UNIVERSITY OF CALIFORNIA
RIVERSIDE

Blind Source Separation of Speech Signals: Exploiting Second Order Statistics

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Electrical Engineering

by

Vishaal Madanagopal

March 2018

Thesis Committee:

Dr. Yingbo Hua, Chairperson
Dr. Vagelis Papalexakis
Dr. Salman Asif

The Thesis of Vishaal Madanagopal is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my advisor, Dr.Yingbo Hua, without whose guidance, support and encouragement, I would not have been able to proceed any further. The numerous discussion sessions and feedback from Dr.Hua has helped me get a good idea on how to proceed further with my thesis. I would also like to thank the lab members , my friends, Qiping Zhu, Reza Sohrabi, Ishmam Zabir for their support and encouragement. Our discussions in the lab has always been more fruitful than any other sources that I may have found. I would like to thank our visiting scholar Piexi Liu for his support and encouragement. I would also like to acknowledge the support given by the department and the department's grad advisors' previously Dr.Anastasios Mourikis and currently Dr.Amit Roy Chowdhury for believing in me and supporting me through all my pits and falls. I would like to thank my parents for being my support system. Finally, I would like to thank all my family,friends and colleagues for motivating me to pursue my ideas and dreams.

To my parents and family members for all the support.

ABSTRACT OF THE THESIS

Blind Source Separation of Speech Signals: Exploiting Second Order Statistics

by

Vishaal Madanagopal

Master of Science, Graduate Program in Electrical Engineering
University of California, Riverside, March 2018
Dr. Yingbo Hua, Chairperson

Blind source separation is a popular technique which is used in the fields of signal processing, audio, video and image processing. BSS is used to separate the mixed signals with only knowing the mixed signals and knowing very little about original signal characteristics. The separated signals should be very good approximations of the source signals. In particular, the blind source separation algorithm tries to estimate the Mixing Matrix. In my thesis, I have studied the blind source separation of signals based on its second order statistics. The problem of blind source separation is studied considering the following cases: when the signal is modelled as non-stationary, cyclo-stationary and quasi-stationary. A closed form solution to the blind source separation of speech signals considering speech to be a quasi-stationary source is studied and implemented.

Contents

List of Figures

Chapter 1

Introduction

Blind source separation is a unique problem where we know very little about the source signals or the mixing matrix, but rather we only have the output signal which is available for further processing and separation. The cocktail party problem is conceptually similar to what the blind source separation attempts to do. The cocktail party effect is the phenomenon of the brain's ability to focus one's auditory attention on a particular stimulus while filtering out a range of other stimuli. It is typical to the brain's function in

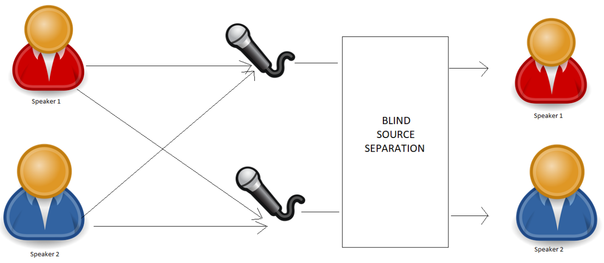

Figure 1.1: Blind source separation problem with two sources

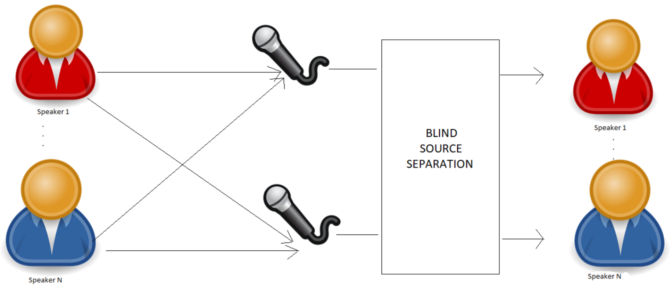

Figure 1.2: Blind source separation problem with N sources

eliminating the external sounds that contribute to unwanted stimuli(which are the sounds that are not of interest). In this way, the blind source separation also identifies only that source signal whose characteristics are more dominant at the time period of interest. The problem can further be refined to a set of sources and microphones separated in space. Each of these microphone receive the signal from all the sources, but there will be only one source whose statistic will dominate at the sensor in the time period of interest. Fig 1.1. illustrates the multi-microphone array set up for two sources and two microphones. Speaker 1 and Speaker 2 speak at the same time instant and their respective signals are received at each microphone. However, only one of the source characteristic is dominant at either microphone. This is the simplest possible set up of multi-microphone array system. However, the system of interest to us is described in Fig.1.2. This system has a total of M sources and N sensors. Blind source separation is used to identify the dominant source at the microphone during the time period of interest. There are a variety of approaches which have been proposed over the years for the blind source separation problem. The solution reviewed and implemented as a part of my thesis will have a set up similar to the one shown in Fig.(1.2). It would have M sources and N sensors. It would focus mainly on the assumption that the speech signals are quasi-stationary and will exploit the values of the second order statistics of speech signals. An estimate of the mixing matrix is obtained which follows from the closed form solution proposed in [9].

Chapter 2 would elaborate the problem statement in detail, listing a few existing modelling techniques for the speech signal and also giving a much more concrete parallel between the blind source separation problem and the cocktail party effect. It would also briefly describe speech acquisition models.

Chapter 3 would focus on the review of previous works that have been done on blind source separation of speech. This involves a brief description of the previous works in the area of blind source separation, considering speech to be a non-stationary source [20] (or) a cyclo-stationary source [2]. However, the solution reviewed as a part of my thesis work will illustrate the blind source separation problem under the assumption that speech is quasi-stationary signal.

Chapter 4 reviews a closed form solution to the blind source separation of speech signals, considering speech to be a quasi-stationary signal [9]. This assumption along withe local dominance assumption is used to implement a closed form solution to the BSS problem.

Chapter 2

Blind Source Separation of Speech Signals

2.1 The cocktail party effect

Human beings, in particular, the human ears have an ability called binaural processing (or) binaural hearing[3]. This is the ability of the human ears to focus on one particular sound source even when it is in an environment where there are a lot of sound sources. This can be explained considering the environment of the cocktail party problem, where there are two people talking to each other. There are also many other people who would have visited that party, however, when the two people are talking to each other, both of them have the ability to focus only on what the other person has to say. This is known as the cocktail party effect and the ability of the human being(the human ear's ability) to do this is known as binaural processing. We can model this to a system which has the following characteristics:

- **An acoustic sensor array** In the example described above, there are two sensors, namely the ears of the two people who are in conversation.
- **A computational processing system** Most auditory models allow for several layers of signal processing considering how the information has to pass from the ear which does lower order computations on the signals that it receives to the brain that does much more complex signal processing with pattern recognition to perceive the speech.

There have been a lot of research going on to try an replicate this ability of the human ear(& brain) to a wide range of problems. One of the most widely researched areas in this aspect is its application to multi-microphone array processing, automatic speech recognition and natural language processing[3]. Potential commerical devices which could benefit from this include: Amazon Alexa, Siri, Google Personal Assistant and so on. Some other potential applications include audio teleconferencing systems and automobile speakerphones. Some of the other applications of blind source separation[7] in the scope of signal processing in other fields include:

- **Machine Monitoring** Signal separation can be used to identify potential mechanical failures by isolating the acoustic feature(in this case: the sound emitted by a mechanical device during damage) from an environment consisting of a mixture of other sources including other potentially normal working parts.
- **Medical diagnosis** Many medical devices often are used to read a lot of signals from the human body(eg.EEG signals, ECG signals). Signal separation can be used to

separate that signal of importance that can be tied to a particular bodily function (or) stimuli that may be present. This would probably prevent misdiagnosis due to influence of noise in the system.

- **Musical Performance** This is a particular application which can be used typically in recording musical performances to focus only on particular instrument sounds and voices in the recordings. This would in particular be used to amplify the sound of a particular instrument, say maybe a percussion which needs to be a little louder than usual.

- **Bio-informatics** Micro-Array data is a rather useful form of encapsulating the information presented in DNA and protein expression data. Blind source separation in multi-microphone array representation can potentially be used to separate the Gene or sequence of importance, to represent rather long sequences, detection of periodicity, clustering and classification of genes.

- **Seismic Monitoring** Long term prediction of seismic activity is often something that can be done in principle of how the tectonic plates of the earth move about. However, the problem of short term prediction is often not that simple. One of the things which usually characterizes an earthquake are modelled as a acoustic system with the signals of interest being acoustic waves and acoustic gravity waves and how they are propagated through the ionosphere. Blind source separation techniques can be used to isolate these signals. Thereby, we can predict the earthquake even a few hours before they occur.

