Blind Source Separation: Recover Signals from Mixtures

Burim Ramadani

b.ramadani@jacobs-university.de

Electrical Engineering
Jacobs University Bremen
Campus Ring 1
28759 Bremen
Germany

Type: Guided Research Thesis
Date: May 6, 2007
Supervisor: Dr. Mathias Bode
Executive Summary

The focus of this paper is on the Blind Source Separation's underdetermined case where three speech signals are mixed in an environment, and the measurements are recorded by two microphones (or sensors). From the research up to date in this area, the assumption that the signals are sufficiently sparse is made. The designing of the binary masks has been made where only one signal was assumed to be active. None of them took into consideration that the musical noise (distortion) was one of the key problems in the separation. In order to solve this problem I propose the combination of the sparseness with the usage of the estimation of mixing matrices. As a first step the geometric approach of finding the times when one source is active is used, and then that gained information is used in the estimation of the mixing matrix, which would lead into the separation of the mixtures into almost noiseless separate signals. The experimental results gained provide evidence that using the mixing matrix estimation and masks for time-frequency leads to much better quality of the separated signal rather than without using the estimating matrix in real world conditions (reverberations present), where in our case the reverberant time was 200ms.
1 Introduction

Blind Source Separation is one of the most researched topics in the field of speech processing, seismic signal processing, antenna arrays for wireless communications and biomedical signal processing. There are two problems that are mainly considered in the aspect of research in BSS, and they are convolutive (where time delays are introduced) and instantaneous BSS (no time delays are introduced). Many algorithms have been proposed for the former one, where mixtures are linear combinations of independent sources.

As mentioned early, there are a variety of applications where we can range. Our attention lies in the application towards hearing aids. Since Cocktail Party Problem is highly connected to what we want, a simple scenario of this problem is given as following. Cocktail Party Problem (Figure 1.) can be described as a situation when we have 2 (or more) speakers talking simultaneously. Each speaker is a source and we record a mixture of the sources using 2 microphones (sensors). Each sensor contains a mixture of voices of 2 (or more) speakers, refers to nxm MIMO system. The task is to recover the individual sources (voices) having only the record of their mixtures (the mixing process of sources is unknown).
Since a lot of words like blind, underdetermined and convolutive will be discussed in the thesis, therefore in this introduction I will focus in each in turn and explain the way they relate to each other.

First, the adjective blind stresses the fact that the source signals are not observed and that no information is available about how the sources are mixed (e.g., [3]). Blind source separation (BSS) actually refers to the problem of recovering signals from several observed linear mixtures. The weakness of the prior information is precisely the strength of the BSS model. In this thesis, we will concentrate on the audio BSS issue, which is currently a major research field. Probably the best known problem in auditory scene analysis is that of the cocktail party.

Recently, many methods have been proposed to solve the BSS problem of audio signals in real environments. A common approach involves using independent component analysis (ICA). However, most of these methods consider a determined problem, where we have as many sources as sensors.

In this thesis in an attempt to deal with a more contemporary issue, we have decided to concentrate on an underdetermined case where there are more sources than sensors, which comes from the fact that in real life, most of the time, there are ambient and uncountable noise sources. Furthermore, manufacturers do not want many sensors for their small and inexpensive products. Therefore, for real applications, we have to consider the underdetermined case. Moreover, to be more realistic, we should consider the mixing process of the speech signal to be convolutive because most speech signals are recorded with their reverberation. Below the focus is on the two remaining keywords and look at the kinds of solutions that have already been proposed for dealing with the issues that the words refer to.

The true obstacle raised by underdetermination relates to the fact that a simple inversion of the mixing matrix will not lead us to a solution in as much as the mixing system is not square and therefore not invertible. Furthermore, even if we knew the mixing matrix exactly, we would not be able to recover the original signals because part of the information is lost during the mixing process. Thus far, solving the BSS
problem in an underdetermined case has mainly consisted in assuming that the speech signals were sufficiently sparse [1], [2], which legitimizes the extraction of each source.

