# BLIND SOURCE SEPARATION OF ACOUSTIC MIXTURES WITH DISTRIBUTED MICROPHONES

*Enrique Robledo-Arnuncio & Biing-Hwang (Fred) Juang*

Center for Signal and Image Processing
Georgia Institute of Technology
75 Fifth Street NW, Atlanta, GA 30308, USA
{era,juang}@ece.gatech.edu

## ABSTRACT

The problem of blind source separation of acoustic mixtures is often addressed using independent component analysis in the frequency domain. One problem with this approach is the inconsistency across frequency in the permutation of the source estimates. Solutions to this problem have been proposed that exploit known properties of both the source signals and the mixing system, but require the microphones to be in a constrained geometry. In this paper a solution is presented that avoids this constraint by extracting information from the magnitude of the mixing system instead of its phase. The new method combines that information with information from the source estimates to provide a reliable permutation alignment.

***Index Terms*—** Blind source separation, Acoustic signal processing, Array signal processing, Independent component analysis

## 1. INTRODUCTION

The separation of sound sources from convolutive mixtures is a challenging problem in the field of acoustic signal processing. The problem is often presented as a case of blind source separation (BSS), since it is commonly assumed that there is no prior knowledge about the sources or about the mixing system.

Several BSS algorithms have been proposed for the case of instantaneous linear mixtures, usually based on the decorrelation of the sources or on their mutual independence. In the later case the problem can be interpreted as an independent component analysis (ICA) problem [1]. The solution to this problem is not unique, since any arbitrary permutation or scaling of a solution is also a valid solution.

The BSS problem becomes much more complex when the mixture is convolutive. Some of the approaches used for the instantaneous case can be extended for the convolutive one, but the computational requirements become too high for real time applications. Alternatively, it is possible to formulate the problem in frequency domain, typically using discrete frequencies [2]. This transforms the convolutions into products, allowing the application of an instantaneous BSS algorithm separately for each frequency.

In the frequency domain BSS approach the permutation ambiguity holds for each frequency bin separately. This is a major problem, since it does not allow to reconstruct the sources from their frequency domain representation. There is no immediate way to know which of the estimated frequency components needs to be selected at each frequency to reconstruct a given source. Some additional properties of the sources or the mixing system need to be used.

This work is supported in part by the National Science Foundation Award IIS-0534221.

Several different approaches have been proposed to deal with the permutation problem. Most of them try to align the permutations after performing the separation, setting the new permutations in such a way that a known property of the sources or the unmixing system is satisfied by the solution. Some properties that have been proposed are the smoothness of the unmixing system [3], the directional information implicit in the unmixing system inverse [4] or the similarity of the time evolution of the sources at different frequencies [5].

The use of information related to the geometry of the mixing system is an effective way to arrange the permutations, as long as some important assumptions about the mixing system are satisfied. One common assumption is that the microphones form an array or, more generally, that distances among them are bounded [6]. Another common assumption is that the direct paths are dominant. When this second assumption is not exactly satisfied, knowledge about the time evolution of the sources can be used to enhance the solution [4].

In this paper an approach to the permutation problem is presented that exploits properties of both the sources and the mixing system. The approach works with any number of sources, and does not require any specific arrangement or any bound to the distance between microphones. This allows one to use it in distributed microphones applications, where previous approaches fail, like remote collaboration where each participant may carry his own microphone via a PDA or other portable device [7]. A simple and effective way to combine the mixing system information and the source information to provide a more reliable permutation alignment is also presented.

## 2. FREQUENCY DOMAIN BSS

In this paper an acoustic mixture is modeled as the linear combination of several statistically independent sources, each of them transformed through a different linear time-invariant acoustic response,

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_{r=0}^{R-1} \mathrm{h}_{jk}(r)\mathrm{s}_k(t-r) \qquad \text{for } j = 1 \dots M, \quad (1)$$

where $t$ is the discrete time index, and $N$ and $M$ are the number of sources and microphones. In the blind separation problem, the mixing filters $\mathrm{h}_{jk}$ and the sources $\mathrm{s}_k$ are unknown, and the goal is to find an estimate of each source $\mathrm{s}_k$ from the measured signals $\mathrm{x}_j$, knowing that the sources are statistically independent from each other. In this paper only the case $N = M$ is considered.