In addition to this, blind source separation can also be used in a gamut of other applications such as :

- **Source (or) Feature Detection in video**
- **Fraud detection in banks and so on**

The rest of this chapter would focus on establishing a correlation between the blind source separation problem in speech and the cocktail party problem by defining an outline to the blind source separation problem and examining its similarities and differences from the cocktail party problem.

2.2 Blind Source Separation of Speech Signals

2.2.1 Problem structure and analogies to cocktail party effect

The authors in [7] define the problem statement of a standard BSS task and relate it to the cocktail party problem defined in the previous section. First, we consider the source signal vector sequence given by $s(k) = [s_1(k)s_2(k)...s_m(k)]^T$ where m is the number of sources and $s_i(k)$ is the i^{th} source signal, which in the cocktail party problem would correspond to a sampled version of the signal measured at its source position. These signals, then pass through a mxn LTI system which has an impulse response $\mathbf{A}_i$. where 0≤i≤ ∞, which would give us the signal that is being measured.

$$x(k) = [x_1(k)...x_n(k)]^T = \sum_{i=0}^{n} \mathbf{A}_i \mathbf{s}(k-i) \tag{2.1}$$

The entries of each (nxm) matrix $\mathbf{A}_i$ is deteremined by a number of things which include source locations, the sensor locations and the acoustical properties and so on. We group all of

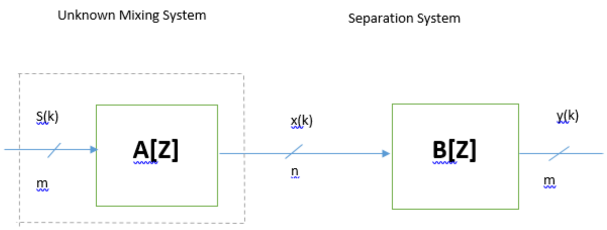

Figure 2.1: Block diagram of the convolutive BSS task

these properties and their impacts into a single value represented as en entry of the matrix $\mathbf{A}_i$. The mixing model is termed convolutive because the model represented in eq.(2.1) describes a multi-channel discrete time convolutive process. This is shown schematically in the Figure (2.1).

The term "blind" in the convolutive BSS task refers to the fact that other than general linear form of the source signals and the mixing system, we often know very little about them. Since, even in the absence of the knowledge about the source signals, their exact temporal(or) statistical properties is available, the blind source separation is a much more preferred and better alternative to the other traditional array processing methodologies. Another assumption made which is consistent to the previous works and my implementation would be that n≥m. i.e. the system is overdetermined. The multi channel- linear representation of such a system can be given below as follows:

$$\mathbf{y}(k) = \sum_{t=0}^{\infty} \mathbf{B}_l \mathbf{x}(k-l) \tag{2.2}$$

The sequence of the matrices describe the separation system and the output vector sequence y(k) contains the estimate of the individual source signals. Now that we have described the model, from the model intuitively used in binaural processing of speech in the human body, it is also imperative to describe certain differences between the two models(see [7]). They are as listed below :

- The processing model used in traditional convolutional BSS is in linear form, however, we do not have any idea of whether the model in binaural processing is linear or not.
- The number of sources m is no greater then the number of sensors n, which may not always be the case when it comes to binaural processing
- The content of the signals does not play any leverage when it comes to blind source separation of speech, however, intuitively the content of the speech signal does play a key factor when it comes to binaural processing of speech bu the human ear.

Considering all of the definitions and assumptions made over the course of this chapter, our next focus would be to define the goal of the BSS algorithm and assumptions made in this regard.

2.2.2 Goal of Convolutive BSS

The overall goal of the convolutive BSS as stated by the authors in [7] :

" Adjust the impulse response of the demixing system such that each output signal $y_i(k)$ contains one filtered version of each source signal $s_j(k)$ without replacement and loss of information".

Mathematically, this can be represented as

$$y_i(k) = \sum_{t=0}^{\infty} d_{ijl} s_j(k-l), 1 \leq i, j \leq m \tag{2.3}$$

Moving forward, the following assumptions are being made,i.e., the mapping from j− >i is arbitrary and the signal values,i.e. the entire signal length at a particular instant represented by the sequence d_{ijl} satisfies the following condition represented as

$$\sum_{l=0}^{\infty} d_{ijl} e^{i\omega l} \neq 0, \quad |\omega\| \leq \pi$$

for each valid pairi,j which would satisfy the above condition, where i=$\sqrt{-1}$. The convolutive BSS task does not make any assumptions about the temporal characteristics of the source signals, however it does make the following assumption about the source signal as detailed by the author in [7]:

Main Assumption: Each $s_i(k)$ is statistically independent of each $s_j(l)$, for all i≠j, all k and all l.

This assumption also implies that, for any two samples $s_1 = s_i(k)$ and $s_2 = s_j(l)$ from any two different source signals within the mixtures, the joint probability density function(p.d.f.) of s_1 and s_2 can be factored into the marginal p.d.f.'s as

$$p_{s_1,s_2}(s_1, s_2) = p_{s_1}(s_1)p_{s_2}(s_2) \tag{2.4}$$

When we look at BSS, statistical independence is a necessary condition, although it alone is not sufficient [7]. Another interesting thing to note is that the statistical independence of the two source signals is represented by the joint p.d.f's as the criteria(see[7]). We would proceed to look at pdfs and other criteria, which may be of use for BSS, in the next chapter.

2.3 Criteria for Blind Source Separation of Speech

The performance of the blind source separation algorithm to a large extent on the criteria based on which the signal is split. We already discussed the density based criteria in the form of pdfs in the previous section. Convolutive BSS criteria can be split into one of the three groups (see [7]) : i)Density modelling criteria, ii)Contrast functions, iii)correlation-based criteria.

Density modeling criteria

This criteria leverages a lot on the concepts from information theory. The amount of shared information between two signals is an important property which can be exploited for BSS. Intuitively, we can say that separation is possible if and only if there is no shared information between any two set of signals. The amount of information that is shared between two sets of signals is essentially characterized by making use of the Kullback Liebler divergence method which is a way of comparing two distributions, a true/actual distribution and an arbitrary/model distribution(see [7]) . The divergence can be modelled mathematically as shown below.

$$d(p_y||\hat{p}_y) = \int p_{\mathbf{y}}(\mathbf{y}) log(\frac{p_{\mathbf{y}}(\mathbf{y})}{\hat{p}_{\mathbf{y}}(\mathbf{y})})\mathbf{dy} \tag{2.5}$$

where $p_{\mathbf{y}}(\mathbf{y})$ and $\hat{p}_{\mathbf{y}}(\mathbf{y})$ are the trueactual and the model distributions respectively. We can write the above equation using expectation formulation E{.} as:

$$d(p_y||\hat{p}_y) = E\{log(\frac{p_{\mathbf{y}}(\mathbf{y})}{\hat{p}_{\mathbf{y}}(\mathbf{y})})\} \tag{2.6}$$

The choice of $\hat{p}_{\mathbf{y}}(\mathbf{y})$ is governed by the assumptions on and a priori knowledge of s(k). If all the $s_i(k)$ are identically distributed for all i, then we get an approximation as shown below

$$\hat{p}_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{m} p_s(y_i) \tag{2.7}$$

which can be used to obtain a maximum likelihood(ML) estimate of the demixing matrix B. The model can further be fine tuned with other considerations to get a better estimate of the demixing matrix B.

Contrast Functions

A contrast function identifies when one output $y_i(k)$ contains elements of only one source signal $s_j(k)$. The main goal in this approach would be to find such a function that depends only on $y_i(k)$ and not on the mixing conditions(see [7]). If we consider the combined system matrix C as a product of the mixing matrix A and an estimate B such that C=AB, then we can express the i^{th} extracted output signal in terms of the elements c_{ij} of C as(see[7])

$$y_i(k) = \sum_{j=1}^{m} c_{ij} s_j(k) \tag{2.8}$$

The contrast function is defined in [7] as follows:"A contrast function is a cost function $\Phi[y_i(k)]$ for which a local maximum over all elements of c_{ij},$1 \leq j \leq m$ corresponds to the separated solution."

$$c_{ij} = \begin{cases} d_l, \text{for a single value of l,} & 1 \leq l \leq m \\ 0, \text{otherwise} \end{cases} \tag{2.9}$$

In practice, this cost function is expressed in terms of the elements of the separation matrix **B** for which the optimization takes place. In the contrast based BSS, the following criteria need to be satisfied. That is the contrast function must obey the following 2 rules(see [7]).

- The contrast function must be simple to evaluate
- The contrast function must identify a separated result for the given source signal statistics through its maxima.

Like density based BSS, contrast based BSS also rely heavily on the spatial independence and non-Gaussianity of the source signals to perform separation. Significant knowledge about the source signal p.d.fs is not required in the case of contrast based criteria(see[7]), hence it is a better choice than density based criteria.

