Blind Source Separation based on Time-Frequency Sparseness in the Presence of Spatial Aliasing

Benedikt Loesch and Bin Yang

Chair of System Theory and Signal Processing, University of Stuttgart
{benedikt.loesch, bin.yang}@LSS.uni-stuttgart.de

Abstract. In this paper, we propose a novel method for blind source separation (BSS) based on time-frequency sparseness (TF) that can estimate the number of sources and time-frequency masks, even if the spatial aliasing problem exists. Many previous approaches, such as degenerate unmixing estimation technique (DUET) or observation vector clustering (OVC), are limited to microphone arrays of small spatial extent to avoid spatial aliasing. We develop an offline and an online algorithm that can both deal with spatial aliasing by directly comparing observed and model phase differences using a distance metric that incorporates the phase indeterminacy of 2π and considering all frequency bins simultaneously. Separation is achieved using a linear blind beamformer approach, hence musical noise common to binary masking is avoided. Furthermore, the offline algorithm can estimate the number of sources. Both algorithms are evaluated in simulations and real-world scenarios and show good separation performance.

Key words: blind source separation, adaptive beamforming, spatial aliasing, time-frequency sparseness

1 Introduction

The task of convolutive blind source separation is to separate M convolutive mixtures $x_m[i], m = 1, \ldots, M$ into N different source signals. Mathematically, we write the sensor signals $x_m[i]$ as a sum of convolved source signals

$$x_m[i] = \sum_{n=1}^{N} h_{mn}[i] * s_n[i], \quad m = 1 \ldots M \tag{1}$$

Our goal is to find signals $y_n[i]$, $n = 1 \ldots N$ such that, after solving the permutation ambiguity, $y_n[i] \approx s_n[i]$ or a filtered version of $s_n[i]$. In the case of moving sources, the impulse responses $h_{mn}[i]$ are time-varying.

Our algorithms cluster normalized phase vectors $\bar{\mathbf{X}}[k, l]$ in the time-frequency domain, where k is the frequency index and l is the time frame index, respectively. Each cluster with the associated state vector $\mathbf{c}[k, \mathbf{p}_n]$ corresponds to a different source n with the associated location or direction-of-arrival (DOA) parameter vector $\mathbf{p}_n$. Different from DUET, our algorithms can use more than two microphones. In contrast to DUET and OVC, our algorithms do not suffer from the spatial aliasing problem. After clustering the phase vectors $\bar{\mathbf{X}}[k,l]$, we apply time-frequency masking or the blind beamformer from [1] to separate the sources.

2 Proposed Offline Algorithm

After a short-time Fourier transform (STFT), we can approximate the convolutive mixtures in the time-domain as instantaneous mixtures at each time-frequency (TF) point $[k,l]$:

$$\mathbf{X}[k,l] \approx \sum_{n=1}^{N} \mathbf{H}_n[k] S_n[k,l] \tag{2}$$

$\mathbf{X} = [X_1, \ldots, X_M]^T$ is called an observation vector and $\mathbf{H}_n = [H_{1n}, \ldots, H_{Mn}]^T$ is the vector of frequency responses from source n to all sensors. We assume that the direct path is stronger than the multipath components. This allows us to exploit the DOA information to perform the separation. Note that we are only interested in separation and not in dereverberation. Hence, we do not aim at a complete inversion of the mixing process. As a consequence, we do not require minimum-phase mixing. The proposed algorithm consists of three steps: normalization, clustering, and reconstruction of the separated signals.