In the case of acoustic mixtures, the length of the mixing filters can be large. To reduce the associated computational cost the problem is moved to frequency domain. Figure 1 shows the different stages of this approach. First, the $L$-point STFT of the measured

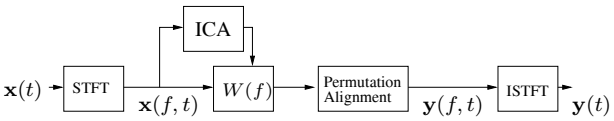


**Fig. 1**. Frequency domain BSS.

signals, $\mathrm{x}_j(t)$, is computed:

$$x_j(f,t) = \sum_{r=-L/2}^{L/2-1} \mathrm{x}_j(t+r)\mathrm{win}(r)e^{-i2\pi fr}, \tag{2}$$

where $f$ is the frequency index, and win is the analysis window. If this window is long enough, the mixing system becomes approximately instantaneous. In vector notation,

$$\mathbf{x}(f,t) = H(f)\mathbf{s}(f,t), \tag{3}$$

where $\mathbf{x}(f,t) = [x_1(f,t)\dots x_M(f,t)]^T$ is the measurements vector at one time-frequency point, $\mathbf{s}(f,t)$ is the corresponding source signal vector, and $H(f)$ is a square scalar (complex) mixing matrix.

Next, ICA is performed separately for each frequency to estimate a separation system $W(f)$ from the vectors $\mathbf{x}(f,t)$. The Complex FastICA algorithm [8] is used for this. The result is

$$\hat{\mathbf{y}}(f,t) = W(f)\mathbf{x}(f,t). \tag{4}$$

If ICA succeeds, the resulting vectors $\hat{\mathbf{y}}(f,t)$ will relate to $\mathbf{s}(f,t)$ through a permutation matrix $Q(f)$ and a diagonal matrix $D(f)$, due to the scaling and permutation ambiguities inherent in ICA:

$$\hat{\mathbf{y}}(f,t) \approx Q(f)D(f)\mathbf{s}(f,t), \tag{5}$$

The permutation matrices $Q(f)$ may be different for different frequencies. To achieve the separation of the sources, it is necessary to permute the vectors $\hat{\mathbf{y}}(f,t)$ so that they all relate to the sources through the same frequency independent permutation matrix.

$$\mathbf{y}(f,t) = P(f)\hat{\mathbf{y}}(f,t) = \hat{P}D(f)\mathbf{s}(f,t) \tag{6}$$

Permutation vectors can be used instead of permutation matrices, so that the reordering in Equation (6) can be written as follows:

$$\mathbf{y}(f,t) = [\hat{y}_{\mathrm{p}_1}(f,t)\dots\hat{y}_{\mathrm{p}_N}(f,t)]^T, \tag{7}$$

where $\mathbf{p}(f) = [\mathrm{p}_1\dots\mathrm{p}_N]$ is a sequence of permutation indexes, i.e. a certain permutation of the sequence $[1\dots N]$. The set of all possible permutation vectors is denoted as $\Pi$.

There are different ways to find the right permutation vectors $\mathbf{p}(f)$. Often this is done using additional information about the mixing system or about the source signals. In [4], both kinds of information are combined to provide a more reliable alignment. In the following sections a related approach is proposed that can be used in situations where some assumptions required in [4] are not satisfied.

## 3. CLUSTERING OF MAGNITUDE RATIOS

When ICA achieves separation at a given frequency bin, it becomes possible to obtain information about the original mixing system from the resulting unmixing matrix. In particular its inverse, $B(f) = W(f)^{-1}$ is related to $H(f)$. This can be shown by noting that

$$\mathbf{x}(f,t) = B(f)\hat{\mathbf{y}}(f,t) \approx B(f)Q(f)D(f)\mathbf{s}(f,t), \tag{8}$$

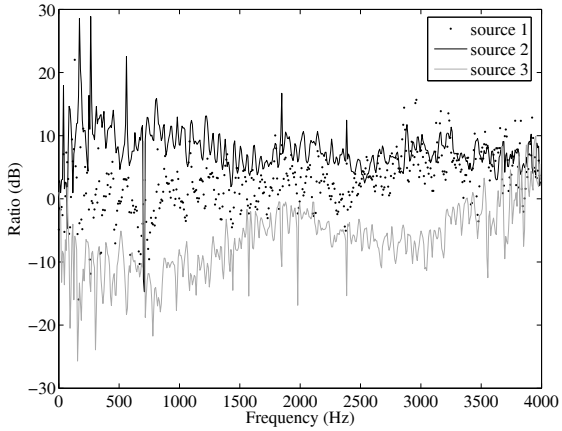


**Fig. 2**. Magnitude ratio of the last two rows of $\mathbf{b}_l(f)$ for $l = 1, 2, 3$, after the separation of a 38 seconds long mixture of 3 sources.