Correlation-Based Criteria

Both the density based BSS and the contrast based BSS approach employ non-quadratic criteria.Whenever, non-quadratic criteria is used, we often need to make use Convergence speed is often an issue in these cases, especially when we are dealing with audio signals as in the case of speech. In [17], the authors have proposed an approach for blind source separation where instead of making use of the density based criteria and the contrast based criteria, they use an approach which employs the correlation of the measured signal $\mathbf{x}$(k) at different time instants. The assumption here is the source signals measured are statistically independent and stationary but temporally correlated, such that the correlation matrix

$$\mathbf{R_{xx}}(k,l) = E\{x(k)x^T(k-l)\} \tag{2.10}$$

exhibits a unique eigenvalue structure fo at least two different time lags l=l_1 and l=l_2. Note that

$$\mathbf{R_{xx}}(k,l) = \mathbf{A}E\{x(k)x^T(k-l)\}\mathbf{A}^T \tag{2.11}$$

where the matrix $E\{x(k)x^T(k-l)\}$ is diagonal due to the independence of the source signals. Define the normalized matrix as follows:

$$\overline{\mathbf{R}}(l_1,l_2) = \mathbf{R_{xx}}(k,l_1)[\mathbf{R_{xx}}(k,l_2)]^{-1} = \mathbf{A}Es(k)s^T(k-l_1)\mathbf{A}^T[\mathbf{A}Es(k)s^T(k-l_2)\mathbf{A}^T]^{-1} \tag{2.12}$$

This simplifies down to $\mathbf{A}\overline{\mathbf{\Lambda}}(l_1,l_2)\boldsymbol{(}A\boldsymbol{)}^{-1}$, where we have defined $\overline{\mathbf{\Lambda}}(l_1,l_2) = E\{s(k)s^T(k-l_1)\}[E\{s(k)s^T(k-l_2)\}]^{-1}$. Since

$\Lambda(l_1,l_2)$ is diagonal, the above equation can be written in the form of an eigen value decomposition of $\overline{\mathbf{R}}(l_1,l_2)$, where the mixing matrix **A** contains the eigenvectors of this matrix. Thus, we can determine **A** from $\overline{\mathbf{R}}(l_1,l_2)$, using well known eigen value procedures from which the separation matrix **B** can be found by simply inverting the mixing matrix A. Joint Diagonalization is often used in conditions where the matrix inversion is very challenging. The efficiency of the blind source separation algorithm is also limited by the following aspects in addition to the ones already mentioned:

- Because the convolutive mixing is a linear process, a multi-channel linear system is sufficient to perform separation. Choice of filter is of critical importance to the algorithm's performance.

- Of all possible implementations, the FIR implementations represent the ideal candidates due to their simplicity and guaranteed stability. Using block implementations th complexity can be made as efficient as $O(log_2 L)$ where L is the system's filter length(see[7]).

- **Room Reverberation** In an ideal environment, there will not be much multipath due to deflections from obstacles. However, source separation is mostly(practically) performed in an environment emulating a room. Hence, we need to take into account the reverberation that takes place which may lead to multi-path propagation. The system in consideration needs to take this into account(see [7]).

- **Stability of the separating system** Most convolutive methods use adaptive procedures for adjusting the system parameters and the system must remain **Bounded Input- Bounded Output** stable during adaptation.

- **Computational Complexity** Room reverberation often adds to the signal being reflected off multiple paths and therefore leads to dispersive effect of the signal. This would lead to potential delays in the signal which may in turn result in the system having an impulse response which would be thousands of taps long to get a good estimate.

2.4 Signal Models

Now, that we have discussed how to model the signals to extract useful information from them for comparison, we would go on to define how acoustic problems in nature can be modelled. There are many ways in which the problems in nature can be modelled, however, the one of importance to us in this case is based on the number of inputs and the number of outputs that are there in the system, i.e, the number of sources and the number of sensors in the system. This would lead to the following four types of calculations as mentioned below(see [12]).They are

- Single-Input Single-Output {**SISO**} model
- Single-Input Multiple-Output {**SIMO**} model
- Multiple-Input Single-Output {**MISO**} model
- Multiple-Input Multiple-Output {**MIMO**} model

The sections that follow would illustrate the representation of each model as a system that performs convolution, a block diagram representation of the transfer functions of the system and the **z** transform representation of the transfer function.

2.4.1 SISO Model

The output signal in the SISO model is given by

$$x(k) = h * s(k) + b(k) \tag{2.13}$$

where h is the impulse response, the symbol represents the linear convolution operator, $s(k)$ is the source signal vector and $b(k)$ is the additive noise vector. Here, we assume that the system is linear and shift invariant, which are routinely used for formulating acoustic signal problems(see [12]). This model can be schematically represented as follows:

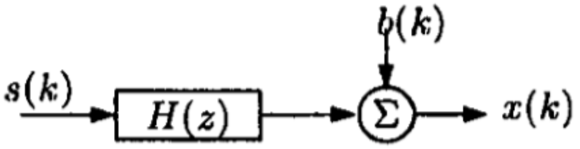

Figure 2.2: Block diagram of a model Single Input Single Output(SISO) system

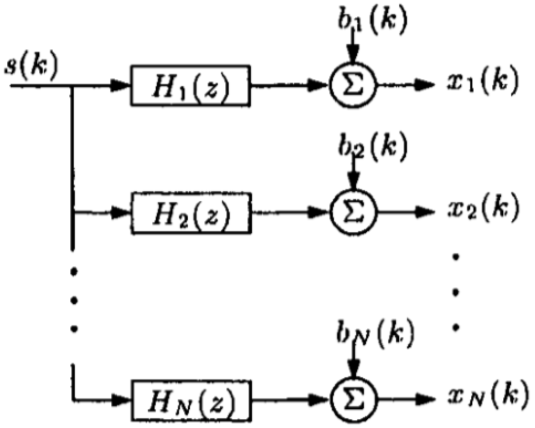

Figure 2.3: Block diagram of a model Single Input Multiple Output(SIMO) system

In the vector/matrix form, the SISO signal model is written as

$$x(k) = h^T s(k) + b(k) \tag{2.14}$$

where h=$[h_0\ h_1\ h_2\ ...\ h_{L-1}]^T$ and s(k)=$[s(k)\ s(k-1)\ ...\ s(k-L+1)]^T$ where h is the impulse response, the symbol represents the linear convolution operator, $s(k)$ is the source signal vector and $b(k)$ is the additive noise vector. Here, we assume that the system is linear and shift invariant, which are routinely used for formulating acoustic signal problems. This model can be schematically represented as follows:

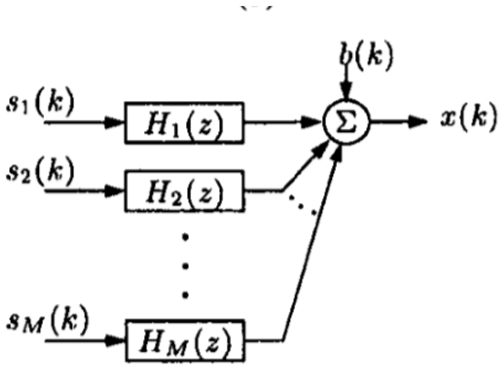

Figure 2.4: Block diagram of a model Multiple Input Single Output(MISO) system

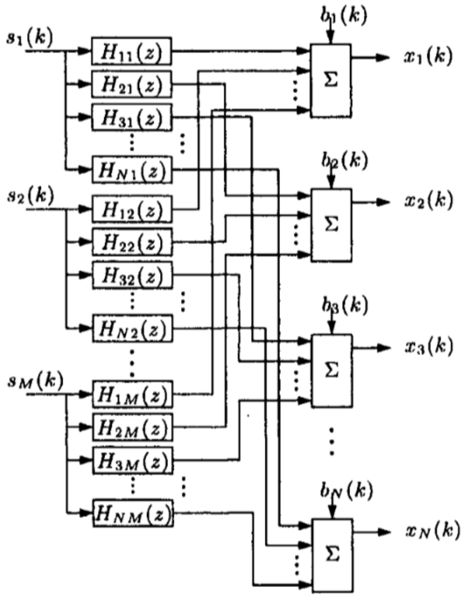

Figure 2.5: Block diagram of a model Multiple Input Multiple Output(MIMO) system

In the vector/matrix form, the SISO signal model is written as

$$x(k) = h^T s(k) + b(k) \tag{2.15}$$

where h=$[h_0\ h_1\ h_2\ ...\ h_{L-1}]^T$ and s(k)=$[s(k)\ s(k-1)\ ...\ s(k-L+1)]^T$. Using the z transform, the SISO signal model can be described as follows:

$$X(z) = H(z)S(z) + B(z) \tag{2.16}$$

where $X(z)$,$S(z)$ and $B(z)$ are the (z)-transforms of x(k),s(k) and b(k) respectively, and H((z))=$\sum_{l=0}^{L-1} h_l z^{-1}$. A schematic representation of the SISO model is given in Fig. 2.2.

