ of source j is then a full-rank matrix defined as the sum of the covariance of the direct part and the covariance of the reverberant part [7]

$$\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^H(f) + \sigma_{\text{rev}}^2 \begin{pmatrix} 1 & \Psi(d, f) \\ \Psi(d, f) & 1 \end{pmatrix} \tag{9}$$

where σ_{rev}^2 is the power of the reverberant part and $\Psi(d, f)$ is a function of microphone distance d and frequency f. This model assumes that the reverberation recorded at both microphones has the same power but is correlated as characterized by $\Psi(d, f)$. This model was employed for single source localization in [7] but its use for the separation of multiple sources has not yet been investigated.

Assuming that the reverberant part is diffuse, *i.e.* its intensity is uniformly distributed over all possible directions, its normalized cross-correlation is real-valued and shown in [8] to be equal to

$$\Psi(d, f) = \frac{\sin(2\pi fd/c)}{2\pi fd/c}. \tag{10}$$

Furthermore, the power of the reverberant part within a rectangular room is given by $\sigma_{rev}^2 = 4\beta^2/(\mathcal{A}\left(1 - \beta^2\right))$ where $\mathcal{A}$ is the total wall area and β the wall reflection coefficient computed from the room reverberation time via Eyring's formula [7].

2.4. Full-rank unconstrained model

In practice, the assumption that the reverberant part is diffuse is rarely satisfied. Indeed, early echoes containing more energy are not uniformly distributed on the walls of the recording room, but at certain positions depending on the position of the source and the microphones. When performing some simulations in a rectangular room, we observed that (10) is valid on average when considering a large number of sources at different positions, but generally not valid for each source considered independently.

Therefore, we also investigate the modeling of each source via an unconstrained spatial covariance matrix $\mathbf{R}_j(f)$ whose coefficients are not related a priori. Since this model is more general than (6) and (9), it allows more flexible modeling of the mixing process and is expected to improve separation performance of real-world convolutive mixtures. However the estimation of its parameters may be more difficult in a blind context.

3. ESTIMATION OF THE MODEL PARAMETERS

In order to use the models for *blind* source separation, we would need to estimate the source variances $v_j(n, f)$ and their spatial parameters $\mathbf{h}_j(f)$, r_{ij}, σ_{rev}^2, d or $\mathbf{R}_j(f)$ from the mixture signal only. Recent evaluations of state-of-the-art algorithms [9] have shown that the estimation of spatial parameters remains difficult for real-world reverberant mixtures, due in particular to the existence of multiple local maxima in the ML criterion and to the source permutation problem arising when the model parameters at different frequencies are assumed to be independent. In the following, we investigate the potential separation performance achievable via each model in a *semi-blind* context, where the spatial covariance matrices $\mathbf{R}_j(f)$ are known but the source variances $v_j(n, f)$ are blindly estimated from the observed mixture. We also compare the models in an *oracle* context, where both the spatial covariance matrices $\mathbf{R}_j(f)$ and the source variances $v_j(n, f)$ are known. The resulting performance figures provide upper bounds of the separation performance achievable in a blind context.

3.1. Blind estimation of the source variances

From now on, we assume that the spatial covariance matrices $\mathbf{R}_j(f)$ are known. Let us denote by $\widehat{\mathbf{R}}_{\mathbf{x}}(n, f)$ the empirical covariance matrix of the mixture signal $\mathbf{x}(n, f)$ in the time-frequency point (n, f). This quantity can be computed by averaging over some time-frequency neighborhood of that point as [4]

$$\widehat{\mathbf{R}}_{\mathbf{x}}(n, f) = \frac{\sum_{n',f'} w(n' - n, f' - f)\mathbf{x}(n', f')\mathbf{x}(n', f')^H}{\sum_{n',f'} w(n' - n, f' - f)} \tag{11}$$

where w is a bi-dimensional window specifying the shape of the neighborhood. The estimation of the source variances $v_j(n,f)$ in the ML sense is equivalent to minimizing the sum over all time-frequency points (n,f) of the Kullback-Leibler (KL) divergence $D_{KL}(\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)|\mathbf{R}_{\mathbf{x}}(n,f))$ between two zero-mean Gaussian distributions with covariance matrices $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)$ and $\mathbf{R}_{\mathbf{x}}(n,f)$ defined in (11) and (4) respectively with

$$D_{KL}(\widehat{\mathbf{R}}|\mathbf{R}) = \frac{1}{2}[\mathrm{tr}(\mathbf{R}^{-1}\widehat{\mathbf{R}}) - \log\det(\mathbf{R}^{-1}\widehat{\mathbf{R}})] - 1. \quad (12)$$