and combining equations (3) and (8), it turns out that

$$B(f) \approx H(f)D(f)^{-1}Q(f)^{-1} \tag{9}$$

This means that each of the columns of $B(f)$ is an scaled version of some column of $H(f)$. The following notation is used for these column vectors:

$$B(f) = (\mathbf{b}_1(f)\dots\mathbf{b}_l(f)\dots\mathbf{b}_N(f)) \tag{10}$$

For this relationship between $B(f)$ and $H(f)$ to be useful to solve the permutation inconsistencies, there needs to be some property of the columns of $H(f)$ that is different for each column (source), constant across frequency, and invariant to the arbitrary scaling. It is possible to find such properties under the following assumption:

**Assumption:** The magnitude of the mixing response consists of a frequency independent matrix multiplied by a scalar function $q(f)$:

$$\|h_{jk}(f)\| = m_{jk}q(f) \tag{11}$$

This assumption is similar, but less restrictive, to the free-field assumption used in beamforming theory, and as such is satisfied exactly in anechoic conditions. In real situations it may be approximately satisfied if the reverberation is not too big, or if the sources are very close to the microphones.

Under the above assumption, the ratios between the magnitudes of any two components of each vector $\mathbf{b}_l(f)$ become frequency independent. For example, the vector

$$\mathbf{r}_l(f) = [b_{2l}(f)/b_{1l}(f)\dots b_{Nl}(f)/b_{N-1l}(f)]^T \tag{12}$$

can be used to align the permutations, since it will be the same for a given source at different frequencies. In the previous equation, $b_{kl}(f)$ are the elements of the vector $\mathbf{b}_l(f) = [b_{1l}(f)\dots b_{Nl}(f)]^T$.

In a real situation where the above assumption is not exactly satisfied, it may be possible to use the vectors $\mathbf{r}_l(f)$ to choose the correct permutations if the frequency dependent variations are smaller than the frequency independent ones. This is broadly the case for the experimental setup described in Section 6, as Figure 2 illustrates.

The procedure to decide the correct permutation vector $\mathbf{p}_a(f)$ for each frequency is the following:

1. Perform $k$-means clustering, with $k = N$ on the set of all the vectors $\mathbf{r}_l(f)$ for all frequencies, and all $l$. This yields $N$ centroid vectors $\mathbf{c}_k^r$ one for each source, and the corresponding cluster standard deviations, $\sigma_k^r$.
2. For each frequency, find the permutation that minimizes:

$$\mathbf{p}_a(f) = \underset{\mathbf{p}\in\Pi}{\operatorname{argmin}} \left( \sum_{k=1}^{N} \frac{\|\mathbf{c}_k^r - \mathbf{r}_{\mathrm{p_k}}(f)\|^2}{(\sigma_k^r)^2} \right) \qquad (13)$$

   This step is necessary because clustering does not guarantee that the column vectors at a given frequency bin will be assigned to different sources.

A probabilistic justification can be given for Equation (13). If the vectors $\mathbf{r}_l(f)$ were drawn from a multidimensional gaussian distribution with mean $\mathbf{c}_k^r$ and a constant diagonal correlation matrix, then $\mathbf{p}_a(f)$ would be the maximum likelihood permutation vector. In practice the gaussian assumption may not be satisfied, so experimental evaluation of the previous algorithm is necessary.

## 4. CLUSTERING OF TIME ENVELOPES

Many sound sources of interest, like speech, are non-stationary. In a time-frequency representation of speech, the time variations are not identical for different frequencies, but they are not independent either. They can be expected to present certain similarity, because silence segments coincide across frequency, and active segments tend to involve many frequencies more or less simultaneously.

These similarities are known to be most clear when looking at adjacent frequency bins, but alignment based on local similarity does not provide adequate results, due to the propagation of errors [9]. Correlation between arbitrary frequencies can be weaker. A way to improve it is to work with magnitude envelopes, instead of using the complex STFT sequences directly.

A clustering algorithm like the one described in the previous section is able to extract global similarities. In this case the vectors would be sequences modeling the time envelopes of the spectrogram, one vector for each frequency bin and for each source. This method is computationally expensive, but its implementation is very simple. More efficient and sophisticated approaches, such as the dyadic sorting proposed in [10], could be used with potentially similar results.

