SOURCE NUMBER ESTIMATION AND CLUSTERING FOR UNDERDETERMINED BLIND SOURCE SEPARATION

Benedikt Loesch and Bin Yang

University of Stuttgart
Chair of System Theory and Signal Processing
http://www.LSS.uni-stuttgart.de

ABSTRACT

Much research has been undertaken in the field of blind source separation (BSS) and a large number of algorithms have been developed. However, most of them assume that the number of sources is known. In this paper we present an algorithm to estimate the number of sources in the (over-)determined and underdetermined case. We call this algorithm NOSET (Number of Sources Estimation Technique). We start from a description of the BSS problem, give a short overview of the so-called observation vector clustering algorithm and then present our approach. It is based on direction-of-arrival (DOA) estimation from reliable time-frequency points and a clustering of the DOA estimates. The estimated DOAs can be used to recover the source signals by performing a nearest-neighbor classification of the observation vectors instead of the conventional k-means clustering procedure which is sensitive to the choice of initial centroids.

1. INTRODUCTION

This paper deals with the estimation of the number of sources for blind source separation. The task of blind source separation is to separate M (possibly) convolutive mixtures $x_m[i], m = 1, \ldots, M$ into N different source signals. Mathematically we can 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 $s_n[i]$. Regarding the number of sources N and the number of sensors/mixtures M, we distinguish between three different cases: overdetermined ($M > N$), determined ($M = N$), and underdetermined ($M < N$).

It is clear that the underdetermined case is the most challenging one since we try to find more signals than we have measurements. Although there is already a big variety of proposed algorithms, most of them assume that the number of sources is known. Up to now there are only a few papers [1, 2] discussing ways to estimate the number of sources in an underdetermined situation where classical model order estimation approaches (such as MDL and AIC using the eigenvalues of the correlation matrix) are not appplicable. [1, 2] make use of the sparsity of speech signals but are limited to the anechoic case. Other algorithms that aim at estimating the number of sources as well as the source signals use hierarchical clustering and l_1-norm minimization [3]. To the authors knowledge, there is no thorough study yet of how reliable these algorithms are.

In our work, we focus on BSS algorithms that make use of the sparsity of speech signals to derive binary time-frequency (TF) masks in order to separate speech signals. Early work using only two microphones resulted in the well-known DUET algorithm and since then many enhancements have been made and led to the so-called observation vector clustering algorithm [4]. Our NOSET algorithm makes use of the observation vector and replaces the conventional k-means clustering (which needs to know the number of sources) with a method to estimate the number of sources and also the clusters. The next section presents a short overview of the observation vector clustering algorithm, before we present the NOSET algorithm in detail in section 3.

2. OBSERVATION VECTOR CLUSTERING

Using a windowed short time Fourier transform (STFT)

$$X_m[k, l] := \sum_{i=0}^{L-1} w[i] x_m[i + lT_0] e^{-j\frac{2\pi ki}{L}} \tag{2}$$

with hop size T_0 and bin width $f_0 = f_s/L$, we can approximate the convolutive mixtures in the time-domain as instantaneous mixtures at each frequency bin k:

$$\mathbf{X}[k, l] \approx \sum_{n=1}^{N} \mathbf{H}_n[k] S_n[k, l] \tag{3}$$

$\mathbf{X} = [X_1, \ldots, X_M]^T$ 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 microphone array is placed in the near-field of the sources which means that we can assume a strong direct-path and weak multipath components. The algorithm consists of three steps: normalization, clustering, and reconstruction of the separated signals.

2.1. Normalization

All observation vectors $\mathbf{X}[k, l]$ are normalized for all frequency bins to form clusters, each of which corresponds to an individual source. The normalization is performed with respect to a reference sensor J

$$\tilde{X}_m[k, l] = |X_m[k, l]| \exp\left[j \frac{\arg[X_m[k, l]/X_J[k, l]]}{4kf_0 c^{-1} d_{\max}}\right] \tag{4}$$

where c is the propagation speed and $d_{\max}$ the maximum distance between any sensor and the reference sensor. After the phase normalization in (4) we apply unit-norm normalization

$$\bar{\mathbf{X}}[k, l] = \tilde{\mathbf{X}}[k, l]/\|\tilde{\mathbf{X}}[k, l]\| \tag{5}$$

Note that observation vector clustering suffers from the same phase ambiguity problem as the DUET algorithm and hence $d_{\max} \leq c/f_s$,

where f_s is the sampling frequency. The reason for the phase normalization by $4kf_0c^{-1}d_{\max}$ in (4) is explained in [4].