2.4.2 SIMO Model

The diagram of a single input multiple output (SIMO) model is shown in Fig.2.3(see [12]). There are N output sources which are obtained from the same sound source and the n^{th} output is expressed as:

$$x_n(k) = h_n^T s(k) + b_n(k), \quad n = 1, 2, ..., N \tag{2.17}$$

where $x_n(k), h_n$ and $b_n(k)$ are defined as in the case of a SIMO model in the previous subsection and L is the longest channel impulse response in the SIMO system. A more comprehensive expression of the SIMO model is given as

$$x(k) = Hs(k) + b(k) \tag{2.18}$$

where x(k)=$[x_1(k)x_2(k)......x_N(k)]^T$ and $\mathbf{H} = \begin{bmatrix} h_{1,0} & h_{1,1} & . & . & . & h_{1,L-1} \\ h_{2,0} & h_{2,1} & . & . & . & h_{2,L-1} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ h_{N,0} & h_{N,1} & . & . & . & h_{N,L-1} \end{bmatrix}_{NXL}$

and b(k)=$[b_1(k)\ b_2(k)\ ...\ b_N(k)]^T$. The SIMO model can be expressed in the z-transform as

$$\overline{\mathbf{X}}(z) = \overline{\mathbf{H}}(z)S(z) + \overline{\mathbf{B}}(z)(k) \tag{2.19}$$

where $\overline{\mathbf{X}}(z)$=$[X_1(z)X_2(z)...X_N(z)]^T$, $\overline{\mathbf{H}}(z)$=$[H_1(z)H_2(z)...H_N(z)]^T$, and $H_n(z) = \sum_{l=0}^{L-1} h_{n,l}z^{-l}$, $n = 1, 2, ..., N$, $\quad \overline{\mathbf{B}}(z)$=$[B_1(z)B_2(z)...B_N(z)]^T$

2.4.3 MISO Model

The diagram of the Multiple Input Single Output(MISO) model is as shown in the Fig.2.4(see [12]). In this type, there are m sources whose signals are grouped to one in the output as :

$$x(k) = \sum_{m=1}^{M} h_m^T s_m(k) + b(k) \tag{2.20}$$

where $h = [h_1^T\ h_2^T\ ...\ h_M^T]^T$ and h_m=$[h_{m,0}\ h_{m,1}\ h_{m,2}\ .\ .\ .\ h_{m,L-1}\ ]^T$ and $s(k) = [s_1^T(k)\ s_2^T(k)\ ...\ s_M^T(k)]^T$ and $s_m(k) = [s_m(k)\ s_m(k-1)\ .\ .\ .\ s_m(k-L+1)]^T$. In the z-transform, the above model would look like

$$\overline{\mathbf{H}}(z) = [H_1(z)H_2(z)...H_M(z)]^T, \tag{2.21}$$

where each $H_m(z) = \sum_{l=0}^{L-1} h_{m,l}z^{-l}, \quad m = 1, 2, ..., M$ and $\overline{\mathbf{S}}(z)[H_1(z)H_2(z)...H_M(z)]^T$

Now, that we have described the first three models which are generally used in acoustic modelling, we go on to the model of importance that is the MIMO model.

2.4.4 MIMO model

A schematic representation of the MIMO model is shown in Fig.2.5. A MIMO system typically has M inputs and N outputs and is referred to as a MxN system(see[12]). At time k, the sytem can be represented as

$$x(k) = \mathbf{H}\mathbf{s}(k) + \mathbf{b}(k) \tag{2.22}$$

where x(k)=$[x_1(k)\ x_2(k)\ .\ .\ .\ x_N(k)]^T$ and $\mathbf{H} = [\mathbf{H}_1\ \mathbf{H}_2\ .\ .\ .\ \mathbf{H}_M]$, and the value for H_m can be given as obtained below:

$$\mathbf{H}_m = \begin{bmatrix} h_{1m,0} & h_{1m,1} & . & . & . & h_{1m,L-1} \\ h_{2m,0} & h_{2m,1} & . & . & . & h_{2m,L-1} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ h_{Nm,0} & h_{Nm,1} & . & . & . & h_{Nm,L-1} \end{bmatrix}_{NXL}$$

for m=1,2,3,4,......, M and b(k)=$[b_1(k)\ b_2(k)\ .\ .\ .\ b_N(k)]^T$ where $h_{nm}(n = 1, 2, 3, ...N, m = 1, 2, 3, ..., M)$ is the impulse response of the channel from input m to output n, and s(k) is defined similarly(see [12]). Transforming into the z- domain we get,

$$\overline{\mathbf{X}}(z) = \mathbf{H}(z)\overline{S}(z) + \overline{\mathbf{B}}(z)(k) \tag{2.23}$$

$$\text{where } \mathbf{H(z)} = \begin{bmatrix} h_{11}(z) & h_{12}(z) & . & . & . & h_{1M}(z) \\ h_{21}(z) & h_{22}(z) & . & . & . & h_{2M}(z) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ h_{N1}(z) & h_{N2}(z) & . & . & . & h_{NM}(z) \end{bmatrix}_{NXL}$$

and $h_{nm}(z) = \sum_{l=0}^{L-1} h_{nm,l} z^{-l}$, n=1,2,....,N m=1,2,....,M. Clearly, the MIMO system is the most general model and we can write all the other models as special case situations of the MIMO model(see [12]). Hence, we derive the BSS algorithm for the MIMO model and the same would hold true for the other three models as well.

Chapter 3

Review of Previous Works

Blind Source Separation of speech has been an area of research for a long period of time. This has resulted in a variety of research papers which illustrate a number of different ways to approach the same problem. This chapter is an exploration of some of the various works done along this line. This chapter would illustrate the previous works by describing their problem statement and a list of solutions that they provided for the same.

3.1 Blind source separation of non stationary Sources

In [20], the authors have tackled the problem of blind source separation by exploiting the non-stationarity of the sources. A least squares optimization technique was proposed to estimate a forward model, to identify the channel components. Similarly, the authors in [20] have proposed a FIR backward model which would generate the well separated model sources. In [20], the authors propose solutions for the estimation of both an Instantaneous Mixture and a Convolutive Mixture. They first propose a backward model for estimation in the instaneous case and then extend it to the convolutive case where it is solved as independent models for each frequency

3.1.1 Instantaneous Mixtures

Forward Propagation

For instantaneous mixtures, an illustration of the forward model used is as shown below[20].

$$x(t) = As(t) + n(t) \tag{3.1}$$

From this definition for the forward model, we can obtain/postulate the covariance $R_x(t)$ of the measured signals at time t with the assumption of independent noise [20] as follows

$$R_x(t) = x(t)x^T(t) = As(t)s^T(t)A^T + n(t)n^T(t) \tag{3.2}$$

From this we can write, the expression for the covariance matrix as being equal to $R_x(t) = A\Lambda_s(t)A^T + \Lambda_n(t)$ where the covariance matrices $\Lambda_s(t)$ is a diagonal matrix[20]. The authors also assumed that the noise is uncorrelated at each sensor, i.e. $\lambda_n(t)$ is also diagonal. For non stationary signals, a set of K equations(for 3.2) for different times $t_1, t_2, t_3, ...t_K$ and the d_s scaling conditions result in a a total of $Kd_x(d_x+1)/2 + d_s$ constraints on $d_sd_x + d_sK + d_xK$ unknown parameters $A, \lambda_s(t_1), ...\lambda_s(t_K), \lambda_n(t_1), ...\lambda_n(t_K)$.

Assuming all the conditions are linearly independent, there will be sufficient conditions if

$$Kd_x(d_x+1)/2 + d_s \succeq d_sd_x + d_sK + d_xK \tag{3.3}$$

When we have the value of dx and ds as eqaul, there is a lack of constraints based upon which we estimate unless $d_x \succeq 4$. which is outlined in [20].The solutions can be found for the square case by considering the non symmetric eigen value problem outlined in [17]. As in [17], we measure the sample estimates $\hat{R}_x(t)$ within some time interval and we use the inaccuracy of that estimation as measurement error

$$E(k) \equiv \hat{R}_x(k) - \lambda_n(k) - A\lambda_s(k)A^T \tag{3.4}$$

The unknown parameters can be estimated by minimizing the total measurement error for a sufficiently large K

$$\hat{A}, \hat{\lambda_n}, \hat{\lambda_s} = \underset{A,\lambda_s,\lambda_n,A_{ii}=1}{\arg\min} \sum_{k=1}^{K} ||E(k)||^2 \tag{3.5}$$

The matrix norm here now is the sum if the absolute squared error of every coefficient which essentially represents a Least Squares(LS) estimation Problem. To find the extrema of the LS cost $J = \sum_{k=1}^{K} ||E(k)||^2$ in (3.5), the gradients with respect to the parameters are calculated as follows

$$\frac{\partial J}{\partial A} = -4\sum_{k=1}^{K} E(k)A\lambda_s(k) \tag{3.6}$$

$$\frac{\partial J}{\partial \lambda_s(k)} = -2diag|A^T E(k)A| \tag{3.7}$$

$$\frac{\partial J}{\partial \lambda_n(k)} = -2diag|E(k)| \tag{3.8}$$

We can find the minimum value for A and $\lambda_s(k)$ with a gradient descent algorithm on (3.6) and (3.7). The optimal $\lambda_n(k)$ can be computed explicitly for every given A and $\lambda_s(k)$ by setting the gradient in (3.8) to zero which would yield us the expression $\hat{\lambda_n}(k) = diag[\hat{R}_x(k) - A\lambda_s(k)A^T]$

Estimation of the source signals

In case of a square and invertible matrix $\hat{A}$, the signal estimates are $\hat{s} = \hat{A}^{-1}x$, whereas in the non square case we can compute the LS estimate as

$$\hat{s}_{LS}(t) = \underset{s(t)}{\arg\min} \|x(t) - \hat{A}s(t)\| = (\hat{A}^T\hat{A})^{-1}\hat{A}^T x(t) \tag{3.9}$$

If the noise is assumed to be Gaussian(not necessarily white or stationary) we can compute the maximum likelihood estimate as

$$\hat{s}_{ML}(t) = \underset{s(t)}{\arg\max} p[x(t)|s(t)l\hat{A}, \hat{\lambda_n}(t)] = [\hat{A}^T\hat{\lambda_n}(t)^{-1}\hat{A}]^{-1}\hat{A}^T\hat{\lambda_n}(t)^{-1}x(t) \tag{3.10}$$

where p() is the Gaussian probability density given by the noise density. The authors also came up with a Maximum-A-Posteriori(MAP) estimate of the source signals in [20]. Assuming that the model defined in(3.1) is correct, we can say that we find the correct estimate $\hat{\mathbf{A}} = \mathbf{A}$. We can represent this as follows:

$$< \hat{s}_{LS}\hat{s}_{LS}^T > \approx < \mathbf{ss}^T > + (\hat{A}^T\hat{A})^{-1}\hat{A}^T\Lambda_n\hat{A}(\hat{A}^T\hat{A})^{-1}$$

The authors in [20] found that the resultant estimates have a degree of correlation in them which is caused due to the noise component and the signal component on the other hand remains uncorrelated.

Backward Model

We can also try to directly estimate the source signals by making use of a model similar to the inverse FIR model estimate as shown in [20] with a model which looks like the one defined below

$$\hat{s}(t) = WAs(t) \tag{3.11}$$

The aim of using the backward model is to find a good approximation of W which would invert the mixing matrix denoted by 'A'(see[20]). In analogy to the equation (3.2) and (3.4), we have

$$< \hat{s}(t)\hat{s}(t)^T >= W[R_x(t) - \lambda_n(t)]W^T \tag{3.12}$$

The aim of the algorithm is to find a W such that $< \hat{s}(t)\hat{s}(t)^T >$ diagonalizes simultaneously for K different times.

The Least square estimate is then

$$\hat{W}, \hat{\lambda_s}, \hat{\lambda_n} = \underset{W,\lambda_s,\lambda_n,W_{ii}}{\arg\min} \sum_{k=1}^{K} ||E(k)||^2 \tag{3.13}$$

where $E(k) = W(\hat{R}_x(k) - \lambda_n(k))W^T - \lambda_s(k)$ Similar to the forward model, the solutions can be found using iterative gradient algorithm.

3.1.2 Convolutive Mixture

Model Description

The model for forward propagation in the convolutive case is as follows:

$$x(t) = A * s(t) + n(t) \tag{3.14}$$

The authors in [20] address the blind source separation problem by transforming it into the frequency domain and solve it for every frequency as in [4][8][13][15][18]. However, every frequency was found to have an arbitrary permutation. Hence, the main goal of Blind Source Separation [20]in this context is:

- Obtain the equivalent equation to (3.2) in the frequency domain
- Choose arbitrary permutations for all individual problems consistently

Cross Correlations

The cross correlation can be represented mathematically as

$$R_x(t, t+\tau) =< x(t)x(t+\tau)^T >$$

. For stationary signals the absolute time does not matter and the cross correlations can therefore be written as $R_x(t, t+\tau) = R_x(\tau)$. The z-transform representaion is obtained and can be expanded using the definition given in (3.2) as follows:

$$R_x(z) = A(z)\Lambda_s(z)A^H(z) + \Lambda_n(z) \tag{3.15}$$

where

- A(z): The matrix of the z-transforms of the FIR filters A(τ)
- $\Lambda_s(z)$ and $\Lambda_n(z)$ are the z-transform of the auto-correlations of the source and the noise. Because of the independence assumption, both $\Lambda_s(z)$ and $\Lambda_n(z)$ are diagonal.

The authors restrict the total number of sampling points of z by taking T equidistant samples on the unit circle, so as to exploit the DFT representation of the auto-correlation function. If the signals are periodic, we can express the circular convolutions as products as represented in (3.15).We can then write the expression as follows

$$x(\omega, t) \approx A(\omega)s(\omega, t) + n(\omega, t), \qquad forP << T \tag{3.16}$$

where $x(\omega, t)$ represents the DFT of the frame of size T starting at T and correspondingly for $s(\omega, t)$ and $A(\omega)$. This is mostly the case for stationary signals. We go for a cross power spectrum average that diagonalizes for the source signals. This is because for the assumptions being made(non-stationarity), the R_x will be time dependent(see [20]). One such average is

$$\overline{R}_x(\omega, t) = \frac{1}{N}\sum_{n=0}^{N-1} x(\omega, t + nT)x^H(\omega, t + nT) \tag{3.17}$$

We can then write for all averages

$$\overline{R}_x(\omega, t) = A(\omega)\Lambda_s(\omega, t)A^H(\omega) + \Lambda_n(\omega, t) \tag{3.18}$$

If N is sufficiently large, then we can model $\Lambda_s(\omega, t)$ and $\Lambda_n(\omega, t)$ as diagonal due to independence assumption. In case of the Convolutive mixture, the forward model **A** does not always guarantee a stable inverse prediction. Therefore, we use the backward model to solve the BSS problem.

Backward Model

In the case of the backward model, we use a model which is similar to what we did for the instantaneous mixture in (3.1.1),i.e., to find model sources with cross power spectra satisfying

$$\Lambda_s(\omega, t) = \mathbf{W}(\omega)[\overline{R}_x(\omega, t) - \Lambda_n(\omega, t)]\mathbf{W}^H(\omega) \tag{3.19}$$

In order to obtain the conditions for BSS, we choose times such that we have non overlapping averaging times for $\overline{R}_x(\omega, t_k)$,i.e.,$t_k$=kTN. A multipath model W that satisfies these equations K times simultaneously can be found again with a LS estimate as given below

$$E(\omega, k) = \mathbf{W}(\omega)[\overline{R}_x(\omega, k) - \Lambda_n(\omega, k)] - \Lambda_s(\omega, k) \tag{3.20}$$

$$\hat{W}, \hat{\Lambda_s}, \hat{\Lambda_n} = \underset{W,\lambda_s,\lambda_n,W(\tau)=0,\tau>Q<<T,W_{ii}(\omega)=1}{\arg\min} \sum_{\omega=1}^{T}\sum_{k=1}^{K} ||E(\omega, k)||^2 \tag{3.21}$$

Like in the instantaneous mixture model, the authors make use of a gradient descent algorithm to find an estimate. For any real valued function f(z) of a complex valued variable z, the gradients with respect to the complex valued coefficent $\mathbf{W}(\omega)$ are obtained by taking derivatives with respect to the conjugate quantities z^*, ignoring the non conjugate occurrences of z., i.e.,

$$\frac{\partial f(z)}{\partial(z)} + i\frac{\partial f(z)}{\partial \Im(z)} = 2\frac{\partial f(z)}{\partial z^*} \tag{3.22}$$

There for the corresponding gradients for the LS cost in(3.21) are as follows:

$$\frac{\partial J}{\partial \mathbf{W}^*(\omega)} = 2\sum_{k=1}^{K} E(\omega,k)\mathbf{W}(\omega)[\overline{R}_x(\omega,k) - \Lambda_n(\omega,k)] \tag{3.23}$$

$$\frac{\partial J}{\partial \Lambda_s^*(\omega,k)} = -diagE(\omega,k) \tag{3.24}$$

$$\frac{\partial J}{\partial \Lambda_n^*(\omega,k)} = -diag[(W)^H(\omega)E(\omega,k)W(\omega)] \tag{3.25}$$

As in the case of the instantaneous mixture, we can find the minimum with respect to $\mathbf{W}(\omega)$ and $\Lambda_n(\omega,k)$ with a constrained gradient descent algorithm using the gradients from (3.23) and (3.25). Similarly, the optimal$\Lambda_s(\omega,k)$ for given $(W)(\omega)$ and $\Lambda_n(\omega,k)$ at every step can be computed explicitly by setting the gradient in (3.24) to zero which yields $\Lambda_s(\omega,k)$=diag$[(W)^H(\omega)E(\omega,k),(W)(\omega)]$.