To our understanding there are two approaches relying on sparseness that will solve the underdetermined BSS. The first involves the clustering of time-frequency points with binary masks [1]. It emerges that if the signals are sufficiently sparse, namely if most of the samples of a signal are almost zero, we can assume that the sources rarely overlap. [1] uses this assumption and extracts each signal using a time-frequency binary mask. The second approach is based on ML (Maximum Likelihood) estimation [4].

Although, dealing with the underdetermined BSS is already very tough, it is still possible to present myself with a greater challenge by considering the underdetermined convolutive BSS issue, a solution to which has already been sought [1]. However, this approach leads to the discontinuous zero-padding of separated signals because all the unknown samples, (e.g. where several sources are active at the same time) are set to zero by a binary mask. Consequently such separated signals are greatly distorted, and therefore a loud musical noise is heard. In contrast to the ML approach, which has excited great interest in recent years, only a few trials have been undertaken related to the convolutive case [11], [12]. Of course, convolutive mixtures can be seen frequency bin by frequency bin as instantaneous mixtures but, as such a conversion results in complex-valued signals, the $l_1$-norm minimization (maximum likelihood estimation under Laplacian noise model) is not applicable straightforwardly and would need to be expanded or alternated. Consequently, due to a high level of complexity, the ML approach is rather avoided in the convolutive case.

The objective of this thesis is to provide a satisfactory solution to the underdetermined convolutive BSS by ensuring that any musical noise is minimized. In this thesis, the focus is on the underdetermined BSS of three speech signals mixed in a real environment from measurements provided by two sensors.
It should be remembered that one of the ways to overcome the BSS issue in a determined problem is to estimate and then invert the mixing matrix modeling of our system [4]. Recently, [5] proposed an approach for estimating the separating matrix in the instantaneous determined BSS problem. However, here, where sources outnumber sensors, the mixing matrix and separating matrix are no longer square and such solutions cannot be used. Nevertheless, this information lead to the idea of combining the sparseness properties of speech signals with an estimation of the mixing matrix.

My suggestion for eliminating the distortion problem in convolutive underdetermined BSS is to combine sparseness with a mixing matrix estimation [6]. First, using the sparseness, we detect the time points when only one source is active. This information is then used to estimate the mixing matrix. Subsequently, by using sparseness investigations, we were able to work under conditions where one of the sources could be omitted by a time-frequency mask. The separation of the residual signals using the inverse of the estimated mixing matrix then follows. By this one source removal, the zero-padding of the separated signal is expected to be less corrupting than with the previous binary mask approach. Results show that we indeed obtain more information about the signals to be separated and significantly lowered the musical noise comparatively to the classical approaches. As a consequence, we can reduce the zero-padding effect from which the musical noise originates. A complementary approach, also relying on one source removal, has been proposed by the same authors [7], where they use the ICA algorithm to the residual signals. That is where by sparseness they find the time points when only one source is active, after which the removal of this single source is being done and then they apply ICA to the remaining sources. Apart from Araki’s approach and so as to overcome the previously mentioned lack of quality, here, I attempt to utilize an estimated mixing matrix to separate the residual.

The organization of this thesis is as follows: In Section 2 the description of problem statements is done, whereas in Section 3 the essential investigation with regard to sparseness is described. Section 4 is a step by step presentation of the proposed method, whose results are provided in Section 5. And at the end the conclusions are provided in Section 6.
2 Statement and Motivation of Research

In this thesis, I consider speech mixtures observed in a real room. In this case, as speeches are mixed with their reverberation, the observed vectors $x_j (j = 1, ... , M)$ can be modeled as convolutive mixtures of the source signals $s_i (i = 1, ... , N)$ as follows:

$$x_j(n) = \sum_{i=1}^{N} \sum_{k} h_{ji}(k)s_i(n - k + 1)$$

(1)

where $h_{ji}$ is the impulse response from a source $i$ to a sensor $j$. As it is the first trial, we decided to deal with the simplest case, i.e. where $N = 3$ sources and $M = 2$ sensors. Moreover, we assume that the source signals are sparse: namely signals have large values at rare sampling points.

