Under-determined convolutive blind source separation using spatial covariance models

Ngoc Duong, Emmanuel Vincent, Rémi Gribonval

UNDER-DETERMINED CONVOLUTIVE BLIND SOURCE SEPARATION USING SPATIAL COVARIANCE MODELS

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

This paper deals with the problem of under-determined convolutive blind source separation. We model the contribution of each source to all mixture channels in the time-frequency domain as a zero-mean Gaussian random variable whose covariance encodes the spatial properties of the source. We consider two covariance models and address the estimation of their parameters from the recorded mixture by a suitable initialization scheme followed by an iterative expectation-maximization (EM) procedure in each frequency bin. We then align the order of the estimated sources across all frequency bins based on their estimated directions of arrival (DOA). Experimental results over a stereo reverberant speech mixture show the effectiveness of the proposed approach.

Index Terms— Convolutive blind source separation, under-determined mixtures, spatial covariance models, EM algorithm, permutation problem.

1. INTRODUCTION

In blind source separation (BSS), the recorded multichannel signal $\mathbf{x}(t)$ is a mixture of several sound sources

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

where $\mathbf{c}_j(t)$ is the spatial image of source j, that is its contribution to all mixture channels. Each $\mathbf{c}_j(t)$ can be modeled via the convolutive mixing process

$$\mathbf{c}_j(t) = \sum_{\tau} \mathbf{h}_j(\tau) s_j(t-\tau) \tag{2}$$

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. Under-determined BSS consists in recovering either the J original source signals or their spatial images given the I mixture channels where $I < J$.

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 is then achieved by estimating the mixing matrices in all frequency bins and deriving the source STFT coefficients under a sparse prior distribution. Popular algorithms include binary masking [1] or ℓ_1-norm minimization [2].

A different framework [3, 4] assumes that the sources are uncorrelated and the vector $\mathbf{c}_j(n,f)$ of STFT coefficients of each spatial source image in time frame n and frequency bin f is modeled as a zero-mean Gaussian random variable with covariance matrix

$$\mathbf{R}_{\mathbf{c}_j}(n,f) = v_j(n,f)\,\mathbf{R}_j(f) \tag{3}$$

where $v_j(n,f)$ is a scalar time-varying variance and $\mathbf{R}_j(f)$ a time-invariant *spatial covariance matrix* encoding the spatial properties of the source. The parameters $v_j(n,f)$ and $\mathbf{R}_j(f)$ can be 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.

This framework was first applied to the separation of instantaneous audio mixtures in [5] using a rank-1 spatial covariance matrix $\mathbf{R}_j(f)$. In [3], we extended this approach to convolutive mixtures and proposed to consider $\mathbf{R}_j(f)$ as a full-rank matrix. This model was shown to improve separation performance of reverberant mixtures in an oracle context, where the true values of $\mathbf{R}_j(f)$ and $v_j(n,f)$ are known, and in a semi-blind context, where $\mathbf{R}_j(f)$ was estimated from single-source training data but $v_j(n,f)$ was blindly estimated from the mixture.

In this paper, we investigate the estimation of both $\mathbf{R}_j(f)$ and $v_j(n,f)$ in a blind context where only the mixture signal is available. For that purpose, we use the expectation-maximization (EM) algorithm and propose an effective parameter initialization scheme. We also solve the source permutation problem arising when model parameters at different frequencies are independently estimated. We argue for the better source separation performance of the full-rank model over the rank-1 model and state-of-the-art algorithms on mixtures with realistic reverberation time.

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 address the blind estimation of the model parameters in Section 3. We compare the source separation performance achieved by rank-1 and full-rank models and by state-of-the-art algorithms over speech data in Section 4. Finally we conclude in Section 5.

2. SPATIAL COVARIANCE MODELS

We investigate two general spatial source models with different degrees of flexibility resulting in a rank-1 or a full-rank spatial covariance matrix.