Permutation and Constraints

The arbitrary permutation of the co-ordinates for each frequency ω will lead to the same error E(ω,k)(see [20]). As a result of this, the total cost will not change if we choose a different permutation of the solutions for each frequency ω. This should not be the case as only consistent permutations of the constraints should be able to correctly reconstruct the source signals. Arbitrary permutations will not satisfy the conditions on the length of the filter,i.e.$\mathbf{W}(\tau) = 0$ for $\tau > Q \ll T$ as stated in [20]. The constraint of $\tau >$Q will restrict the solution to be continuous (or) "smooth" in the frequency domain. This constraint links the otherwise independent frequencies, and solves the frequency permutation problem. These are then enforced by properly projecting the unconstrained gradient to the subspace of permissible solutions[20]. The projection operator that zeros the appropriate delays for every channel $W_{ij} = [W_{ij}(0), ..., W_{ij}(\omega), ..., W_{ij}(t)]^T$ is

$$P^{(2)} = FZF^{-1} \tag{3.26}$$

where the DFT is given by $F_{ij}=\frac{1}{\sqrt{T}}e^{-\mathbf{i}2\pi ij}$ and Z is diagonal with Z_{ii}=1 for $i < Q$ and Z_{ii}=0 for i$\geq$ Q[20].

To sum up the contributions of this paper, we solve the problem by obtaining a constrained LS cost that is optimal to the desired solutions[20].

3.2 Blind Source Separation of Speech using Second Order Cyclo-stationary Statistics

3.2.1 Problem Statement

In [2], the authors defined the problem statement as follows : "Assume that m source signals impinge on an array of n sensors where n$\geq$ m. The output of each sensor is modeled as a weighted sum of the source signals corrupted by additive noise. This can be expressed as

$$x(t) = y(t) + w(t) = As(t) + w(t) \tag{3.27}$$

where $s(t) = [s_1(t), ..., s_m(t)]^T$ is the mx1 complex source vector and the nx1 complex noise vector is W(t)=[$w_1(t)$,...,$w_n(t)$]^T , $A = [a_a, ..., a_m]$ is the unknown nxm full rank mixing matrix where T: transpose of the matrix/vector".

The source signal vector s(t) is modeled as a cyclo-stationary complex stochastic process. That authors in [2] assume that the component processes mutually independent with zero mean. This would essentially mean that the following conditions are interpreted to hold true

$$< e^{J\beta_i t} s_i(t+\tau) s_j^*(t) >= 0 \qquad if i \neq j \tag{3.28}$$

$$< e^{J\beta_i t} s_i(t+\tau) s_i^*(t) >= 0 \qquad if \beta_i \neq \beta_j \tag{3.29}$$

$$< e^{J\beta_i t} s_i(t) s_i^*(t) >> 0 \qquad \forall i \tag{3.30}$$

Here J=$\sqrt{-1}$ and $< . >$ denotes the time averaging operator(see [2]). Further more each β_i is a non zero cycle frequency of each source i. The cyclic auto-correlation function $\rho_i(\tau)$ is defined to be as follows

$$\rho_i(\tau) \overset{\text{def}}{=} < s_i(t+\tau) s_i^*(t) e^{J\beta_i t} > \qquad with \rho_i(0) > 0 \tag{3.31}$$

Before proceeding with the rest of this section, a few notations which are used are listed as follows : * denotes the complex conjugate whereas $\star$: denotes the complex conjugate transpose of a vector(see [2]). The additive white noise is also modelled as

$$< e^{J\beta_i t} w_i(t+\tau) w_i^*(t) >= 0 \qquad \forall i, \tau \tag{3.32}$$

The output cyclic correlation function $R_x^{(\beta_i)}(\tau)$ is defined to be

$$R_x^{(\beta_i)}(\tau) \overset{\text{def}}{=} < e^{J\beta_i t} x(t+\tau) x^\star(t) > \tag{3.33}$$

Under the assumptions mentioned above, the cyclic auto-correlation function takes a form as shown below

$$R_x^{(\beta_i)}(\tau) = \sum_{j|\beta_j=\beta_i} \rho_j(\tau) a_j a_j^\star \tag{3.34}$$

where the sum is over all sources with frequency β_i(see[2]). When $\beta_i \neq \beta_j$, for i≠j, then only one source i contributes to $R_x^{(\beta_i)}(\tau)$ which can be written as

$$R_x^{(\beta_i)}(\tau) = \rho_i(\tau) a_i a_i^\star. \tag{3.35}$$

The authors in [2], give the following definition for the problem of blind source separation : i.e. To find a mxn separating matrix **B**=$[b_1, ..., b_n]$,such that $\hat{s}(t) = \mathbf{B}x(t)$ is an estimate of the source signal. An assumption made to get a close approximation of the separating matrix is shown as follows : We assume that the emitter(source) signals have unit-norm zero-lag cyclic auto-correlation coefficients, i.e.

$$\rho_i(0) =< e^{J\beta_i t} s_i(t) s_i^*(t) >= 1 \tag{3.36}$$

The authors in [2] propose to determine **B** up to a permutation and scaling of its columns, i.e. **B** is a separating matrix if **B y(t)**=$\mathbf{P}\Lambda s(t)$ where P is a permutation matrix and Λ is a unitary diagonal matrix(see[2]). If all sources have distinct cyclic frequencies, then the mxing matrix is defined as **By(t)**=Λ**s(t)** for given unitary matrix Λ.

3.2.2 Assumptions and Remarks

- For simplicity we use the definition for cyclo-stationary sources in [11]. However, the authors in [2] have given a more rigorous definition of a cyclo-stationary sources as given below

“A zero-mean cyclo-stationary process s(t) is characterized by the property that its time varying auto-correlation $r_s(t, \tau) = E(s(t + \tau)s * (t))$ varies periodically with respect to time”.

Thus, it accepts a Fourier series representation given by

$$r_s(t, \tau) = \sum_{\beta \in C} r_s^{\beta}(\tau) e^{-j\beta t} \tag{3.37}$$

$$r_s^{\beta}(\tau) = \lim_{t \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} r_s(t, \tau) e^{j\beta t} \tag{3.38}$$

where the Fourier Coefficients r_s^{β} are called the cyclic auto-correlation at cycle frequency β, and

$$C = \{\beta | 0 \leq \beta < 2\pi and r_s^{\beta}(\tau) \neq 0\} \tag{3.39}$$

is the cyclic frequency of the set s(t)

- This solution takes into consideration only one cycle frequency for each source signal. However, in practice, the sources’ energy may be distributed to more than one cycle frequency[2]. In this case , we can replace $R_x^{(\beta_i)}(\tau)$ by a linear combination of cyclic correlation matrices that adds coherently the energy of the considered source over its different cycle frequencies (or) Alternatively, we can use several cycle frequencies for each source signal.

- The mutual independence of the sources is a fundamental condition for separation of sources

3.2.3 Condition for identifiability

This section states some of the essential conditions for blind source separation via second order cyclo-stationary statistics of the array output as proposed by the group in [2].The condition for identifiability follows from the two theorems that are given below(see[2]). It is a necessary and sufficient condition for BSS using only the cyclic correlation matrices $R_X^{(\beta_i)}(\tau)$, i=1,.......,m at time lags 0,τ_1,τ_2,.......,τ_K. The theorems and the corresponding identifiablility condition(see[2]) are as follows:

- **Theorem 1(ATH1):** Assume that the cyclic frequencies of the source signals are distinct. For any matrix B, define **z**(t) to be the mx1 vector given by **z**(t)=**Bx**(t). In addition, define its cyclic cross correlation $r_{ij}(\tau) \overset{\text{def}}{=} < z_i(t+\tau)z_j^*(t)e^{J\beta_i t} >$. then, B is a separating matrix if and only if $r_{ij}(0) = 0$ and $r_{ii}(0) = 1$ for all $1 \leq i \neq j \leq m$(see [2]).