2.2. Clustering

The next step is to find clusters $C_1, \ldots, C_N$ formed by all normalized observation vectors $\bar{\mathbf{X}}[k,l]$. This can be done with the k-means clustering algorithm [4] using the cost function

$$\mathcal{J} = \sum_{n=1}^{N} \mathcal{J}_n, \quad \mathcal{J}_n = \sum_{\bar{\mathbf{X}} \in C_n} \|\bar{\mathbf{X}} - \mathbf{c}_n\|^2 \tag{6}$$

where $\mathbf{c}_n$ is the centroid of cluster C_n.

2.3. Reconstruction

To reconstruct the signals, the authors in [4] designed a binary TF mask $M_n[k,l]$ that extracts the TF points in each cluster and obtained the separated signals $Y_n[k,l]$ by

$$Y_n[k,l] = M_n[k,l]X_{J'}[k,l] \tag{7}$$

where $J' \in \{1, \ldots, M\}$ can be arbitrarily chosen. To obtain the separated signals in the time-domain, we perform an inverse STFT. Note that, in addition to the simple binary masking, we can also perform blind beamforming as discussed in [5]. This approach has the advantage of reducing or completely removing musical noise artifacts which are common in binary TF mask based separation.

2.4. Problems of k-means Algorithm

We want to note that the k-means clustering has the following drawbacks that limit its applicability in real world situations:

- The number of clusters needs to be known a priori. Usually we do not know the number of sources and hence cannot specify the number of clusters to be created.
- The clustering procedure is sensitive to initial centroids. Traditional approaches to run the algorithm several times with different initial centroids and select the result with the lowest squared error $\mathcal{J}$ might not be optimal in terms of signal-to-interference ratio of the demixed signals.
- We have observed cases where k-means clustering moves the centroids away from their true positions, resulting in very poor separation performance. This can be explained such that, in reality, clusters might overlap by a great amount and have "heavy tails" since not all TF points return a reliable observation vector.

In fact, the goal of a sparsity based BSS algorithm is to reliably estimate the number of sources and to find cluster centroids with high density of points rather than minimizing the overall squared error $\mathcal{J}$. In section 3 we will present the NOSET algorithm which aims to achieve this. Our contribution is two-fold: The NOSET algorithm estimates the number of sources by selecting reliable TF points for DOA estimation and uses the peaks of the DOA histogram as centroids for a one-step clustering procedure for all TF points. Thus we are able to overcome the limitations of the conventional k-means clustering. The motivations for the NOSET algorithm are:

- by selecting a subset of reliable TF points, we are able to estimate DOAs more reliably and reduce at the same time the amount of computations.
- in the DOA domain, we can estimate the source number more conveniently than in the observation vector domain.
- by considering all TF points in the source separation stage, we can achieve a high signal quality of the demixed signals.

3. SOURCE NUMBER ESTIMATION

Our algorithm NOSET to estimate the number of sources is based on a DOA estimation. Due to the convolutive mixing process and the fact that the sources are not perfectly disjoint in the TF domain, the observation vectors that belong to one source are spread around the "true" DOA. As a consequence, clusters overlap by a great amount and we cannot find disjoint clusters. In order to overcome this problem, we perform a pre-processing step to select only the "reliable" TF points. A reliable DOA estimation using phase differences is only possible if the following two conditions are fulfilled:

- The phase difference among different sensors is large enough. In the low-frequency region, this is not the case and the phase estimate is rather noisy.
- Only one source is dominant at a TF point $[k,l]$.

The first step in finding reliable TF points is hence to limit the frequency region: $f > f_l$. The next step is to detect TF points where one source is dominant. We will call those points one-source TF points in the following.

3.1. Selection of One-Source TF Points

In this section we show a simple and efficient way to find one-source TF points.

Analysis of source activity: We assume that we have perfect knowledge about the contribution of each individual source at each sensor. Let s_{Jn}^{img} denote the *image* of source n at reference sensor J: $s_{Jn}^{img}[i] = h_{Jn}[i] * s_n[i]$. The power of source n at TF point $[k,l]$ is denoted as $P_n[k,l] = 20\log_{10}|S_{Jn}^{img}[k,l]|$. These signals are available in simulations. We consider a source s_n to be active at a TF point $[k,l]$ if

- $P_n[k,l] > t_{\text{noise}}$ where t_{noise} is the noise floor in dB.
- $P_n[k,l] > (\max_r P_r[k,l]) - t_1, r = 1, \ldots, N$

In other words, the power of the active source is above the noise floor and its power is not less than t_1 dB below the power of the strongest source. Let $\hat{N}_{\text{cp}}[k,l]$ be the number of strong sources with comparable power at TF point $[k,l]$, i.e. the number of sources satisfying the above two properties. This number can be calculated in simulations. The probability of $\hat{N}_{\text{cp}}[k,l] = n$ conditioned on total power $P[k,l] = 20\log_{10}|X_J[k,l]| > t$ is

$$Pr[\hat{N}_{\text{cp}} = n | P > t] = \frac{Pr[P > t | \hat{N}_{\text{cp}} = n] Pr[\hat{N}_{\text{cp}} = n]}{Pr[P > t]} \tag{8}$$

$$Pr[P > t | \hat{N}_{\text{cp}} = n] = \int_t^\infty p(P = b | \hat{N}_{\text{cp}} = n) db \tag{9}$$

$$Pr[P > t] = \int_t^\infty p(P = b) db \tag{10}$$

The pdf $p(P = b)$ and conditional pdf $p(P = b|\hat{N}_{\text{cp}} = n)$ are approximated in simulations by power histograms over all TF points or all TF points where $\hat{N}_{\text{cp}} = n$. $Pr[\hat{N}_{\text{cp}} = n]$ is the fraction of TF points with $\hat{N}_{\text{cp}} = n$. Note that $\hat{N}_{\text{cp}}[k,l]$ does not denote the conventional number of active sources at a TF point, but rather the number of strong sources with comparable power.

Example: We play back $N = 2$ speech signals in a real office room. Figure 3.1 shows the conditional probability $Pr[\hat{N}_{\text{cp}} = n | P > t]$ for $n = 0, 1, 2$. Obviously, if the total power P exceeds a certain threshold (e.g. 25 dB), it is much more likely that we have one dominant source at a TF point rather than no source or two strong sources with comparable power.

Selection of reliable TF points: Motivated by this observation, we propose to select reliable TF points using a combined power and frequency criterion

$$\mathcal{I} = \{[k,l] \mid P[k,l] > t_2 \wedge f > f_l\} \tag{11}$$

We note that this criterion selects only a small subset (e.g. 25%) of the TF points. Very weak sources might have only a small number of TF points or even no TF points at all with enough signal power to fulfill the power criterion and hence might be discarded. In reality, this is usually not a problem, since the signal quality would be very bad if we still try to demix those very weak signals.

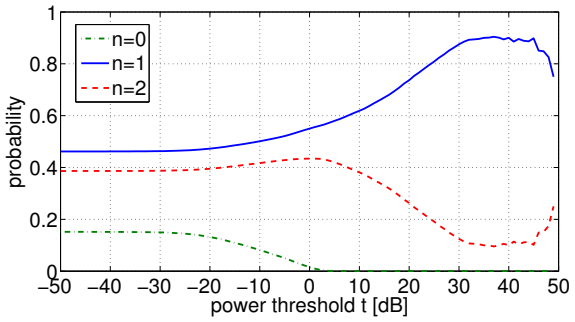


Fig. 1. Probability $Pr[\hat{N}_{\text{cp}} = n \mid P > t]$ as a function of t

Comparison with other criteria: In order to evaluate the selection of one-source TF points, we compare our selection criterion (11) with three other methods:

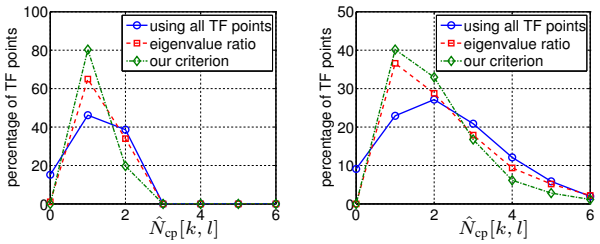
- No selection of TF points at all, i.e. all TF points are used to calculate the observation vector in (4) or other features.
- Use the eigenvalue ratio $\lambda_1 / \sum_{i=2}^{N} \lambda_i$ from [2] to select single-source TF points. In this case, λ_i are the decreasingly sorted eigenvalues of a correlation matrix calculated over a small TF region around the TF point of interest. A large value of this ratio indicates a single dominant source. We selected the same total number of TF points as with our criterion.
- Estimate the number of active sources by using AIC or MDL from the eigenvalues of the above described correlation matrix and select the one-source TF points based on this order estimation. Unfortunately, both AIC and MDL tend to overestimate. They never returned TF points with order estimate of one. This is due to the fact that they are both derived for the anechoic case and tend to overestimation in a reverberant environment.

Figure 2 shows the percentage of TF points with different number of active sources $\hat{N}_{\text{cp}}[k,l]$. In the left and right plot, $N = 2$ and 6 speech signals are played back in a real office room. Clearly, our selection criterion (11) achieves the highest percentage of one-source TF points among all studied methods. In addition, the criterion (11) is very simple to implement in comparison to the two other criteria requiring eigenvalue decomposition. Our criterion improves DOA estimation while reducing simultaneously the computational complexity since we consider a smaller number of TF points.

3.2. DOA Estimation

Starting from the observation vectors $\bar{\mathbf{X}}$ in (5) at the selected TF points $[k,l] \in \mathcal{I}$, we now determine the DOA of the single dominant source from the phase differences. Assuming a single source anechoic scenario, the time delay δ_m for sensor m can be written as a scalar-product of the microphone position vector $\mathbf{d}_m$ and the source direction vector $\mathbf{q}$:

$$\delta_m = \mathbf{d}_m^T \mathbf{q}, \quad \mathbf{q} = [\sin\theta\cos\phi, \cos\theta\cos\phi, \sin\phi]^T \tag{12}$$



(a) $N = 2$ (b) $N = 6$

Fig. 2. Effect of different TF selection criteria

According to eq. (4), we obtain

$$\arg([\bar{\mathbf{X}}]_m) = \frac{2\pi k f_0 (\delta_m - \delta_J)}{4k f_0 c^{-1} d_{\max}} = \frac{\pi}{2c^{-1} d_{\max}} (\mathbf{d}_m - \mathbf{d}_J)^T \mathbf{q} \tag{13}$$

This equation also holds approximately for a convolutive near-field scenario. We estimate the source direction vector by the least-squares approach:

$$\tilde{\mathbf{q}} = \frac{2c^{-1} d_{\max}}{\pi} \mathbf{D}^\dagger \mathbf{r}, \quad \hat{\mathbf{q}} = \frac{\tilde{\mathbf{q}}}{\|\tilde{\mathbf{q}}\|} \tag{14}$$

$\mathbf{D} = [\mathbf{d}_1 - \mathbf{d}_J, \cdots, \mathbf{d}_M - \mathbf{d}_J]^T$ is a matrix of the relative sensor locations, $\mathbf{r} = [\arg([\bar{\mathbf{X}}]_1), \cdots, \arg([\bar{\mathbf{X}}]_M)]^T$, and $\mathbf{D}^\dagger$ is the Moore-Penrose pseudo-inverse of $\mathbf{D}$. In this paper, we consider 2D localization only ($\phi = 0$). The bearing θ is thus estimated by

$$\theta = \arctan \frac{[\hat{\mathbf{q}}]_1}{[\hat{\mathbf{q}}]_2} \tag{15}$$

3.3. Source Number Estimation

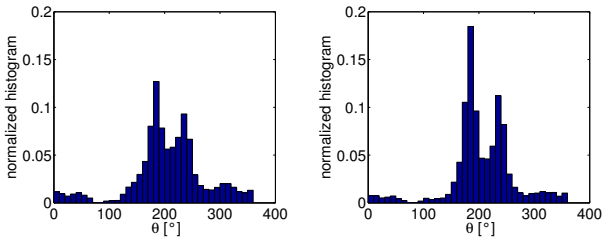
The estimation of the source number $\hat{N}$ is done in 5 steps:

1. Select reliable TF points $\mathcal{I}$ using (11).
2. Estimate the bearing θ from all TF points in $\mathcal{I}$ using (14)-(15).
3. Form a histogram $R[\nu]$ of all θ with a certain bin width. $R[\nu]$ is the number of bearing estimates that fall into bin number ν. Then we subtract off the minimal value of $R[\nu]$. This reduces the noise and improves the quotient of peak values calculated in the next step. Next we find all peaks $R_n, n = 1, \ldots, Q$.
4. Order the peaks in decreasing order ($R_{n+1} \leq R_n$) and calculate $p[n] = R_n / R_1$.
5. The estimated number of sources $\hat{N}$ is determined by counting the number of peaks where $p[n]$ is larger than a threshold t_3. Mathematically: $\hat{N} = \max\{n \mid p[n] \geq t_3\}$. This operation aims to remove weak peaks which might be the result of source overlap or reverberation.

Figure 3 shows a comparison of the DOA histogram with and without (w/o) our selection criterion (11) for $N = 2$. We can see that our criterion improves the DOA histogram.

3.4. Clustering

After estimating the number of sources $\hat{N}$, we convert the DOAs of the peaks $R_n, n = 1, \ldots, \hat{N}$ back to the observation vectors $\mathbf{c}_n$ using eqs. (13) and (15), assuming equal amplitude for all components. $\mathbf{c}_n$ are normalized to unit length and serve as the centroids for $\hat{N}$ clusters. Then we employ a one-step clustering which calculates cluster membership by minimizing the euclidean distance between $\bar{\mathbf{X}}[k,l]$ and $\mathbf{c}_n, n = 1, \ldots, \hat{N}$. This corresponds to the first step of k-means without changing the centroids, an operation which turns



(a) w/o selection, 33125 TF points (b) with selection, 7524 TF points

Fig. 3. Effect of our selection criterion (11) on DOA histogram

out to be critical. All TF points are used in clustering. Afterwards we perform binary TF masking or blind beamforming to reconstruct the separated signals.

4. EXPERIMENTAL EVALUATION

For the experimental evaluation we used a sampling frequency of $f_s = 8\,\mathrm{kHz}$ and a cross-array with $M = 5$ microphones (⁘) with uniform spacing of $d = 4\,\mathrm{cm} < \lambda_{min}/2$. Sources had equal power and were placed approximately $80\,\mathrm{cm}$ from the center of the microphone array, which justifies the assumption of a strong direct-path and weak multipath components made in section 2. We evaluated NOSET by using 16 sets of 6 speech signals (3 male, 3 female, different for each of the 16 sets) from the TIMIT database [6]. For each number of sources $N = 1, \ldots, 6$, we created all possible combinations of signals within each set. We considered different angular separations $(36^\circ, 45^\circ, 60^\circ)$ of sources and different rotations of both microphone array and sources. All evaluations were done using real recorded signals in an office room with a reverberation time of $T_{60} = 520\,\mathrm{ms}$. The average SNR was between 20 and 30 dB.