2.1. Rank-1 model

Most under-determined BSS approaches model the convolutive mixing process (2) in the frequency domain as $\mathbf{c}_j(n,f) = \mathbf{h}_j(f)s_j(n,f)$, 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 $s_j(t)$ [2]. The spatial covariance matrix of source j is then equal to the rank-1 matrix

$$\mathbf{R}_j(f) = \mathbf{h}_j(f)\mathbf{h}_j^H(f) \tag{4}$$

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

2.2. Full-rank model

The above rank-1 model assumes that the sound of source j as recorded on the microphones comes from a single spatial position at each frequency f, as specified by $\mathbf{h}_j(f)$. But in practice, reverberation increases the spatial spread of each source due to echoes at many different positions on the walls. Therefore, we also investigate the modeling of each source via a full-rank spatial covariance matrix $\mathbf{R}_j(f)$ without any constraint on its entries [3]. Since this model is more general than the rank-1 model in (4), it allows more flexible modeling of the mixing process.

3. ESTIMATION OF MODEL PARAMETERS

We estimate the model parameters $\mathbf{h}_j(f)$, $\mathbf{R}_j(f)$ and $v_j(n,f)$ from the recorded mixture using a three-step procedure: initialization by hierarchical clustering, iterative estimation via the EM algorithm, and permutation alignment. The overall process is shown in Fig. 1.

3.1. Initialization by hierarchical clustering

Preliminary experiments showed that the initialization of the model parameters greatly affects the separation performance resulting from the EM algorithm. In the following, we propose a hierarchical clustering-based initialization scheme inspired from the algorithm in [2].

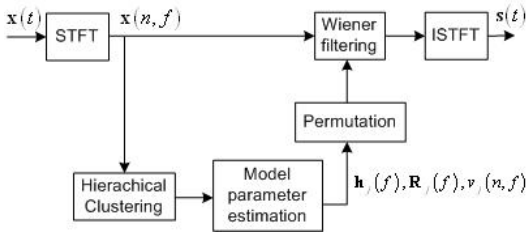

Fig. 1. Flow of the proposed BSS approach.

This scheme relies on the assumption that the sound from each source comes from a certain region of space at each frequency f, which is different for all sources. The vectors $\mathbf{x}(n,f)$ of mixture STFT coefficients are then likely to cluster around the direction of the associated mixing vector $\mathbf{h}_j(f)$ in the time frames n where source j is predominant.

In order to estimate these clusters, we first normalize the vectors of mixture STFT coefficients as

$$\bar{\mathbf{x}}(n,f) \leftarrow \frac{\mathbf{x}(n,f)}{\|\mathbf{x}(n,f)\|_2} e^{-i\arg(x_1(n,f))} \tag{5}$$

where $\arg(.)$ denotes the phase of a complex number and $\|.\|_2$ the Euclidean norm. We then define the distance between two clusters C_1 and C_2 by the average distance between the associated normalized mixture STFT coefficients

$$d(C_1,C_2) = \frac{1}{|C_1||C_2|} \sum_{\bar{\mathbf{x}}_1 \in C_1} \sum_{\bar{\mathbf{x}}_2 \in C_2} \|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\|_2 \tag{6}$$

At a given frequency, the vectors of mixture STFT coefficients on all time frames are first considered as clusters containing a single item. The distance between each pair of clusters is computed and the two clusters with the smallest distance are merged. This "bottom up" process called linking is repeated until the number of clusters is less than a predetermined threshold K. This threshold is usually much larger than the number of sources J [2], so as to eliminate outliers. We finally choose the J clusters with the largest number of samples and calculate the initial values as

$$\mathbf{h}_j^{\text{init}}(f) = \frac{1}{|C_j|} \sum_{\mathbf{x}(n,f) \in C_j} \mathbf{x}(n,f) e^{-i\arg(x_1(n,f))} \tag{7}$$

$$\mathbf{R}_j^{\text{init}}(f) = \frac{1}{|C_j|} \sum_{\mathbf{x}(n,f) \in C_j} \mathbf{x}(n,f)\mathbf{x}(n,f)^H \tag{8}$$

Note that, contrary to the algorithm in [2], we define the distance between clusters as the average distance between the normalized mixture STFT coefficients instead of the minimum distance between them. Besides, the mixing vector $\mathbf{h}_j^{\text{init}}(f)$ is computed from the phase-normalized mixture STFT coefficients as (7) instead of the normalized coefficients as (5). These modifications were found to provide better initial approximation of the mixing parameters in

our experiments. We also tested random initialization and direction-of-arrival (DOA) based initialization, *i.e.* where the mixing vectors $\mathbf{h}_j^{\text{init}}(f)$ are derived from known source and microphone positions assuming no reverberation. Both schemes were found to result in slower convergence and poorer separation performance for both the rank-1 and full-rank model than the proposed scheme.

3.2. Maximum likelihood estimation with EM

Under model (3), the STFT coefficients of the mixture signal are zero-mean Gaussian with covariance matrix

$$\mathbf{R}_{\mathbf{x}}(n,f) = \sum_j \mathbf{R}_{\mathbf{c}_j}(n,f). \tag{9}$$

In each frequency bin f, we now wish to estimate both the mixing parameters $\mathbf{h}_j(f)$ or $\mathbf{R}_j(f)$ and the source variances $v_j(n,f)$ in the ML sense. The EM algorithm is well-known as an appropriate choice in this case [6].

For the rank-1 model, the EM algorithm is derived based on the *complete data* $\{\mathbf{x}(n,f), s_j(n,f)\ \forall j,n\}$ that is the set of STFT coefficients of all mixture channels and all sources on all time frames. Due to lack of space, we do not provide the EM parameter updates for this model here. Similar updates can be found in [5, 7].

For the full-rank model, the *complete data* becomes $\{\mathbf{c}_j(n,f)\ \forall j,n\}$ that is the set of STFT coefficients of the mixture and all source images on all channels and all time frames. The details of one EM iteration for each source j are as follows.

In the E-step, the Wiener filter $\mathbf{W}_j(n,f)$ and the mean $\widehat{\mathbf{c}}_j(n,f)$ and the covariance matrix $\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)$ of the estimated source image are computed as

$$\mathbf{W}_j(n,f) = \mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f) \tag{10}$$

$$\widehat{\mathbf{c}}_j(n,f) = \mathbf{W}_j(n,f)\mathbf{x}(n,f) \tag{11}$$

$$\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f) = \widehat{\mathbf{c}}_j(n,f)\widehat{\mathbf{c}}_j^H(n,f) + (\mathbf{I} - \mathbf{W}_j(n,f))\mathbf{R}_{\mathbf{c}_j}(n,f) \tag{12}$$

where $\mathbf{I}$ is the $I \times I$ identity matrix, $\mathbf{R}_{\mathbf{c}_j}(n,f)$ is defined in (3) and $\mathbf{R}_{\mathbf{x}}(n,f)$ in (9).

In the M-step, $\mathbf{R}_j(f)$ and $v_j(n,f)$ are updated as [3]

$$v_j(n,f) = \frac{1}{I}\,\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)) \tag{13}$$

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f) \tag{14}$$

where $\mathrm{tr}(.)$ denotes the trace of a square matrix.

3.3. Permutation alignment

Since the model parameters are estimated independently in each frequency bin f, they should be ordered so as to correspond to the same source across all frequency bins. In order to solve this so-called permutation problem, we apply the DOA-based algorithm described in [8] for the rank-1 model. Given the geometry of the microphone array, this algorithm computes the DOAs of all sources and permutes the model parameters by clustering the estimated mixing vectors $\mathbf{h}_j(f)$ normalized as in (5).

Regarding the full-rank model, we first apply principal component analysis (PCA) to calculate the first principal component $\mathbf{w}_j(f)$ of the spatial covariance matrix $\mathbf{R}_j(f)$ of each source j in each frequency bin f. This vector is conceptually equivalent to the mixing vector $\mathbf{h}_j(f)$ of the rank-1 model. Thus, we can apply the same procedure to solve the permutation problem. Fig. 2 depicts the phase of the second entry $w_{2j}(f)$ of $\mathbf{w}_j(f)$ before and after solving the permutation for a stereo mixture of three sources with room reverberation time $T_{60} = 130$ ms, where $\mathbf{w}_j(f)$ has been normalized as in (5). This phase is unambiguously related to the source DOAs below 5 kHz. Above that frequency, spatial aliasing occurs. We can see that the source order is globally aligned for most frequency bins after solving the permutation.

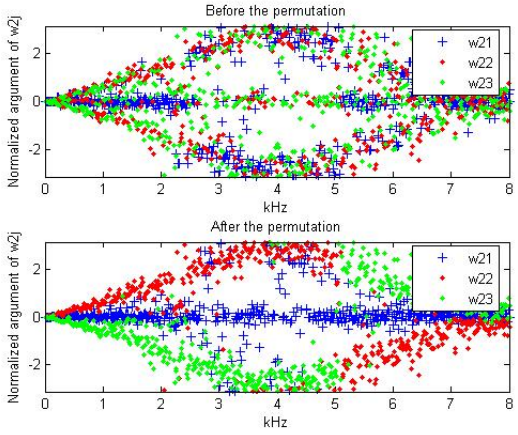

Fig. 2. Normalized argument of $w_{2j}(f)$ before and after permutation alignment for stereo mixture of three sources.

4. EXPERIMENTAL RESULTS

We evaluated the blind source separation performance of the proposed algorithm for the rank-1 and the full-rank model over stereo mixtures of three speech sources with different reverberation times. The mixtures were generated by convolving 8 s speech signals sampled at 16 kHz with room impulse responses simulated via the source image method. The STFT was computed with a sine window of length 1024. The distance between two microphones was 5 cm and the distance from sources to microphones 50 cm. The number of clusters was set to $K = 30$ and the number of EM iterations to 20. We also computed the performance of binary masking and ℓ_1-norm minimization using the same mixing parameters as those estimated from the hierarchical clustering step. Separa-

tion performance was evaluated using the signal-to-distortion ratio (SDR) criterion measuring overall distortion, as well as the signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR) and source image-to-spatial distortion ratio (ISR) criteria in [9], averaged over all sources, and shown in Table 1. The separation results for ℓ_1-norm minimization were not included in the table since they are about 0.3 dB below that of the rank-1 model in terms of SDR regardless of reverberation time.

$T_{60}(\mathrm{ms})$	Approach	SDR	SIR	SAR	ISR
50	Binary masking	8.9	18.2	9.4	18.3
	Rank-1 model	11.4	17.2	12.8	20.8
	Full-rank model	10.6	16.8	11.9	17.7
130	Binary masking	7.2	14.3	7.8	14.7
	Rank-1 model	7.4	11.4	10.0	14.2
	Full-rank model	8.8	13.8	11.2	15.2
250	Binary masking	5.2	10.9	6.0	11.0
	Rank-1 model	4.0	7.9	7.5	9.2
	Full-rank model	6.7	10.4	10.0	10.9
500	Binary masking	2.3	6.1	4.2	7.4
	Rank-1 model	0.9	3.6	6.4	5.7
	Full-rank model	3.8	5.8	7.5	7.2

Table 1. Average source separation performance

In a low reverberant environment, *i.e.* $T_{60} = 50$ ms, the rank-1 model provides the best SDR and SAR among the three remaining approaches. This is consistent with the fact that the direct sound part contains most of the energy received at the microphones, so that the rank-1 spatial covariance matrix provides similar modeling accuracy than the full-rank model with fewer parameters. However, in an environment with realistic reverberation time, *i.e.* $T_{60} \geq 130$ ms, the full-rank model outperforms both the rank-1 model and binary masking in terms of SDR and SAR and results in a SIR very close to that of binary masking. For instance, with $T_{60} = 250$ ms, the SDR achieved via the full-rank model is 2.7 dB and 1.5 dB larger than that of the rank-1 model and binary masking, respectively. These results confirm the effectiveness of our proposed parameter estimation approach and also show that full-rank spatial covariance matrices better approximate the mixing process in a reverberant room.

5. CONCLUSION

In this paper, we investigated the blind source separation performance stemming from rank-1 and full-rank models of the source spatial covariances. For that purpose, we addressed the estimation of model parameters by maximizing the likelihood of the observed mixture data using the EM algorithm with a proper initialization scheme. Experimental results over speech data confirm that the full-rank model outperforms both the rank-1 model and state-of-the-art approaches, *i.e* binary masking and ℓ_1-norm minimization, in a reverberant environment. Future work will take into account background noise within the models and validate the performance of the proposed algorithms over real-world recordings with a larger number of sources. We will also consider combining the proposed models with models of the source spectra, such as those proposed in [7].

6. REFERENCES

[1] 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, 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] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial covariance models for under-determined reverberant audio source separation," in *Proc. WASPAA*, 2009, pp. 129–132.

[4] E. Vincent, M. G. 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*, W. Wang, Ed. IGI Global, to appear.

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

[6] A. P. Dempster, N. M. Laird, and B. D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, pp. 1–38, 1977.

[7] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation.," in *Proc. ICASSP*, 2009, pp. 3137–3140.

[8] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, July 2007.

[9] E. Vincent, H. Sawada, P. Bofill, S. Makino, and JP. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Proc. ICA*, 2007, pp. 552–559.