2.1 Normalization

From the observation vectors $\mathbf{X}[k,l]$, we derive the normalized phase vectors $\bar{\mathbf{X}}[k,l]$ which contain only the phase differences of the elements of $\mathbf{X}[k,l]$ with respect to a reference microphone J:

$$\bar{\mathbf{X}}[k,l] = \left[e^{j \cdot \arg(X_m[k,l]/X_J[k,l])} \right], \quad m = 1, \cdots, M \tag{3}$$

For a single active source, the phase of the ratio of two elements of $\mathbf{X}[k,l]$ is a linear function of the frequency index k (modulo 2π):

$$\arg\left(X_m[k,l]/X_J[k,l]\right) = 2\pi \Delta f k \tau_m + 2\pi o, \quad o \in \mathbb{Z} \tag{4}$$

where Δf is the frequency bin width and τ_m is the time-difference of arrival (TDOA) of the source with respect to microphone m and J. If there is no spatial aliasing (i.e. $o = 0$), we can cluster the TDOAs at all TF points because of $\tau_m = \frac{1}{2\pi\Delta f k} \arg\left[\frac{X_m[k,l]}{X_J[k,l]}\right]$. However, in the case of spatial aliasing ($o \neq 0$), we can no longer cluster $\frac{1}{2\pi\Delta f k} \arg\left[\frac{X_m[k,l]}{X_J[k,l]}\right]$ directly. Instead we would need to take into account all possible values of o. However, we can avoid this problem by directly comparing the observed phase difference and the model phase difference for multiple microphone pairs using the distance metric

$$\|\bar{\mathbf{X}}[k,l] - \mathbf{c}[k,\mathbf{p}]\|^2 = 2M - 2 \cdot \sum_{m=1}^{M} \cos\left(\arg\left[\frac{X_m[k,l]}{X_J[k,l]}\right] - 2\pi\Delta f k \tau_m(\mathbf{p}) \right) \tag{5}$$

with the state vector $\mathbf{c}[k, \mathbf{p}] = [c_m]_{1\leq m\leq M} = \left[e^{j2\pi \Delta f k \tau_m(\mathbf{p})}\right]_{1\leq m\leq M}$. $\mathbf{c}[k, \mathbf{p}]$ contains the expected phase differences for a potential source at $\mathbf{p}$ with respect to microphones $m = 1, \cdots, M$ and J.

The distance metric (5) allows an estimation of the location or DOA parameters $\mathbf{p}_n$ of all sources even if spatial aliasing occurs. This is achieved by considering all frequency bins simultaneously: Due to the spatial aliasing, (5) contains location ambiguities for higher frequencies. However, these ambiguities are removed by summing across all frequency bins. We define $\mathcal{J}(\mathbf{p})$

$$\mathcal{J}_l(\mathbf{p}) = \sum_k \rho(\|\bar{\mathbf{X}}[k,l] - \mathbf{c}[k,\mathbf{p}]\|^2), \quad \mathcal{J}(\mathbf{p}) = \sum_l \mathcal{J}_l(\mathbf{p}), \tag{6}$$

which has maxima for $\mathbf{p} = \mathbf{p}_n$. $\rho(t)$ is a monotoneously decreasing nonlinear function in the range $[0, 1]$ that reduces the influence of outliers and increases spatial resolution.

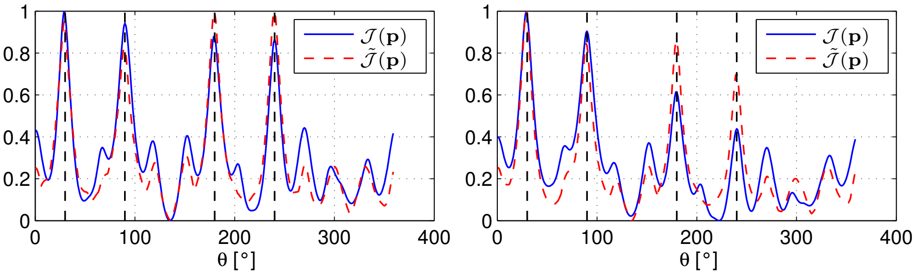
We estimate $\mathbf{p}_n$ by looking for the maxima of $\mathcal{J}(\mathbf{p})$ and then cluster the TF points as described in the next section.