- **Theorem 2(ATH2):** Assume that the identifiability condition is satisfied,that is, if $\beta_i = \beta_j$, then ρ_i, and ρ_j are linearly independent. then **B** is a separating matrix if and only if $r_{ij}(k) = 0$ and $r_{ii}(0) = 1$, for all $1 \leq ij \leq m$. This is just a general case of the previous theorem(see[2]). Both these can be used to define the identifiability condition which is given below.

- **Identifiability condition :** "For any K $\geq$ 0, blind source separation can be achieved using the output cyclic auto-correlation matrices $\{R_x^{(\beta_i)}(\tau)|i = 1, 2, 3, ...m; \tau_1, \tau_2, ..., \tau_K\}$ if and only if there exists two distinct source signals s_i(t) and $s_j(t)$ whose cyclic frequencies are the same(β_i=β_j) and auto-correlation vectors are linearly independent"

.

The authors in [2] also define conditions for partial identifiability of the source signals in a manner analogous to the one they followed for complete identifiability. These algorithms are respectively termed in [2] as ATH3 and ATH4. Theorem 1 and theorem 2 have an iterative implementation being deived in [2]. Further, an adaptive version of the above algorithms can also be derived using the approaches from [5], [6] and [22]. To sum up the contributions of the paper done by the group in [2], it can be cited as follows

- An iterative algorithm(ATH2)(see [2]) was derived to separate the sources even when they do not have distinct cyclic frequencies.
- If the cyclic frequencies are distinct, then the ATH2 simplifies to ATH1(see [2]).
- A non iterative algorithm (ATH4)(see [2]) was derived by the authors to separate only those sources of a particular cycle frequency.
- When all the source signals have distinct cycle frequencies ATH4 simplifies to ATH3(see [2]).

Chapter 4

Speech as a Quasi-stationary source

This chapter focuses on the implementation that has been done over the course of my graduate studies at UCR. The problem formulation for the blind source separation of speech signals that is used for my implementation follows the Blind source Separation of quasi stationary sources(BSS-QSS) formulation given in [9], where in the observed signals are a linear instantaneous mixture of sources. This is represented as follows:

$$x(t) = \sum_{k=1}^{K} a_k s_k(t) = As(t) \tag{4.1}$$

where x(t) ϵ R^N is the observed signal vector and $s_k(t)$ is the kth source signal and s(t)=[$s_1(t)$ $s_2(t)$ $s_3(t)$ $s_4(t)$ ....., $s_K(t)$] ϵ R^k and a_k ϵ R^N is the system response vector of the kth source, and the mixing matrix A can be represented as follows $A = [a_1, a_2, ..., a_K]$ ϵ R^{NxK} where N and K are the number of sensors and sources respectively. The algorithm that I have implemented makes the following assumptions and implements the closed form solution to

Blind Source Separation problem as derived in [9]. The following are the assumptions made :

1)The source signals are assumed to be zero mean wide sense quasi stationary sources with frame length L i.e.

$$E\{s_k(t)^2\} = d_k(m); \quad \text{for any (m-1)L+1} \leq t \leq mL, (4.2)$$

where m is the local time frame index. This means that the source second order statistics are constant within the time frame defined by m and that it varies from one time frame to another. If we denote the local covariance of the observed signals in frame m as R(m), then we can define it as shown in the following expression:

$$R(m) = E\{x(t)x(t)^T\}; \quad (m-1)L + 1 \leq t \leq mL. \tag{4.3}$$

Such local covariances are in fact obtained by local averaging. Applying this, the equation further becomes(see [9]):

$$R(m) = \frac{1}{L} \sum_{t=(m-1)L+1}^{mL} x(t)x(t)^T; \quad m = 1, 2, 3, ..., M \tag{4.4}$$

Now, using the expressions for x(t) obtained from equations (4.1), and the value of their expectation obtained from (4.2), we would be getting

$$R(m) = \sum_{k=1}^{K} d_k[m] a_k a_k^T; \quad m = 1, 2, 3, ..., M \tag{4.5}$$

where M is the total number of available frames.

2)Now that, the characteristics of the source signals have been defined, we next proceed to define the local dominance assumption which is going to form the crux of the solution that is being implemented(See [9]).

A1) (local dominance assumption) For each source index k, there exists a time frame indexed by m_k such that $d_k[m_k] > 0\, and\, d_l[m_k] = 0$, for all l≠k.

What this really means is that there exists certain time indices in the speech signal where only one source dominates. Temporally sparse signals such as speech signals that we have taken often tend to satisfy this property. Under local dominance[9], the expression of covariance matrix becomes the following

$$R[m_k] = d_k[m_k]a_k a_k^T \tag{4.6}$$

where k is the source which is locally dominant within the time frame 'm'. We can obtain the a_k's as principal eigenvalues of R(m). However, in order to do so, we need to know where the locally dominant points are (see [9] and [10]).

Traditional approach: Clustering Techniques

Other blind source separation techniques would predominantly make use of the following approach(see [9]):

- Detect Locally dominant points by evaluating the ranks of all R[m]'s
- Extract the principal eigenvector of each detected R(m)
- Apply a K-means clustering algorithm to the obtained principal eigenvectors to construct the mixing Matrix A.

This is the traditional algorithm used in most of the papers. However, the algorithm implemented does not follow this approach, we follow a slightly different approach which is simple and the arithmetic operations involved can be done using simple 2-norm computations and linear projections mostly. This reduces the overall complexity of the algorithm over the ones which have been previously used.

Closed Form solution

Instead of using the rank as used in a clustering based technique, the authors in [9] employ another idea where they use the non negativity of the local source covariances $d_k(m)$ together with the assumption of local dominance of a particular source covariance to come up with an alternate solution to identify the locally dominant points. The algorithm employed makes use of the successive search strategy and does not require clustering. We consider the over-determined solution that is N > K.

Again as in the previous technique we need to make a few assumptions with this one. The assumptions being made are listed below as follows(see [9]):

A1)Local dominance: The local dominance assumption is the same as what was made in the previous case for the BSS. This means that there would exist some local time frame where only one source dominates. This assumption applies very well to the speech signals because speech signal also contains may unvoiced segments between utterances.

A2)The mixing matrix A is orthonormal: That is the matrix A is a collection of k vectors such that $a_1, a_2, a_3, a_4, ..., a_k$ has a unit norm 1 , i.e., $\|a_i\|$=1 and the vectors are mutually orthogonal with each other.

Given these two assumptions are holding strongly, we can simplify the problem further as follows. By applying pre-whitening, we can convert the mixing model with a full rank mixing matrix A into an equivalent model where A is orthonormal.Now, from the local covariance model as seen in (4.6), we can apply vectorization to our output model to obtain

$$y[m] = vec(R_m) = \sum_{k=1}^{K} d_k(m) vec(a_k a_k^T) = Hd[m] \tag{4.7}$$

where $H = [h_1, ..., h_K]$, $d[m] = [d_1[m], ..., d_K[m]]^T$ and $h_k vec(a_k a_k^T) = a_k \otimes a_k$ with $\otimes$ being the Kronecker function. Because we assumed that the original mixing matrix A is orthonormal, it can be very easily shown that H is also orthonormal. This means that we have from equation(4.7),

$$\|y(m)\|_2 = \|d(m)\|_2 \leq \|d(m)\|_1 \tag{4.8}$$

where $\|.\|_2$ and $\|.\|_1$ are the 2 norm and the 1 norm respectively, this equality will be satisfied only if the 2-norm is less than or equal to the 1-norm and if the vector whose norm is being taken is a scaled unit vector(See [9]). The reason the 1-norm of the signal coefficient is being chosen was explained below.

It is important to note that $||\tilde{s}_i||_0$ is discontinuous and may be difficult to optimize for any $\tilde{s}_i$. Also, the l_0 is highly sensitive to noise, in that even a tiny amount of noise could make all the samples non zero. Therefore, in most cases as in the case of this two source and two mixture model, we make use of the l_1 norm as a very good substitute for the objective function we choose to minimize[13][14]. We know that the one norm can also be written as follows

$$\|d_1(m)\|_1 = \sum_{k=1}^{K} d_k[m] = Tr(R(m)) \tag{4.9}$$

where Tr is the trace of the covariance matrix R(m). Also, we have the assumption of non negativity in the space from (m-1)L+1 to mL as $d_k(m) \geq 0$ where the $d_k(m)$ are modeled as local source covariances. From (4.8) and (4.9), we can obtain

$$\frac{\|y(m)\|_2}{Tr(R[m])} \leq 1 \tag{4.10}$$

which is true when d[m] is a scaled unit vector, i.e., y(m) is locally dominant taking the form $y[m] = d_k[m]h_k$ for some k(see [9]). As a result, we can obtain, the locally dominant time frame for each source k from the following expression as

$$\hat{m} \in \max_{\text{m=1, 2, ..., M}} \frac{\|y(m)\|_2}{Tr(R(m))} \tag{4.11}$$

This is iteratively done again and again to provide a search for all h_k for all the k sources(see [9]). This search can be done as follows. Suppose that we have obtained the first k-1 columns for the matrix H as $H_{\text{1:k-1}} = [h_1, ..., h_k - 1]$, and the projection vector onto this can be explained as $P_X^{\perp} = I - X(X^T X)^{\dagger} X^T$ where $P_X^{\perp}$ is the orthogonal complement projector of its argument X. We can find the similar locally dominant points[9] using the expression shared below

$$\hat{m} \in \max_{\text{m=1, 2, 3,...., M}} \frac{\|P_{\text{H}_{\text{1:k-1}}}^{\perp} y(m)\|_2}{Tr(R(m))} \tag{4.12}$$

which are the locally dominant points corresponding to their respective vectors[9].

Based on these findings the algorithm 1 was formulated in [9] as follows

Algorithm

```
1 input:R[1],..........,R[M];
2 y[m]= vec(R[m]),    z[m]=Tr(R[m]),   m=1,2,3,4,..........., M;
$\hat{h_1} = y[\hat{m_1}]$ where $\hat{m_1} \in \max_{\text{m=1,2,3,....., M}} \|y[m]\|_2/z[m]$
3 Obtain $\hat{a_1}$ as the principal eigen vector of $vec^{\text{-1}}(\hat{h_1})$
4 for k=2,......., K do
```

```
    $\hat{h_k} = y[\hat{m_k}]$ where
    $\hat{m_k} \in \underset{\text{m=1,2,3,....., M}}{argmax} \frac{\|P^{\perp}_{\text{H}_{\text{1:k-1}}} y(m)\|_2}{z[m]}$
    Obtain $\hat{a_k}$ as the principal eigen vector of $vec^{\text{-}1}[\hat{h_k}]$
end
output: $\hat{A} = [\hat{a_1}, ........, \hat{a_k}]$
```

The final output is an estimate of the mixing matrix and the efficiency of the overall method is calculated using a Mean Square Error(MSE) estimate.

The mean Square error is calculated as follows:

$$MSE = \sum_{k=1}^{1000} ||A_k - H_k||_2^2 / size(A_k)$$

where A_k is the original mixing matrix given and H is the prediction of the mixing matrix. The estimate is run for a total of 1000 times.

The average Mean square error estimate is then calculated and was -41.2036. The results obtained after the simulation are represented pictorially as shown below.

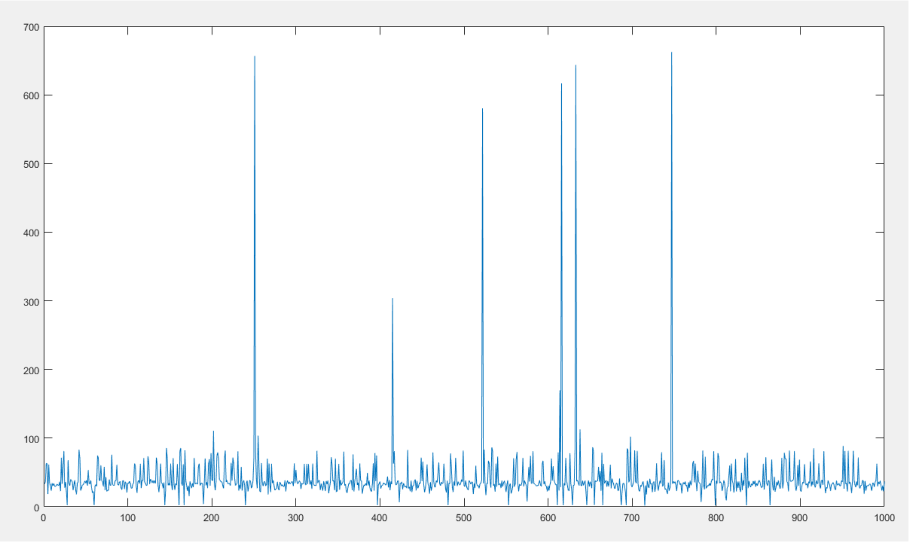

Figure 4.1: Mean Square error estimates for 1000 iterations

Chapter 5

Conclusions

The problem of Blind source separation of speech for a Multi-Microphone Array system has been discussed and analyzed in detail. The algorithm is being reviewed for the Multiple-Input Multiple-Output(MIMO) model of acoustic system as it was shown that the algorithm would hold true for the other models ;namely Single-Input Single-Output(SISO), Single-Input Multiple-Output(SIMO) and Multiple-Input Single-Output(MISO) models; if it was shown to be true for the MIMO model. The blind source separation algorithm was discussed for three different cases:

- The speech signal is considered to be a non-stationary source, with a backward model being proposed for the instantaneous case and the convolutive case[20].
- The speech signal is considered to be cyclo-stationary and a solution is being derived for full identifiability and partial identifiability, for case when the source signals all have the same cyclic frequencies and the case when all the source signals have distinct cyclic frequencies[2].

- When the speech signal is considered to be a quasistationary signal, then the local dominance of a source signal at any given instant can be used to obtain a correlation metric which can be used to derive a solution for the blind separation[9].

An implementation of the case 3 was done and the output is shown in chapter 4. Although, the implementation was done for speech signals, there are other applications where the above algorithm would work well, like the seismic activity recognition by separating the sounds of interest.

Bibliography

[1] *Matrix Computations.* John Hopkins University Press, 1996.

[2] K. Abed-Meraim, Yong Xiang, J. H. Manton, and Yingbo Hua. Blind source-separation using second-order cyclostationary statistics. *IEEE Transactions on Signal Processing*, 49(4):694–701, Apr 2001.

[3] Benesty, Sondhi, Jacob Huang, Mohan M, and Yiteng. *Handbook of Speech Processing.* Springer, 2008.

[4] V. Capdevielle, C. Serviere, and J. L. Lacoume. Blind separation of wide-band sources in the frequency domain. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 2080–2083 vol.3, May 1995.

[5] J. F. Cardoso, A. Belouchrani, and B. Laheld. A new composite criterion for adaptive and iterative blind source separation. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, volume iv, pages IV/273–IV/276 vol.4, Apr 1994.

[6] J. F. Cardoso and B. H. Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030, Dec 1996.

[7] Scott C. Douglas. *Blind Separation of Acoustic Signals*, pages 355–380. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[8] F. Ehlers and H. G. Schuster. Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment. *IEEE Transactions on Signal Processing*, 45(10):2608–2612, Oct 1997.

[9] X. Fu and W. K. Ma. A simple closed-form solution for overdetermined blind separation of locally sparse quasi-stationary sources. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2409–2412, March 2012.

[10] X. Fu and W. K. Ma. Blind separation of convolutive mixtures of speech sources: Exploiting local sparsity. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4315–4319, May 2013.

[11] W. A. Gardner. Exploitation of spectral redundancy in cyclostationary signals. *IEEE Signal Processing Magazine*, 8(2):14–36, April 1991.

[12] Huang, Benesty, Yiteng Chen, Jacob, and Jingdong. *Acoustic MIMO signal Processing.* Springer, 2006.

[13] Christian Jutten. *Calcul neuromimetique et traitment du signal: Analyze en composantes independantes.* PhD thesis, INP-USM Grenoble, 1987.

[14] Steven M Kay. *Fundamentals of Statistical Signal Processing.* Prentice Hall, 1993.

[15] Te-Won Lee, Anthony J. Bell, and Russel H. Lambert. Blind separation of delayed and convolved sources. In *Proceedings of the Conference on Neural Information Processing Systems, NIPS 1996*, Denver, Colorado, USA, December 2-5 1996.

[16] Albert Leon-Garcia. *Probability, Statistics and Random Processes for Electrical Engineering.* 2008.

[17] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72:3634–3637, Jun 1994.

[18] Noboru Murata, Shiro Ikeda, and Andreas Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1):1 – 24, 2001.

[19] Linh-Trung Nguyen, A. Belouchrani, K. Abed-Meraim, and B. Boashash. Separating more sources than sensors using time-frequency distributions. In *Proceedings of the Sixth International Symposium on Signal Processing and its Applications (Cat.No.01EX467)*, volume 2, pages 583–586 vol.2, 2001.

[20] L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8(3):320–327, May 2000.

[21] Saeed V Vaseghi. *Multimedia Signal Processing.* Wiley, 2007.

[22] Yong Xiang, K. Abed-Meraim, and Yingbo Hua. Adaptive blind source separation by second order statistics and natural gradient. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, volume 5, pages 2917–2920 vol.5, 1999.