We use the Short Time Fourier Transform (STFT) to convert our problem into a linear instantaneous mixture problem as well as to improve the sparseness of the speech signals [1]. In the time-frequency domain, our system becomes:

$$X(f, m) = H(f)S(f, m)$$

(2)

where $f$ is the frequency, $m$ the frame index, $H(f)$ the $2 \times 3$ mixing matrix whose $(j, i)$ component is a transfer function from a source $i$ to a sensor $j$. $X(f, m) = [X_1(f, m)X_2(f, m)]^T$ and $S(f, m) = [S_1(f, m)S_2(f, m)S_3(f, m)]^T$, namely the Fourier transformed observed signals and source signals, respectively. Our aim is to estimate three speech signals from measurements provided by two sensors.
3 Sparseness

Sparseness becomes greater as the number of zero samples contained in a source increases, which means that the sources overlap at infrequent intervals. Interesting investigations of the sparseness property of speech signals have been proposed in [1] for anechoic speech signals. Here, the detailed investigations for three speech signals are given.

In the real world it can be seen that there are many time points at which no sources (speech signals) are active and few where three sources are active. We can infer from the old investigations by others that the signals are sparse and that three signals rarely overlap. So that the speech signals are sparse is an assumption that has been taken in this thesis.

Measurement of the Overlapping

For determining the best representation, we investigated the sparseness more closely and checked the degree of signal overlap by utilizing a criterion called Approximate W-Disjoint Orthogonality (WDO) defined by Yilmaz and Rickard [1]. We use a mask:

\[ \phi_{(j,x)}(f,m) = \begin{cases} 1, & \text{if } 20 \log \left( \frac{|S_j(f,m)|}{|Y_j(f,m)|} \right) > x \\ 0, & \text{otherwise} \end{cases} \tag{3} \]

where \( Y_j(f,m) \) is the DFT of \( y_j(n) = \sum_{i=1,i\neq j}^{N} s_i(n) \) \( (4) \) i.e. \( y_j(n) \) is the summation of the sources interfering with the source \( j \). The approximate WDO is defined as:

\[ r_j(x) = 100 \frac{\left\| \phi_{(j,x)}(f,m) S_j(f,m) \right\|^2}{\left\| S_j(f,m) \right\|^2} \tag{5} \]
where \( \|f(x, y)\|^2 = \sum_y \sum_x |f(x, y)|^2 \), where \( f \) is a function of \( x \) and \( y \).

This measures the percentage \( r_j \) of source \( j \) energy for time-frequency points where this source dominates the other signals by \( r_j \) % at \( x \) dB. From this criterion it emerges that, if we can predict the time-frequency points at which a source dominates the others by \( r_j \) % at \( x \) dB, we should be able to recover \( r_j \) % of the energy of the original sources. If \( r_j \) is sufficiently large, we can separate signals with little distortion and vice-versa.
4 Method

As seen in the previous papers (e.g. [2]), a drawback of the usual approaches was the noise. In order to overcome this issue, the proposition of a methods made of three steps is proposed. As a first step starting with the usage of the sparseness of speech signals, we adopt a geometrical approach to determine the time-frequency masks that extract the time points \( m \) when only one source is active, after that as a second step then the estimation of the mixing matrix is being done and as a last step the reconstruction of the signals when two sources are active.

- The geometric approach

As a first step the detection of the frame indices \( m \) when only one of the three sources is active for each frequency bin \( f \). Scatter plot of the measurement, as shown in Fig. 2, comprise three main lines in case the sources are sparse enough. According to a paper from Pereda [8], these lines represent the directions defined by the column vectors of the mixing matrix. In other words, they can be seen as a representation of each source existing alone. In between two given directions, we find the time-frequency points modeling our system when two sources (those linked to the above directions) are active simultaneously.
With the setting of narrow areas each containing only one line, such as areas 1, 2 and 3 in Fig. 2, it is able to determine when only one source is active and at the same time we can reconstruct the signals for these time-frequency points. From which we can say that three time-frequency binary masks can be designed, they are assigned to the value 1 in areas 1, 2 or 3 and to the value 0 otherwise. This method was described in the paper by Yilmaz [1]. However, as expected when using such a rough approach, the quality of the separated signals is unsatisfactory. Since the rate of recoverable energy is very low, we cannot avoid an important zero padding, which makes the signals insufficiently continuous. As a result, we hear considerable distortion i.e., loud musical noise. To recover this lack of quality, we attempt to complete our separation relying on the knowledge of the mixing matrix.