2.2 Source Number Estimation and Clustering

We need to estimate the number of sources $\hat{N}$ and then find clusters $C_1, \ldots, C_{\hat{N}}$ of $\bar{\mathbf{X}}[k, l]$ with centroids $\mathbf{c}[k, \hat{\mathbf{p}}_n]$. Unlike [2–4], we achieve the clustering by a direct search over all possible TDOAs or DOAs $\mathbf{p}$ and do not use iterative approaches such as k-means or expectation-maximization (EM). This has the advantage, that we are guaranteed to find the global optima of the cost function. Inspired from [5], we propose to use $\rho(t) = 1 - \tanh(\alpha\sqrt{t})$ as the nonlinear function $\rho(t)$ in (6). Independently of our research, [6] proposed a similar cost function $\mathcal{J}_l$ for only two microphones and without the summation over time for localization purposes.

Another advantage of the direct search is that we do not need to know the number of sources beforehand as in [2]. Instead, we can count the number of significant and distinct peaks of $\mathcal{J}(\mathbf{p})$: This is done by finding all peaks $\mathbf{p}_n$ of $\mathcal{J}(\mathbf{p})$ with $\mathcal{J}(\mathbf{p}) > t$ and sorting the peaks in descending order $\mathcal{J}(\mathbf{p}_i) > \mathcal{J}(\mathbf{p}_{i+1})$. Then we start from the first peak and accept the next peak if the minimum distance to a previously accepted peak is larger than a certain threshold t_2. The number of estimated sources $\hat{N}$ is then given as the number of accepted peaks.

Since in (6) the peak height is a function of the amount of source activity it might be difficult to count the number of sources if the amount of source activity differs a lot among the sources. One way to solve this problem is to use the max-approach from [6] to estimate the number of sources by replacing $\mathcal{J}(\mathbf{p})$ by $\tilde{\mathcal{J}}(\mathbf{p}) = \max_l \mathcal{J}_l(\mathbf{p})$. This modified function is less sensitive to the amount of source activity because the peak height is proportional to the coherence of the observed phase. So if for each source there is at least one time frame where it is the single active source, $\tilde{\mathcal{J}}(\mathbf{p})$ will yield a large peak for this source. Fig. 1 shows $\mathcal{J}(\mathbf{p})$ and $\tilde{\mathcal{J}}(\mathbf{p})$ for two scenarios with different amounts of source activity. In Fig. 1(b) the max-approach is clearly superior because the contrast beetween true peaks and spurious peaks is larger. Furthermore, the max-approach improves TDOA estimation for closely spaced microphones: It selects time frames with high coherence of the observed phase, i.e. a single active source.



(a) Sources always active (b) Src 3: start 10 s, Src 4: start 14 s

Fig. 1. Source Number and DOA Estimation for different source acitivity (length 24 s)

The positions/DOAs $\hat{\mathbf{p}}_n, n = 1, \cdots, \hat{N}$ of the sources are given by the relevant peaks of $\mathcal{J}(\mathbf{p})$ or $\tilde{\mathcal{J}}(\mathbf{p})$. For each source, we generate the corresponding state vectors $\mathbf{c}[k, \hat{\mathbf{p}}_n], n = 1, \cdots, \hat{N}$ and assign all TF points to cluster n for which $\|\bar{\mathbf{X}}[k, l] - \mathbf{c}[k, \hat{\mathbf{p}}_n]\|^2$ is minimal.

Comparison with EM algorithm with GMM model : [4] uses a related approach for two microphones: They use an EM algorithm with a Gaussian mixture model (GMM) for the phase difference between the two microphones at each TF point. The phase difference of each source is modelled as a mixture of $2K_f + 1$ Gaussians with mean $2\pi k \cdot (1 + \Delta f \mu_q), k = -K_f, \cdots, K_f$, where μ_q is the mean of the q-th component. The observed phase difference for all sources is then described as a mixture of Q such models. Furthermore they use a Dirichlet prior for the mixture weights $\alpha_q, q = 1, \cdots, Q$ to model the sparsity of the source directions, i.e. to represent the phase difference with a small number of Gaussians with large weight α_q. After convergence of the EM algorithm the mean μ_q of the Gaussians with large weight α_q and small variance σ_q^2 reflect the estimated TDOAs of the N sources. Our approach differs in a number of ways:

- We use a direct search instead of an iterative procedure to estimate the source parameters $\mathbf{p}_n$.
- We are guaranteed to find the global optima of the function $\mathcal{J}(\mathbf{p})$, whereas the EM algorithm could converge to local optima.
- We estimate the number of sources by counting the number of significant and distinct peaks instead of checking the weights and variance of the components of a GMM model.
- We do not model the phase difference of each source using $2K_f + 1$ Gaussians. Instead we use the distance metric (5) which incorporates the phase wrapping.

- Our approach is computationally less demanding: We need $N_{\text{grid}} \cdot K \cdot L$ function evaluations, while the EM algorithm requires $N_{\text{iter}} \cdot Q \cdot (2K_f + 1) \cdot K \cdot L$ function evaluations. For a typical scenario $N_{\text{grid}} = 180, N_{\text{iter}} = 10, Q = 8, K_f = 5$ and assuming comparable computational cost for each function evaluation, our approach would be about 6 times faster.

The differences between [4] and our approach can be summarized as differences in the model (wrapped Gaussians vs. 2π-periodic distance function) and in the clustering algorithm (iterative EM algorithm vs. direct search and simple clustering).

The authors would like to thank Dr. Shoko Araki for running our proposed algorithm on her dataset from [4]. Using $t = 0.5$ and $t_2 = 5°$, our algorithm estimates the number of sources correctly for all tested cases.

2.3 Reconstruction

To reconstruct the separated signals, we use the blind beamforming approach discussed in [1]. This approach reduces or completely removes musical noise artifacts which are common in binary TF mask based separation. After the beamforming step, additional optional binary TF masks can be used to further suppress the interference. Then we convert the separated signals back to the time domain using an inverse STFT.

3 Online Separation Algorithm

The online algorithm operates on a single frame basis and uses a gradient ascent search for the TDOAs/DOAs of the sources with prespecified maximum number of sources N. Inactive sources result in empty clusters. The cost function $\mathcal{J}_l$ for the l-th STFT frame is

$$\mathcal{J}_l(\mathbf{p}_n) = \sum_k \rho(\|\bar{\mathbf{X}}[k,l] - \mathbf{c}[k,\mathbf{p}_n]\|^2) \tag{7}$$

Its gradient vector with respect to $\mathbf{p}_n$ is

$$\frac{\partial \mathcal{J}_l}{\partial \mathbf{p}_n} = -\sum_k \Re \left\{ j2\pi \Delta f k \cdot \frac{\partial \rho(t)}{\partial t} \cdot \frac{\partial \boldsymbol{\tau}}{\partial \mathbf{p}_n} \cdot \mathbf{c}[k,\mathbf{p}_n] \odot 2(\bar{\mathbf{X}}[k,l] - \mathbf{c}[k,\mathbf{p}_n]) \right\}, \tag{8}$$

where $\boldsymbol{\tau} = [\tau_m]_{1 \leq m \leq M}$. Similar to [7], we use a time-varying learning rate which is a function of the amount of TF points associated with source n. The separation steps are similar to the offline algorithm. A computationally efficient version of the blind beamforming algorithm can be obtained by using recursive updates of the parameters as in [8].

4 Experimental Evaluation

For the experimental evaluation, we used a sampling frequency of $f_s = 16\,\text{kHz}$, a STFT with frame length 1024 and 75% overlap. SNR was set to 40 dB. We selected six speech signals from the short stories of the CHAINS corpus [9].