The procedure to decide the correct permutation vector $\mathbf{p}_a(f)$ for each frequency using time envelopes is the following:

1. Compute the magnitude envelopes of the output spectrogram $\hat{\mathbf{y}}(f,t)$, by filtering each $\|\hat{y}_l(f,t)\|$ with a smoothing window, and downsampling the result to build the vectors $\mathbf{e}_l(f)$.

$$\{\mathbf{e}_l(f)\}_t = \sum_r \|\hat{y}_l(f,r)\| \mathrm{ewin}(St - r) \qquad (14)$$

   where $S$ is the downsampling factor and ewin is the smoothing window.
2. Perform $k$-means clustering, with $k = N$ on the set of all the vectors $\mathbf{e}_l(f)$ for all frequencies, and all $l$. This yields $N$ centroid vectors $\mathbf{c}_k^e$ one for each source, and the corresponding cluster standard deviations, $\sigma_k^e$.
3. For each frequency, find the permutation that minimizes:

$$\mathbf{p}_a(f) = \underset{\mathbf{p}\in\Pi}{\operatorname{argmin}} \left( \sum_{k=1}^{N} \frac{\|\mathbf{c}_k^e - \mathbf{e}_{\mathrm{p_k}}(f)\|^2}{(\sigma_k^e)^2} \right) \qquad (15)$$

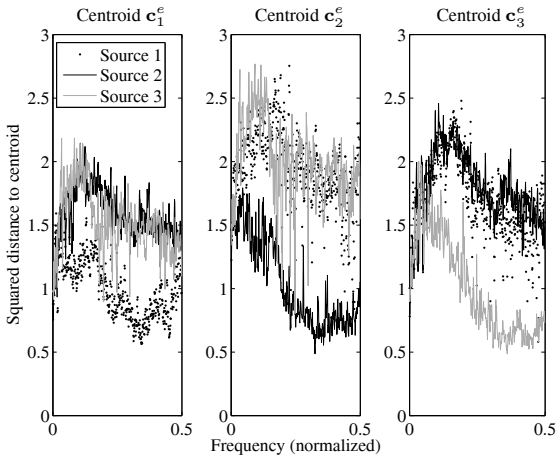


**Fig. 3**. Squared distances from each of the envelope vectors to each of the centroids $\mathbf{c}_k^e$, after separation of a 38 seconds mixture. Vectors are grouped according to the source they have been assigned to.

As the experiments described in Section 6 show, this mechanism indeed allows to identify the frequency components that correspond to each source. This can be seen in Figure 3, which shows the resulting distances to the cluster centroids for one of the experiments.

## 5. COMBINING SYSTEM MAGNITUDE AND SOURCE ENVELOPES

The two approaches described in the previous sections, using the mixing system magnitude and the source time envelopes, are based on unrelated concepts, and it can be expected that better results will be obtained by combining both.

A simple way to achieve this combination is to perform the clustering operations described before, and then to choose each permutation vector using the following rule:

$$\mathbf{p}_a(f) = \underset{\mathbf{p}\in\Pi}{\operatorname{argmin}} \left( \sum_{k=1}^{N} \frac{\|\mathbf{c}_k^r - \mathbf{r}_{\mathrm{p_k}}(f)\|^2}{(\sigma_k^r)^2} + \sum_{k=1}^{N} \frac{\|\mathbf{c}_k^e - \mathbf{e}_{\mathrm{p_k}}(f)\|^2}{(\sigma_k^e)^2} \right) \qquad (16)$$

This is again the maximum likelihood choice if the clusters follow certain gaussian model. But as Figure 3 suggests, this is sometimes clearly not the case. This may be a problem, specially if the cluster distributions of $\mathbf{r}_l(f)$ and $\mathbf{e}_l(f)$ are very different. That can make Equation (16) biased towards one of the two contributions.

## 6. EXPERIMENTAL RESULTS

Several experiments have been carried out to analyze the performance of the proposed approach. The mixture segments were prepared from a set of clean speech recordings mixed through one set of room response recordings, using three sources and three microphones. The mixture segment lengths used were 38 and 5 seconds.

For the mixing system, the set of room responses were recorded in a controlled room configured to emulate a remote collaboration site, with a reverberation time of 300ms. Three loudspeakers were located around a conference table, and three omnidirectional microphones were placed on the table, attached to the side of three PDAs in front of each of the loudspeakers. The nine acoustic responses were estimated using the maximum length sequence method.

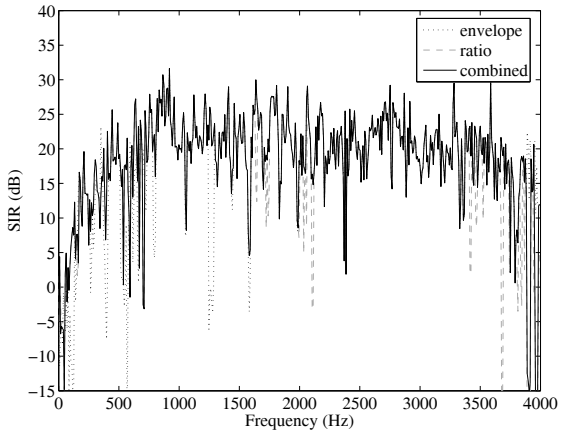


**Fig. 4**. SIR frequency for 5 seconds signals, using envelopes and magnitude ratios clustering, separately and jointly.

The speech signals were constructed by concatenating clean sentence recordings taken from the TIMIT database to form three segments of 38 seconds, each from a different speaker in the database. The speech segments and the room responses were downsampled to 8KHz. No noise was added to the mixture. The separation experiments were performed on the 38 seconds segments, and were repeated using only the first 5 seconds.

In the separation algorithm, a hanning window of length $L = 1024$ samples with a 75% overlap was used for the STFT analysis. The time envelopes were computed using a triangular window of length 6 and a downsampling factor of $S = 3$.

Separation performance was evaluated using the signal to interference ratio (SIR) at the outputs, computed both globally and across frequency. For comparison, performance was evaluated before separation, and also after separation followed by an optimum alignment performed with knowledge of the mixing system.

The following table shows the aggregated SIR results in dB.

| Length | Input | Mag. ratio | Envelope | Combined | Optimum |
|---|---|---|---|---|---|
| 5 sec | 5.75 | 15.97 | 13.73 | 16.15 | 17.85 |
| 38 sec | 5.79 | 19.38 | 20.92 | 19.53 | 20.94 |

In the 5 seconds experiment (first row), clustering of magnitude ratios gives a better alignment than envelope clustering, but both do provide some improvement. When they are combined, the results are further improved, although the optimum alignment is not reached. The results across frequency are shown in Figure 4. Most of the pronounced dips in the SIR curve indicate a permutation error.

In the experiment with long signals (second row), envelope clustering provides an almost optimum alignment, while clustering of magnitude ratios is slightly less effective. The combination rule seems to be favoring the magnitude clustering results in this case, even at some frequency bins where they are pointing to a wrong alignment, leading to worse results than with envelope clustering.

The short data situation is more interesting for practical applications with a slowly varying mixing system, so these results show the usefulness of the combined algorithm. Future work should aim at improving the permutation decision rule.

## 7. CONCLUSION

A novel approach to the permutation alignment problem in blind source separation has been presented in this paper. The approach combines prior knowledge about general properties of both the mixing system and the time evolution of the sources.

A simple assumption about the magnitude of the mixing system has been shown to allow certain degree of alignment without the need to add any constraint on the location of the microphones. Regarding the time evolution of the sources, it has been found that for long enough measurements time envelopes can be clustered to obtain a very reliable alignment.

Finally, a way to combine both approaches that provides an improvement with respect to to either of them for short measurements has been presented. This combination is not desirable or necessary for longer signals, where time envelopes seem to provide a good enough permutation alignment.

## 8. REFERENCES

[1] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, *Independent Component Analysis*, John Wiley, 2001.

[2] Paris Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21 – 34, Nov. 1998.

[3] Lucas C. Parra and Clay Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[4] Hiroshi Sawada, Ryo Mukai, Shoko Araki, and Shoji Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[5] Jörn Anemüller and Birger Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," Helsinki, Finland, June 2000, pp. 215–220.

[6] Hiroshi Sawada, Shoko Araki, Ryo Mukai, and Shoji Makino, "Blind Extraction of a Dominant Source Signal from Mixtures of Many Sources," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, PA, USA, Mar. 2005, pp. 61–64.

[7] T. S. Wada, E. Robledo-Arununcio, G. Yue, and B. H. Juang, "Immersive acoustic signal processing for intelligent collaboration," in *Proceedings from the 9th Western Pacific Acoustics Conference*, Seoul, South Korea, June 2006, p. 653.

[8] E. Bingham and Aapo Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *International Journal of Neural Systems*, vol. 10, no. 1, pp. 1–8, 2000.

[9] E. Robledo-Arnuncio and B. H. Juang, "Using Inter-Frequency Decorrelation to Reduce the Permutation Inconsistency Problem in Blind Source Separation," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech)*, Lisbon, Portugal, Sept. 2005.

[10] Kamran Rahbar and James P. Reilly, "A Frequency Domain Method for Blind Source Separation of Convolutive Audio Mixtures," *IEEE transactions on speech and audio processing*, vol. 13, no. 5, pp. 832–844, Sept. 2005.