- **Estimating the Mixing Matrix**

As seen at Deville [9], he recovers the mixing matrix by estimating a certain cross correlation parameter ratio over time frequency zones where only one source exists. After some proofs the ratio was equal to \( \frac{H_{2i}}{H_{1i}} \) \((i = 1, 2, 3)\). In contrast to the work of Deville, here we are dealing with the underdetermined convolutive issue, however his approach
enlightened to the idea of modeling our system in the time frequency domain by:

\[
\begin{pmatrix}
X_1(f, m) \\
X_2(f, m)
\end{pmatrix} =
\begin{pmatrix}
1 & 1 & 1 \\
H_{21} & H_{22} & H_{23}
\end{pmatrix}
\begin{pmatrix}
H_{11}(f)S_1(f, m) \\
H_{12}(f)S_2(f, m) \\
H_{13}(f)S_3(f, m)
\end{pmatrix}
\] (6)

So, by using the time points that are estimated in the first step when only one of \( S_i \) \((i = 1, 2, 3)\) is active, we have:

\[
\begin{align*}
X_1(f, m) &= H_{1i}(f)S_i(f, m) \\
X_2(f, m) &= H_{2i}(f)S_i(f, m)
\end{align*}
\] (7)

whose ratio \( X_2(f, m)/X_1(f, m) \) provides one of the components of the mixing matrix \( H_{2i}(f) \). For a stable estimation of the mixing matrix coefficients, we estimated the expectation of the ratio \( X_2/X_1 \):

\[
\frac{H_{2i}(f)}{H_{1i}(f)} = \mathbb{E}[X_2(f, m)/X_1(f, m)]
\] (8)

where \( \mathbb{E} \) is the expectation when \( X_1(f, m) \) and \( X_2(f, m) \) are in area \( i \) at time \( m \).

- **The reconstruction of the points in time frequency when two out of three sources are active**

At this point of time, knowing the mixing matrix doesn’t help us to make the separation of the signals when all the three sources are active. And the reason is because the mixing matrix isn’t square, i.e. doesn’t have an inverse. In the paper of Deville [9], it can be noticed that he applied his method only to a square mixing matrix. Nevertheless, it is still possible to rebuild the time-frequency points when two sources are active, providing that for each frequency bin, we know the frame indices for which this case occurs. Once more this information is provided by the geometrical approach employed in the first step. But this time, instead of
setting the limits close enough to the observed directions, much wider areas are considered so as to enclose the points located between two given directions. That is to say that utilization of wider time-frequency masks has been done. Indeed let us suppose that, for an estimated \((f, m)\) detected during the first step, \(S_3(f, m)\) is zero, in this area our system becomes:

\[
\begin{bmatrix}
X_1(f, m) \\
X_2(f, m)
\end{bmatrix} = \begin{bmatrix}
\frac{1}{H_{21}} & \frac{1}{H_{22}} \\
H_{11} & H_{12}
\end{bmatrix} \begin{bmatrix}
H_{11}(f)S_1(f, m) \\
H_{12}(f)S_2(f, m)
\end{bmatrix} \quad (9)
\]

And now since the mixing matrix is square and can be inverted, leading to \(H_{11}(f)S_1(f, m)\) and \(H_{12}(f)S_2(f, m)\):

\[
\begin{bmatrix}
H_{11}(f)S_1(f, m) \\
H_{12}(f)S_2(f, m)
\end{bmatrix} = \begin{bmatrix}
\frac{1}{H_{21}} & \frac{1}{H_{22}} \\
H_{11} & H_{12}
\end{bmatrix}^{-1} \begin{bmatrix}
X_1(f, m) \\
X_2(f, m)
\end{bmatrix} \quad (10)
\]