This criterion is defined only when $\mathbf{R}$ and $\widehat{\mathbf{R}}$ are both full-rank matrices. It is always nonnegative and equal to zero if and only if $\mathbf{R} = \widehat{\mathbf{R}}$.

This minimization with respect to the source variances was performed iteratively using the expectation-mazimization (EM) algorithm in [3] and a faster conjugate gradient algorithm in [6]. We observed that the resulting source separation performance is quite sensitive to the initial parameter values, since both algorithms may converge to a local minimum of the criterion. The issue of finding an appropriate initialization was addressed in [4] for rank-1 instantaneous models. We now extend this approach to rank-1 and full-rank convolutive models. To simplify the notation, we omit time and frequency indexes hereafter, since the estimation of the source variances is achieved separately in each time-frequency point.

We choose the initial source variances v_j as the global minimum of the KL divergence under the constraint that at most two sources are active in the considered time-frequency point, *i.e.*

$$\exists j_1, j_2 \text{ such that } v_j = 0 \; \forall j \notin \{j_1, j_2\}. \quad (13)$$

We compute this global minimum via a non-iterative approach. We distinguish two cases: either this global minimum involves a single active source or it involves two active sources. Let us consider first the case when the global minimum of the KL divergence under constraint (13) involves a single active source indexed by j. By computing the derivative of $D_{KL}(\widehat{\mathbf{R}}_{\mathbf{x}}|\mathbf{R}_{\mathbf{x}})$ with respect to v_j and equating it to zero, we get

$$v_j = \frac{1}{2}\,\mathrm{tr}(\mathbf{R}_j^{-1}\widehat{\mathbf{R}}_{\mathbf{x}}). \quad (14)$$

Let us now assume that the global minimum of the KL divergence under constraint (13) involves two active sources indexed by j_1 and j_2. We use the fact that there exists an invertible matrix $\mathbf{A}$ and two diagonal matrices $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ such that $\mathbf{R}_{j_1} = \mathbf{A}\mathbf{\Lambda}_1\mathbf{A}^H$ and $\mathbf{R}_{j_2} = \mathbf{A}\mathbf{\Lambda}_2\mathbf{A}^H$. When $\mathbf{R}_{j_1}$ or $\mathbf{R}_{j_2}$ have full rank, the columns of $\mathbf{A}$ can be computed as the eigenvectors of $\mathbf{R}_{j_2}\mathbf{R}_{j_1}^{-1}$ or $\mathbf{R}_{j_1}\mathbf{R}_{j_2}^{-1}$ as shown in [10]. When both $\mathbf{R}_{j_1}$ and $\mathbf{R}_{j_2}$ have rank 1, the columns of $\mathbf{A}$ are the vectors $\mathbf{h}_{j_1}$ and $\mathbf{h}_{j_2}$ such that $\mathbf{R}_{j_1} = \mathbf{h}_{j_1}\mathbf{h}_{j_1}^H$ and $\mathbf{R}_{j_2} = \mathbf{h}_{j_2}\mathbf{h}_{j_2}^H$. Since the KL divergence is invariant under invertible linear transforms, it can be rewritten as

$$D_{KL}(\widehat{\mathbf{R}}_{\mathbf{x}}|\mathbf{R}_{\mathbf{x}}) = D_{KL}(\mathbf{A}^{-1}\widehat{\mathbf{R}}_{\mathbf{x}}(\mathbf{A}^H)^{-1}|v_{j_1}\mathbf{\Lambda}_1 + v_{j_2}\mathbf{\Lambda}_2). \quad (15)$$

By computing the gradient of this expression with respect to v_{j_1} and v_{j_2} and equating it to zero, we obtain

$$\begin{pmatrix} v_{j_1} \\ v_{j_2} \end{pmatrix} = \begin{pmatrix}\mathrm{diag}(\mathbf{\Lambda}_1) & \mathrm{diag}(\mathbf{\Lambda}_2)\end{pmatrix}^{-1} \mathrm{diag}(\mathbf{A}^{-1}\widehat{\mathbf{R}}_{\mathbf{x}}(\mathbf{A}^H)^{-1}) \quad (16)$$

where $\mathrm{diag}(.)$ denotes the column vector of diagonal entries of a matrix. Note that this equation may result in negative variances v_{j_1} or v_{j_2}, which implies that the KL divergence is not minimized when these two variances only are nonzero. To sum up, the global minimum of the KL divergence under the constraint that at most two sources are active is obtained by considering all possible indexes j and pairs of indexes (j_1, j_2) of active sources, deriving the optimal source variances via (14) and (16) and selecting the index or the pair of indexes resulting in the smallest KL divergence.

Once a suitable initial value of the source variances has been found, the KL minimization problem can be solved via any Newton-based optimizer given the gradient and Hessian of the criterion. In the following, we use Matlab's fmincon optimizer which is based on a subspace trust region. Preliminary experiments showed that convergence is achieved in less than five iterations in most time-frequency points.

3.2. Oracle estimation of the source variances

In addition to semi-blind separation, we evaluate the separation performance achieved by each model when the "true" source variances $v_j(n,f)$ are known. These variances can be derived from the true source spatial images $\mathbf{s}_j^{\mathrm{img}}(n,f)$ when available. Let us denote by $\widehat{\mathbf{R}}_{\mathbf{s}_j^{\mathrm{img}}}(n,f)$ the empirical covariance matrix of the spatial image of source j, which can be computed as in (11) by replacing $\mathbf{x}$ with $\mathbf{s}_j^{\mathrm{img}}$. For full-rank models, the "true" source variances can be computed in the ML sense by minimizing the KL divergence $D_{KL}(\widehat{\mathbf{R}}_{\mathbf{s}_j^{\mathrm{img}}}(n,f)|\mathbf{R}_{\mathbf{s}_j^{\mathrm{img}}}(n,f))$, which gives

$$v_j(n,f) = \frac{1}{2}\,\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{R}}_{\mathbf{s}_j^{\mathrm{img}}}(n,f)). \quad (17)$$

For rank-1 models, this divergence is undefined since $\mathbf{R}_{\mathbf{s}_j^{\mathrm{img}}}(n,f)$ is not invertible. We compute the true source variances instead as

$$v_j(n,f) = \frac{|\mathbf{h}_j^H(f)\mathbf{s}_j^{img}(n,f)|^2}{\|\mathbf{h}_j(f)\|_2^2}. \quad (18)$$

3.3. Computation of the source spatial covariance matrices

The estimation of the source variances in Sections 3.1 and 3.2 relied on the knowledge of $\mathbf{R}_j(f)$. For rank-1 models and for the full-rank direct+diffuse model, $\mathbf{R}_j(f)$ was computed from the geometry setting or from the mixing filters as explained in Sections 2.1, 2.2 and 2.3. For the full-rank unconstrained model in Section 2.4, $\mathbf{R}_j(f)$ was computed by iterative minimization of $D_{KL}(\widehat{\mathbf{R}}_{\mathbf{s}_j^{\mathrm{img}}}(n,f)|\mathbf{R}_{\mathbf{s}_j^{\mathrm{img}}}(n,f))$ by alternate application of (17) and

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\widehat{\mathbf{R}}_{\mathbf{s}_j^{\mathrm{img}}}(n,f) \quad (19)$$

where N is the total number of time frames. The minimization was initialized by $\mathbf{R}_j(f) = \frac{1}{N}\sum_{t=1}^{N}\widehat{\mathbf{R}}_{\mathbf{s}_j^{\mathrm{img}}}(n,f)$ and convergence was typically achieved in two or three iterations.

4. EXPERIMENTAL EVALUATION

We evaluated the source separation performance achieved by each of the models in Section 2 over a three-source stereo reverberant speech mixture using the semi-blind and oracle parameter estimation algorithms described in Section 3. The mixture was generated by convolving 5 s speech signals sampled at 16 kHz with room

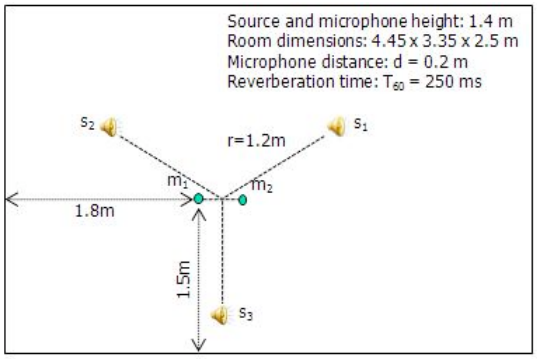

Figure 1: Geometry setting of the test mixture.

Source variance estimation	Covariance model	Rank	SDR	SIR	ISR
Blind	anechoic	1	0.9	1.7	4.8
	convolutive	1	4.0	6.4	8.5
	direct+diffuse	2	3.1	6.1	7.3
	unconstrained	2	5.8	10.3	10.5
Oracle	anechoic	1	0.4	4.4	7.5
	convolutive	1	4.2	10.2	6.2
	direct+diffuse	2	10.2	17.3	11.7
	unconstrained	2	10.9	17.9	12.5

Table 1: Average source separation performance.

impulse responses simulated via the source image method so that the geometry setting, *i.e.* showned in Fig 1, is known exactly. The STFT was computed with a sine window of length 1024 (64 ms). The bi-dimensional window w defining time-frequency neighborhoods in (11) was chosen as the outer product of two Hanning windows with length of 3 [4]. Computation time was on the order of 5 min per model in the semi-blind case using Matlab on a 2.4 GHz computer. Separation performance was evaluated using the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and source image-to-spatial distortion ratio (ISR) criteria in [9], averaged over all sources.

The results are shown in Table 1. The rank-1 anechoic model has lowest performance because it only accounts for the direct path. In a semi-blind context, the full-rank direct+diffuse model results in a SDR decrease of 1 dB compared to the rank-1 convolutive model. This decrease appears surprisingly small when considering the fact that the former involves 8 spatial parameters (6 distances r_{ij}, plus σ^2_{rev} and d) instead of 3078 parameters (6 mixing coefficients per frequency bin) for the latter. The full-rank unconstrained model improves the SDR by 2 dB and 2.5 dB when compared to the rank-1 convolutive model and binary masking method respectively. In an oracle context, full-rank models clearly outperform rank-1 models by 6 dB or more regarding all criteria. Also, the performance of the full-rank direct+diffuse model is very close to that of the unconstrained model.

5. CONCLUSION

In this paper, we proposed to model the spatial properties of sound sources by full-rank spatial covariance matrices and studied a possible parameterization of these matrices stemming from the theory of statistical room acoustics. We derived algorithms to estimate the source variances and perform source separation either in a semi-blind or in an oracle setting. Experimental results over speech data confirm that full-rank spatial covariance matrices better account for reverberation and potentially improve separation performance compared to rank-1 matrices. Future work will validate the performance of the proposed algorithms over real-world recordings. Moreover, we will investigate blind learning of full-rank spatial covariance matrices from the mixture signal. In order to address the permutation problem, we will take into account dependencies between the model parameters in different frequency bins by investigating both advanced models of the source variances in the spirit of [5] and alternative parameterizations of the spatial covariance matrices providing more flexibility than the current direct+diffuse parameterization, *e.g.* by learning the value of $\Psi(d, f)$ from the data instead of defining it as in (10).

6. REFERENCES

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

[2] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007, article ID 24717.

[3] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models," in *Proc. 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 78–81.

[4] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 775–782.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation." in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2009.

[6] J. F. Cardoso and M. Martin, "A flexible component model for precision ICA," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 1–8.

[7] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 791–803, Nov 2003.

[8] H. Kuttruff, *Room Acoustics*, 4th ed., New York, 2000.

[9] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 552–559.

[10] A. Yeredor, "On using exact joint diagonalization for noniterative approximate joint diagonalization," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 645–648, Sep 2005.