4.1 Stationary Sources

First, we evaluate the offline algorithm using the measured impulse responses of room E2A ($T_{60} = 300\,\text{ms}$) from the RWCP sound scene database [10]. This database contains impulse responses for sources located on a circle with radius 2.02 m with an angular spacing of 20°. The microphone array is shifted by 0.42 m with respect to the center of the circle. We consider a two-microphone array with the spacings $d = 11.3\,\text{cm}$ and $d = 33.9\,\text{cm}$. All sources have equal power and a length of 5 s. For $d = 11.3\,\text{cm}$, $\mathcal{J}(\mathbf{p})$ sometimes does not show N distinct maxima. Hence, we use $\tilde{\mathcal{J}}(\mathbf{p})$ for $d = 11.3\,\text{cm}$ since it provides increased resolution by using frames with a single active source for the localization. We have tried different signal combinations and DOA scenarios (A,B,C,D): For scenarios A,B,C we varied the DOAs between $30^\circ \dots 150^\circ$ and tested different angular spacings between the sources. For case D we distributed the sources with maximum angular spacing between $10^\circ \dots 170^\circ$. The angular spacings between the sources are summarized in Table 1. For each scenario, we considered all signal combinations of $N = 2, 3, 4$ out of the 6 source signals. Table 2 summarizes the performance of the source number estimation and the signal-to-interference (SIR) gain for the different cases. For the source number estimation, we used thresholds $t = 0.46$ for $d = 11.3\,\text{cm}$ and $t = 0.55$ for $d = 33.9\,\text{cm}$. t_2 was set to 10° for both microphone spacings d. The performance is evaluated by the percentage of estimations for which $\hat{N} = n, n = 1, \cdots, 5$. This is known as the confusion matrix.

Table 1. Considered scenarios

	A	B	C	D
$N = 2$	20°	–	40°	160°
$N = 3$	20°/40°	40°/20°	40°/40°	80°/80°
$N = 4$	20°/40°/40°	40°/20°/40°	40°/40°/40°	60°/40°/60°

Table 2. Source number estimation and SIR gain

		$\hat{N}$ accuracy (%)					SIR gain [dB]			
N	spacing	1	2	3	4	5	A	B	C	D
2	$d = 11.3\,\text{cm}$	0	97	3	0	0	4.6	–	7.9	15.9
	$d = 33.9\,\text{cm}$	0	100	0	0	0	6.0	–	9.0	13.0
3	$d = 11.3\,\text{cm}$	0	1	98	1	0	3.7	5.7	8.7	11.4
	$d = 33.9\,\text{cm}$	0	0	100	0	0	6.2	8.9	9.4	13.1
4	$d = 11.3\,\text{cm}$	0	0	2	95	3	2.8	6.0	7.1	6.4
	$d = 33.9\,\text{cm}$	0	0	0	100	0	6.5	8.9	8.0	8.4

Our offline algorithm estimates the number of sources correctly in almost all cases and shows a good separation performance. A larger microphone spacing achieves a better separation performance for closely spaced sources (case A and B) or for large source numbers. On the short two-source-two-microphone mixtures of the SISEC2010 campaign (`http://irisa.fr/metiss/SiSEC10/short/short_all.html`), separation performance is comparable to other algorithms.

4.2 Moving Sources

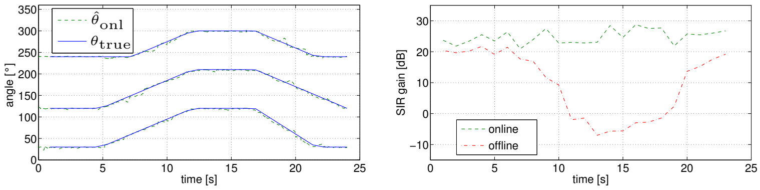
In order to have a precise reference for moving source positions, we performed simulations using the MATLAB ISM RoomSim toolbox [11]. The considered room was of size 5 m × 6 m × 2.5 m and we chose reverberation times of $T_{60} = 100, 200, 300\,\mathrm{ms}$. We used a cross-array (⁜) with $M = 5$ microphones. The microphone spacing was $d = 10\,\mathrm{cm}$, so spatial aliasing occurs above 1700 Hz . We have $N = 3$ sources that move along a circle with radius 1.0 m. Source 1 moves from $\theta_1 = 30°$ to $\theta_1 = 120°$ and back, source 2 moves from $\theta_2 = 120°$ to $\theta_2 = 210°$ and back and source 3 moves from $\theta_3 = 240°$ to $\theta_3 = 300°$. The total simulation time is 24 s. Fig. 2(a) shows the estimated angles $\hat{\theta}_{\mathrm{onl}}[l]$ using our online algorithm as well as the reference angles $\theta_{\mathrm{true}}[l]$. The online-algorithm was initialized with the true angles $\theta_{\mathrm{true}}[0]$. As we see, the online algorithm accurately tracks the sources. During speech pauses, angle estimates are not updated. The separation performance of the online and offline algorithm is summarized in Table 3. As expected, the offline algorithm fails to separate the source signals while our gradient-based online algorithm achieves good results. The reason for the failure of the offline algorithm is that it averages the DOAs of the moving sources. The estimated DOAs are 63°, 119° and 240°. Two of the three estimated DOAs match the initial and final DOAs of the two sources and hence separation quality for these two sources (2 and 3) is acceptable for $T_{60} = 100\,\mathrm{ms}$. However, the separation performance drops significantly when the sources start moving as shown in Fig. 2(b). It shows the local SIR gains which are calculated over non-overlapping segments of 1 second and averaged over the three sources.

Table 3. Global SIR gain in dB for moving sources

T_{60}	algorithm	source 1	source 2	source 3	mean
100 ms	offline	4.5	18.2	10.1	10.9
	online	25.9	28.1	27.8	27.3
200 ms	offline	5.0	15.9	8.5	9.8
	online	24.6	22.4	24.1	23.7
300 ms	offline	4.9	10.7	6.9	7.5
	online	17.3	16.5	18.0	17.3

5 Conclusion

In this paper we have presented two blind source separation algorithms based on TF sparseness that are able to deal with the spatial aliasing problem by using



(a) DOA tracking (b) Local SIR gain

Fig. 2. Fixed number of moving sources, $T_{60} = 200\,\mathrm{ms}$

a distance metric which incorporates phase wrapping (mod 2π) and averaging all frequency bins for the estimation of the location or DOA parameters of the sources. The offline algorithm reliably estimates the source number and achieves the clustering using a direct search. The online algorithm assumes a prespecified maximum number of sources and is able to track moving sources. Both algorithms show good separation performance in midly reverberant environments.

References

1. J. Cermak, S. Araki, H. Sawada, and S. Makino, “Blind source separation based on a beamformer array and time frequency binary masking,” *Proc. ICASSP*, 2007.
2. S. Araki, H. Sawada, R. Mukay, and S. Makino, “Underdetermined blind sparse source separation of arbitrarily arranged multiple sensors,” *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.
3. H. Sawada, S. Araki, R. Mukay, and S. Makino, “Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.
4. S. Araki, T. Nakatani, H. Sawada, and S. Makino, “Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem,” *Proc. ICA*, 2009.
5. F. Nesta, M. Omologo, and P. Svaizer, “A novel robust solution to the permutation problem based on a joint multiple TDOA estimation,” *Proc. International Workshop for Acoustic Echo and Noise Control (IWAENC)*, 2008.
6. Z.E. Chami, A. Guerin, A. Pham, and C. Serviere, “A phase-based dual microphone method to count and locate audio sources in reverberant rooms,” *Proc. WASPAA*, 2009.
7. S. Rickard, R. Balan, and J. Rosca, “Real-time time-frequency based blind source separation,” *Proc. ICA*, 2001.
8. B. Loesch and B. Yang, “Online blind source separation based on time-frequency sparseness,” *Proc. ICASSP*, 2009.
9. F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, “The CHAINS corpus (characterizing individual speakers),” `http://chains.ucd.ie/`, 2006.
10. Real World Computing Partnership, “RWCP Sound Scene Database in Real Acoustic Environment,” `http://tosa.mri.co.jp/sounddb/indexe.htm`, 2001.
11. E. Lehmann, “Image-source method for room impulse response simulation (room acoustics),” `http://www.watri.org.au/~ericl/ism_code.html`, 2008.