4.1. Source Number Estimation

Table 1 shows the confusion matrix of the source number estimation averaged over all experiments. The thresholds for the different criteria used in the NOSET algorithm have been determined through a series of experiments. Typical values are: $f_l = 250\,\mathrm{Hz}$, $t_2 = 20\,\mathrm{dB}$, $t_3 = 0.2$. The performance of our order estimator NOSET is excellent for the whole range of $N = 1, \ldots, 6$, although it slightly decreases with increasing N. The average probability of correct source number estimation is 92.0%. Experiments with unequal source power show that the algorithm also works satisfactory for source power differences of up to $10\,\mathrm{dB}$. On average, the probability of correct source number estimation drops to a value between 65% and 90% depending on the amount of source power difference.

		$\hat{N}$						
		1	2	3	4	5	6	7
N	1	**99.6%**	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%
	2	0.1%	**99.1%**	0.8%	0.0%	0.0%	0.0%	0.0%
	3	0.0%	0.5%	**96.2%**	3.3%	0.0%	0.0%	0.0%
	4	0.0%	0.0%	1.5%	**92.0%**	6.5%	0.0%	0.0%
	5	0.0%	0.0%	0.0%	4.3%	**86.0%**	9.2%	0.5%
	6	0.0%	0.0%	0.0%	2.0%	8.9%	**79.5%**	9.8%

Table 1. Overall confusion matrix for NOSET

4.2. Source Separation

As we have seen in section 2, source separation using the conventional k-means clustering has three drawbacks:

- Number of clusters needs to be known
- Sensitivity to initial centroids
- Occasional convergence to bad clusters.

This is illustrated in table 2 with $N = 2$ source speech signals at the true DOAs θ_{true}. The k-means algorithm applied to the observation vectors of all TF points with perfect knowledge of the number of sources and perfect initialization of the cluster centroids with θ_{true} converges to wrong cluster centroids corresponding to $\theta_{\text{k-means}}$ because the k-means algorithm moves the centroids away from their true positions. Correspondingly, the gain in signal-to-interference ratio (SIR) after the blind source separation SIR_{gain} is pretty poor. In comparison, by using our NOSET algorithm for order estimation and the much simpler clustering in section 3.4, the DOA estimates θ_{our} by using a DOA histogram bin width of 5° are more accurate. In addition, the gain in SIR after BSS is improved by $6.8\,\mathrm{dB}$ in average.

	Source 1	Source 2	Average
θ_{true}	299.5°	358.2°	
$\theta_{\text{k-means}}$	312.5°	154.7°	
θ_{our}	302.5°	357.5°	
SIR_{in}	$-1.99\,\mathrm{dB}$	$1.99\,\mathrm{dB}$	$0.00\,\mathrm{dB}$
SIR_{gain} k-means	$2.09\,\mathrm{dB}$	$2.67\,\mathrm{dB}$	$2.38\,\mathrm{dB}$
SIR_{gain} NOSET	$12.67\,\mathrm{dB}$	$5.71\,\mathrm{dB}$	$9.19\,\mathrm{dB}$

Table 2. Comparison of DOA estimation and SIR gain for binary TF masking using k-means and our algorithms

5. CONCLUSION

In this paper, we have presented the NOSET algorithm to estimate the number of sources in blind source separation. It relies on DOA estimation at selected one-source TF points and works in both overdetermined and underdetermined situations. We also showed a simple clustering method which solves the occasional convergence of the k-means clustering to bad clusters. The combination of NOSET and the proposed clustering method is computationally simple and achieves better performance than k-means clustering based blind source separation.

6. REFERENCES

[1] R. Balan, "Estimator for number of sources using minimum description length criterion for blind sparse source mixtures," *Int. Conf. Independent Component Analysis and Blind Source Separation*, 2007.

[2] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear anechoic mixture," *Proc. ICASSP*, 2007.

[3] 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*, 2007.

[4] S. Araki, H. Sawada, R. Mukay, and S. Makino, "Underdetermined blind sparse source separation of arbitrarily arranged multiple sensors," *Signal Processing*, 2007.

[5] 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.

[6] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallet, and Nancy S. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CDROM," `www.ldc.upenn.edu/lol/docs/TIMIT.html`, 1986.