From which we can see that the separated signals can be obtained in the following manner:

\[
\begin{bmatrix}
C_1(f, m) \\
C_2(f, m)
\end{bmatrix} = \begin{bmatrix}
\frac{1}{H_{21}} & \frac{1}{H_{22}} \\
H_{11} & H_{12}
\end{bmatrix}^{-1} \begin{bmatrix}
X_1(f, m) \\
X_2(f, m)
\end{bmatrix} \quad (11)
\]

To elaborate more in this area, because the signals \(H_{11}(f)S_1(f, m)\) and \(H_{12}(f)S_2(f, m)\) are not too greatly zero-padded due to the wide mask, we expect that the distortion of the estimated \(H_{11}(f)S_1(f, m)\) and \(H_{12}(f)S_2(f, m)\) will not be very large. We proceed in the same way when \(S_1(f, m)\) is zero and obtain the estimates of \(H_{12}(f)S_2(f, m)\) and \(H_{13}(f)S_3(f, m)\). It should be noted that in Section 3, we have already confirmed that we do not often have three sources active simultaneously. The method proposed could also be extended to a general case where there are \(N\) sources and \(M\) sensors. In such a case, we would rewrite (6) as:
\[
\begin{pmatrix}
X_1(f, m) \\
\vdots \\
X_M(f, m)
\end{pmatrix}
= \mathbf{H}'(f) \begin{pmatrix}
H_{k1}(f)S_1(f, m) \\
\vdots \\
H_{kN}(f)S_N(f, m)
\end{pmatrix}
\] (12)

where

\[
\mathbf{H}'(f) = \begin{cases}
1 & j = k, \\
\frac{H_{ji}}{H_{ki}} & j \neq k.
\end{cases}
\] (13)

And \(\frac{H_{ji}}{H_{ki}}\) could be estimated by \(E[X_j(f)/X_k(f)]\) for an area \(i\). Then we define a wide area so that we can redefine the square part of \(\mathbf{H}'(f)\) as in (9), then separate the signals in wide area by inverting the square part of \(\mathbf{H}'(f)\) as it done in (10).
5 Experimental Results

5.1 Conditions

The speech recordings were made in a small room with reverberation approx. RT=200 ms [10] using a two-element array of directional microphones 10 cm apart. The speech signals, sampled at 8 kHz, came from three directions: 120° female, 90°-male and 50° - male and the distance between the sources and the sensors was \( L = 80 \text{ cm} \) (see, Fig 3).

![Figure 3 - Model of the room situation when recordings were made](image)

5.2 Check for the Stability of Estimated Matrix Coefficients and the Results

To evaluate the efficiency of our method, we need to know about the stability of the mixing matrix we estimated in the 2nd stage. In Fig. 4, we plot the amplitude and phase of the three coefficients \( \frac{H_{2i}(f)}{H_{1i}(f)} \) \((i = 1, 2, 3)\) in (8). It is seen that our estimation generally offers a great stability in the whole, except for the low frequencies, where the time delay between the two microphones, which are positioned very close to each other, is harder to calculate with accuracy.
However we can observe the constant amplitude of the coefficients.

Figure 4 - Matrix coefficients represented, (amplitude), for $H_{23}(f)/H_{13}(f)$-female, $H_{22}(f)/H_{12}(f)$-male, $H_{21}(f)/H_{11}(f)$-male

Having achieved this, the coefficients are all ready. The only thing left is to apply the formulas described under (6), (7), (8), (9) and finally (10), and to get the estimated sources when two signals are active and the third one is not. This was tried in non-laboratory conditions and the time required to reach the results was quite long. That was the reason for me to use small amount of samples in order to get the above shown graphs.
6 Conclusions

We believe we achieved to propose a separation method for use when there are more speech signals than sensors by combining the estimation of the mixing matrix and the sparseness approach. Even though we realize that the approach of STFT is not 100% correct for the mixing matrix, because of the finite duration of the windows. In case the microphones were far away from each other, than the delay would have been much longer, and the information on the two windows would have been misleading since they would have contained different information of the source signals. The experimental results are encouraging, and suggest that the combination of mixing matrix estimation and a time-frequency mask is an approach that deserves serious investigation.
References:


