Speech Enhancement using Constrained-ICA with Bessel Features

Balakrishna Lavu
Venkata Ashokkumar Potnuru

This thesis is presented as part of Degree of Master of Science in Electrical Engineering

Blekinge Institute of Technology
Thesis Series No: MEE10:114
January 2011
Abstract

This thesis explores the method of using Bessel features to constrained Independent Component Analysis (cICA) algorithm to extract a subset of desired independent source signals from a set of mixture of source signals.

In our thesis the cICA is extended to use the Bessel coefficients of the observed signals and the reference signal as they converge faster than other transformations. To overcome the problem of designing the reference signal from the desired source signal we used a different speech signal generated by the same physical source.

Simulation results and comparison of the results with a few pre-existing methods the proposed method shows the usefulness in the applications like speaker identification, voice recognition applications etc.
Acknowledgement

We would like to thank our supervisor Dr. Suryakanth V. Gangashetty for his extensive support, encouragement and guidance to us in a right way during the thesis work. He provided all necessary resources and guidance during the thesis which helped us a lot to complete the thesis successfully.

We would like to thank our examiner Dr. Nedelko.Grbic for giving us the chance to work on such a wonderful topic and for his extended support, guidance during the thesis work.

We would like to thank all members of Speech and Vision lab, IIIT, HYDERABAD for their support and help during thesis.

We would like to further thank our parents and relatives for their love and belief, and special thanks to our friends who were always there for us.
# Contents

1. **Introduction**  
   1.1 Problem statement ............................................. 6  
   1.2 Aim of the thesis work ........................................ 6  
   1.3 Thesis outline .................................................. 7  

2. **Literature Review**  
   2.1 Fundamentals of Speech Processing  
      2.1.1 Speech Communication .................................... 8  
      2.1.2 Speech Production System  
          2.1.3 Different types of Speech ................................ 11  
   2.2 Independent Component Analysis (ICA)  
      2.2.1 ICA Algorithm ............................................. 12  
   2.3 Fourier Bessel Features  
      2.3.1 Bessel functions .......................................... 15  
      2.3.2 Orthogonality of Bessel Functions ......................... 17  
      2.3.3 Effect of Number of Terms on Fourier-Bessel Representation  
      2.4 Constrained Independent Component Analysis (cICA) .......... 20  
   2.5 The Convolutive BSS Problem  
   2.6 Empirical Mode decomposition (EMD)  
      2.6.1 Speech extraction using ICA-R with EMD based Reference  
   2.7 Ambiguities of ICA-R ......................................... 25  

3. **Problem Formulation**  
   3.1 Speech extraction using ICA-R with Fourier Bessel  
   3.2 Proposed Method ............................................... 26  
   3.3 Proposed Algorithm ............................................ 27  

4. **Implementation and Simulation**  
   4.1 Data Acquisition ............................................... 28  
   4.2 SNR and Mean Square Error Analysis  
   4.3 Test Scenarios .................................................. 29  
      4.3.1 Instantaneous Mixing ...................................... 29  
      4.3.2 Convolutive Mixing ........................................ 31  
      4.3.3 Speech Signals along with Noise Signals ................. 35  

5. **Summary and Conclusion**  
   .......... 40
Chapter 1

Introduction

Independent Component Analysis (ICA) belongs to a class of Blind Signal Separation (BSS) methods for separating data into underlying informational components, where such data can take the form of images, sounds etc. As the name suggests, ICA separates a set of signal mixtures into a corresponding set of statistically independent component signals or source signals. Here the property of being independent is of fundamental importance, because it can be used to separate a mixture of source signals.

The defining feature of the extracted signals is that each extracted signal is statistically independent of all other extracted signals i.e. each extracted signal will be generated by a different physical process [2].

Compared to the conventional ICA algorithm, the technique of the cICA also called as ICA with Reference (ICA-R) provides a general framework to incorporate these additional requirements or prior information, e.g., statistical properties or rough templates of the sources. With prior information, the cICA algorithm usually has a better performance. In the algorithm, the norm between the references and the estimated signals is used as the inequality constraints in the contrast function to extract the desired signals.

1.1 Problem statement

The cICA algorithm extracts the desired signals by incorporating some prior information into the separation process. The crucial problem is the design of reference signal in advance which is close to the desired signal when the desired source signal is very weak in the mixed signals and also when there is no prior information about the desired source signal.

1.2 Aim of the thesis work

The main aim of developing this thesis is to overcome the problem of designing the reference signal when there is no prior information available of the desired source signal. The main objectives are listed below:

- To overcome the problem when there is no prior information available about the desired signal we propose to use a different speech signal generated by the same physical source as the reference signal for the algorithm.
- The ICA-R algorithm is extended to use Bessel coefficients of the observed signals and the reference signal as they are more efficient in representing
speech-like waveform and also converges faster than the other transformations [2].

- The final objective is to analyze the performance of the proposed method on different mixing scenarios like instantaneous, random, convolutive mixing etc...

### 1.3 Thesis outline

In chapter 2, we describe the introduction to speech processing, Convolutive mixtures, Independent component analysis (ICA), constrained ICA (cICA) and Bessel expansion of signals. Chapter 3 describes the speech extraction using ICA-R algorithm with Bessel features. In chapter 4, we discuss the initial measurements of the ICA-R algorithm and results. In chapter 5 we provide summary and conclusions.
Chapter 2

Literature Review

2.1 Fundamentals of Speech Processing

2.1.1 Speech Communication

Speech is a medium of communication between a speaker and a listener. Human speech production is originated with a message that the speaker wants to convey to a listener.

In the process of speech production, the speaker conveys a thought through a series of neurological processes and muscular movements to produce an acoustic sound pressure wave that is received by a listener's auditory system, processed, and converted back to neurological signals. For this, a speaker first forms a message to convey and then convert that message into a linguistic structure by choosing appropriate words or phrases based on learned grammatical rules associated with the particular language, and finally adds any additional local or global characteristics such as pitch intonation or stress to emphasize aspects important for overall meaning. After this, the human brain produces a sequence of motor commands that move the various muscles of the vocal system, produce the desired sound pressure wave. This acoustic wave is received by the talker's auditory system and converted back to a sequence of neurological pulses that provide necessary feedback for proper speech production. This allows the talker to continuously monitor and control the vocal organ by receiving his or her own speech as feedback. Any delay in this feedback to our own ears can also cause difficulty in proper speech production. The acoustic wave is also transmitted through a medium, which is normally air, to a listener's auditory system.

The speech perception process begins when the listener collects the sound pressure wave at the outer ear, converts this into neurological pulses at the middle and inner ear, and interprets these pulses in the auditory cortex of the brain to determine what message was received.

We can see that in both production and perception, the human auditory system plays an important role in the ability to communicate effectively [1].
2.1.2 Speech Production System

The speech waveform is an acoustic sound pressure wave that originates from voluntary movements of anatomical structures which make up the human speech production system. In Figure 2.1 portrays a medium sagittal section of the speech system in which we view the anatomy, midway through the upper torso as we look on from the right side. The gross component of the system is the lungs, trachea (windpipe), vocal folds, epiglottis, oral cavity (mouth), and nasal cavity (nose).

Figure 2.1: A schematic diagram of the human speech production mechanism [1].
Technically the pharyngeal and oral cavities are usually grouped into one unit referred to as the vocal tract, and the nasal cavity is often called the nasal tract. The vocal tract begins at the output of the larynx, and terminates at the input to the lips. The nasal tract begins at the velum and ends at the nostrils of the nose. Finer anatomical features critical to speech production include the vocal folds or vocal cords, soft palate or velum, tongue, teeth and lips. The soft tip of the velum, which may be seen to hang down in the back of the oral cavity when the mouth is wide open, is called the uvula. These finer anatomical components move to different positions to produce various speech sounds and are known as articulator's by the speech scientists. The mandible or jaw is also considered to be an articulator, since it is responsible for both gross and fine movements that affect the size and shape of the vocal tract as well as the position of the other articulators. The three main cavities of the speech production system (vocal plus nasal tracts) comprise the main acoustic filter. The filter is excited by organs below it and is loaded its main output by radiation impedance due to the lips. The articulators, these are associated with filter itself. Figure 2.2 shows the block diagram of human speech production [1].

The acoustic filter of the system is contributed to the resonant structure of human speech. The average adult male and female, length of the vocal tract is 17cm and 14cm. The average vocal tract length of a child is 10cm. The nasal tract constitutes an auxiliary path for the transmission of sound. The average length of the nasal tract for adult male is 12cm. Acoustic coupling between the nasal and vocal tracts is controlled by the size of the opening at the velum. The nasal coupling can substantially influence the frequency characteristics of the sound radiated from the mouth. If the velum is lowered, the nasal tract is acoustically coupled to produce the nasal sounds of the speech. The average adult male velar opening can be from 0 to 5cm. In the production of non-nasal sounds, the velum is drawn up tightly towards the back of the pharyngeal cavity, effectively sealing off the entrance to the nasal cavity and decoupling it from the speech production system. The larynx has a simple, but highly significant, role in speech production. Its function is to provide a periodic excitation to the system for speech sounds. That is often known as voiced speech/sound. The periodic vibration of the vocal folds is responsible for this voicing. The larynx is an intricate and complex organ that has been studied extensively by anatomists and physiologists [1].
2.1.3 Different Types of Speech

The principle feature of the any speech sound is the manner of excitation.

Excitations are of following types:

(1) Voiced
(2) Unvoiced
(3) Mixed
(4) Plosive
(5) Whisper
(6) Silence
Last four are the combination of the voiced, unvoiced and silence. During the excitation of a particular speech sound or class of sounds one or more of these excitation types get blended.

The production of voiced sound is made by forcing air through the glottis or an opening between the vocal folds. The vocal cords vibrate in an oscillatory fashion. Voice or phonation is the sound produced by larynx. The term voice is frequently used which means ‘speech’.

Unvoiced sounds are generated by forming a constriction at some point along the vocal tract, and forcing air through the constriction to produce turbulence. The combination of voiced and unvoiced sound is denoted as Mixed.

Some speech sounds are composed of a short region of silence, followed by a region of voice speech, unvoiced speech, or both, which are called plosive sounds. These sounds are formed by making a complete closure, building up air pressure behind the closure, and suddenly releasing it. ‘Silence’ is one type of excitation, whatever the source of excitation, the vocal tract act as a filter [1].

2.2 Independent Component Analysis (ICA)

Independent component analysis aims to recover a set of unknown mutually independent sources signals from there linear mixtures are observed without knowing the mixing coefficients[2], [3]. We assume that you are in room where the two speakers are speaking simultaneously and we have two microphones. These are placed in different locations. The microphones give you the recorded time signals, $x_1(t)$ and $x_2(t)$, where $t$ is time index. These recorded signals are a weighted sum of the speech signals emitted by the two speakers $s_1(t)$ and $s_2(t)$ given by:

$$x_1(t) = a_{11}s_1 + a_{12}s_2$$

$$x_2(t) = a_{21}s_1 + a_{22}s_2$$

(1) (2)

Where $a_{11}$, $a_{12}$, $a_{21}$ and $a_{22}$ are parameters that depend on the distance of microphones from the speakers. It would be very useful if you can estimate the two original speech signals $s_1(t)$ and $s_2(t)$ by using recorded signals $x_1(t)$ and $x_2(t)$. This is called the Cocktail-Party problem. If we knew the parameters $a_{ij}$, we could solve the linear equation in (2) by classical methods. However since we don’t know the $a_{ij}$, these problems are challenging. Solving these problems recently developed technique of “Independent Component Analysis” can be used to estimate the $a_{ij}$ based on information of their independence, using this to separate the two original source signals from their mixtures [2], [3], [12].
ICA recovers a set of unknown mutually independent source signals from their observed linear mixtures. Assume that we observe \( n \) linear mixtures \( x_1, \ldots, x_n \) of \( n \) independent components. Suppose \( M \) independent source signals are \( s(t) \), and \( N \) observed mixture of source signals are \( x(t) \), then

\[
\begin{align*}
    s(t) &= [s_1(t) \ldots s_M(t)]^T \\
    x(t) &= [x_1(t) \ldots x_N(t)]^T
\end{align*}
\] (3) (4)

The linear ICA assumes that these mixtures are linear, instantaneous, and noise less. The vector –matrix notation is given by:

\[
x = As(t)
\] (5)

The columns of matrix \( A \); denoting them by \( a_i \) the model can also be written as

\[
x = \sum_{i=1}^{n} a_i s_i
\] (6)

ICA is a very simple assumption that the components \( s_i \) are statistically independent and also assume that the independent components must have non-Gaussian distributions. In the basic model we do not assume these distributions. In the basic model (6), \( A \) is a \( M \times N \) mixing matrix that contains the mixing coefficients. The goal of ICA is to find to a \( N \times M \) de mixing matrix \( W \) such that \( M \) output signals,

\[
s = Wx(t)
\] (7)

ICA is very closely related to the method called Blind Source Separation (BSS) or Blind Signal Separation. Without knowing information about the original signal, we are separating the mixed signals.

Pre-processing for ICA:

Before we are applying data to ICA, it is usually very useful to do some pre-processing. The pre-processing techniques are centring and whitening, that make the problem of ICA estimation simpler and better conditioned [2].

Centring:

The necessary pre-processing is to centre the mixed signal \( x \). Centring means, subtract it mean vector \( m = E\{x\} \) from \( x \), so as to make \( x \) a zero mean variable. This pre-processing is made solely to simplify the ICA algorithms: It does not mean that the mean could not be estimated. After estimating the mixing matrix \( A \) with centred data, we can complete the estimation by adding the mean vector of \( s \) back to centred estimates of \( s \). The mean vector is given by \( A^{-1}m \), where \( m \) is the mean that was subtracted in the pre-processing [2], [3].
Whitening:

Another pre-processing strategy in ICA is to first whiten the observed variables. After the centring the data and before the application of the ICA algorithm, we transform the observed vector $\mathbf{x}$ linearly so that we obtain a new vector $\tilde{\mathbf{x}}$ which is whitened. Its components are uncorrelated and their variances are equal to unity. The covariance matrix $\tilde{\mathbf{x}}$ equals the identity matrix.

$$E\{\tilde{\mathbf{x}} \tilde{\mathbf{x}}^T\} = I$$

The whitening transformation is always possible. One most important method for whitening is to use the Eigen-value decomposition (EVD) of the covariance matrix $E\{\tilde{\mathbf{x}} \tilde{\mathbf{x}}^T\} = \mathbf{EDE}^T$, where $\mathbf{E}$ is the orthogonal matrix of eigenvectors of $E\{\tilde{\mathbf{x}} \tilde{\mathbf{x}}^T\}$ and $\mathbf{D}$ is the diagonal matrix of its Eigen-values. Whitening transforms the mixing matrix into a new one $\tilde{\mathbf{A}}$.

$$\tilde{\mathbf{x}} = \mathbf{ED}^{-1/2} \mathbf{E}^T \mathbf{A} \mathbf{s} = \tilde{\mathbf{A}} \mathbf{s}$$

The whitening reduces the number of parameters to be estimated. Instead of having to estimate $n^2$ parameter that is the elements of the original matrix $\mathbf{A}$, we need to estimate orthogonal mixing matrix $\tilde{\mathbf{A}}$. Whitening is very simple and standard procedure for any ICA algorithm; it is a good idea to reduce the complexity [2], [3].

### 2.2.1 ICA algorithm:

Given $N$ mixed signals $\mathbf{x}(t) = [\mathbf{x}_1(t) \ldots \mathbf{x}_N(t)]^T$ of length $L$.

1. Initiate $\mathbf{W}$ to the identity matrix ($N$ by $N$), set $n = 1$

2. Calculate the output $\mathbf{s}(n)$: $\mathbf{s}(n) = \mathbf{Wu}(n)$

3. Update $\mathbf{W}$ according to above algorithm

4. $\mathbf{W}^{(n+1)} = \mathbf{W}^n + \alpha [\mathbf{I} - \tanh(\mathbf{s}) \mathbf{s}^T] \mathbf{W}^n$

   $n = n + 1$ go to 2 until $n = L$ then stop

Separated signals are $\mathbf{s}(t) = [\mathbf{s}_1(t) \ldots \mathbf{s}_N(t)]^T$
2.3 Fourier Bessel Features

The practical signals that we encounter in our daily lives, like speech and images are highly redundant. In the conventional way of representing a signal by its amplitude verses time plots, such redundancies are not very evident. Also, such a representation is not suited for the purpose of pattern classification and recognition, since it is computationally infeasible to deal with signal in its natural form. Therefore, ever since the advent of digital communications and pattern recognition by the machines, a constant effort of the Electrical and Computer Engineers has been the task of finding transforms that can exploit the redundancy of the practically encountered signals, and express them in as few vectors as possible. Fourier Transforms have been widely used for transforming the signals into the frequency domain, primarily due to the ease of their computation (through the FFT Algorithm), and the reasonably good performance that they yield in data compression, and pattern recognition. In the field of speech processing, frequency domain coding schemes have employed the Fourier Transform for Speech Compression and speech enhancement.

The presented work “Speech Enhancement using Constrained ICA with Bessel features” aimed at exploring the applications of Fourier-Bessel coefficients in the field of Speech enhancement. The Bessel functions arise as solutions to wave equations inside cylindrical tubes includes the first kind of Bessel functions. Since the vocal tract can be approximated as organ pipe-like cylindrical tubes with a sound source at one end (the larynx or voice box) and open at other end (lips and nose), it appears reasonable to use the Bessel functions for analyzing human speech [14], [17]. Moreover, the speech signal has regular zero-crossing and decaying amplitudes. The Bessel functions of the 1st kind, $J_n(x)$, has regular zero-crossing and decaying sinusoids shown in Figure 2.3. Therefore, they exhibit quasi-periodicity unlike the pure sinusoids. Thus, there is another very good motivation to use them for analyzing speech signals, given that they are structurally more similar to speech signals than the pure sinusoids [15-19]. This work also aims at exploring the use of Fourier-Bessel Coefficients for speech enhancement.

2.3.1. Bessel Functions

Bessel Functions arise as solutions to the following differential equation

$$x^2 y'' + x y' + (x^2 - n^2) y = 0, \quad n \geq 0 \quad (10)$$

A general solution of (10) is given by the following expression

$$y = C_1 J_n(x) + C_2 Y_n(x), \quad n \geq 0 \quad (11)$$

In (11), $J_n(x)$ is called Bessel Function of the first kind of order $n$, and $Y_n(x)$ is called Bessel function of the second kind. The term $J_n(x)$ is given by [17].
\[ J_n(x) = \sum_{l=0}^{\infty} \frac{(-1)^l}{2^{2l+m} l!(m+l)!} x^{2l+m} \] (12)

To appreciate the usefulness of \( J_n(x) \) in representing the Bessel basis function, we need to look at the plots of \( J_n(x) \) for different values of \( n \) in Figure 2.3. As evident in the figure, the plot of \( J_n(x) \) is decayed amplitudes (damped sinusoid). We can approximate this behavior as being quasi-periodic, in the sense that for a small interval, the damped sinusoid can be approximated as a sinusoid. A voiced speech utterance’s waveform is shown in Figure 2.4. Thus, a structural similarity between \( J_n(x) \) and \( s(t) \) is a good motivation behind using the Fourier-Bessel decomposition of speech signals.

![Figure 2.3: Bessel Function of the first kind, \( J_n(x) \) for values of \( n=0, 1, \) and 2](image)
Figure 2.4: Waveform of a sample speech utterance. Observe the evident quasi-periodicity.

### 2.3.2 Orthogonality of Bessel Functions

It can be shown that the Bessel functions are orthogonal with respect to \( x \) [15-18]. We can write

\[
\int_0^1 x J_n(\alpha x) J_n(\beta x) \, dx = \frac{\beta J_n(\alpha)J_n(\beta) - \alpha J_n(\beta)J_n(\alpha)}{\alpha^2 - \beta^2}, \quad \alpha \neq \beta
\]  \hspace{1cm} (13)

Also,

\[
\int_0^1 x J_n^2(\alpha x) \, dx = \frac{1}{2} \left[ J_n^2(\alpha) + \left( 1 - \frac{n^2}{\sigma^2} \right) J_n^2(\alpha) \right]
\]  \hspace{1cm} (14)
If $\alpha$ and $\beta$ are different roots of $J_n(x) = 0$, it can be written that

$$\int_0^1 x J_n(\alpha x) J_n(\beta x) \, dx = 0, \quad \alpha \neq \beta$$

(15)

Thus $J_n(\alpha x)$ and $J_n(\beta x)$ are orthogonal with respect to $x$. Any arbitrary function $f(x)$ can be written in terms of Bessel functions with the form

$$f(x) = \sum_{m=1}^{\infty} C_m J_n(\lambda_m x)$$

(16)

In (16) $\lambda_1, \lambda_2 \ldots$ are the positive roots of $J_n(x) = 0$. The coefficients, $C_m$ are given by

$$C_m = \frac{2}{J''_{n+1}(\lambda_m)} \int_0^1 x f_n(\lambda_m x) f(x) \, dx$$

(17)

We can expand a given function $f(t)$ over some arbitrary interval $(0,a)$ using the zero order Bessel series expansion [14], [16], [17] as

$$f(t) = \sum_{m=1}^{\infty} C_m J_0(\lambda_m t), \quad 0 < t < a$$

(18)

The coefficients $C_m$ appearing in (17) are calculated using the following expression [16-18], and have been generally referred to in the literature as Fourier-Bessel (FB) coefficients [14], [16], [17].

$$C_m = \frac{2 \int_0^a t f(t) J_0(\lambda_m t) \, dt}{a^2 [J_1(\lambda_m a)]^2}$$

(19)

The terms $\lambda_m, m = 1, 2, \ldots$ are the ascending order positive roots of $J_0(a \lambda_m) = 0$. Further discussion on the computational aspects of using (19) can be found in [15]. Figure 2.5(a) shows a speech signal $s(t)$ Figure 2.5(b) shows its Fourier-Bessel coefficients, while Figure 2.5(c) shows the speech signal reconstructed back from the Fourier-Bessel Coefficients. As observed from Figure 2.5 a speech utterance can be accurately reconstructed back from its Fourier-Bessel coefficients. Another point to be made here is that for practical purposes, we do not need an infinite number of terms in (19). In Figure 2.5(b), for example, a good reconstruction has been achieved by using the first eight thousand terms. In general, when using a frame size of $x$ samples, it is sufficient to use $m=x$. 

18
2.3.3 Effect of Number of Terms on Fourier-Bessel Representation

To appreciate the effect ‘m’ on the Fourier-Bessel representation of a given signal as given in (29), Figure 2.6 shows the Bessel basis functions for a signal for varying values of m. An immediate consequence of these waveforms is that as the value of m increases, the frequency of the Basis function increase. This, in turn, means that a Basis function with a higher value of m captures the higher frequency content of a signal. Thus, it becomes necessary to keep the value of m high enough, if we desire to capture the high frequency content of a signal.

Also, by using selective values of m, it is possible to extract certain specific frequencies and leave the rest. This has been used in works like speaker identification, speaker separation, and speech compression. At this point, it should be mentioned that
a robust algorithm to selectively pick up desired values of m for speech compression or speech separation is desirable.

![Figure 2.6: Zero-order Bessel basis functions for different values of 'm']

**2.4 Constrained Independent Component Analysis**

The ICA – R also called as constrained ICA (cICA) is good for extracting several source signals from a large number of observed mixtures by providing a reference signal which should be closely related to the desired source signal [4,5,6,7]. The adaptive solutions using the Newton-like learning rule is used to solve the constrained optimization problem. The cICA is used to provide a systematic and flexible method to incorporate more assumptions and prior information into contrast function so the ill-posed ICA is converted to one with better condition number. Constraints are adopted to reduce the dimensionality of the output of the ICA. ICA-R is very useful in many applications as Automatic Speech Recognition and speaker identification etc.

Instead of separating all M number of independent sources from N mixed signals, ICA-R extracts L (L<M) number of desired sources from N mixed signals by incorporating some prior information about target signal into the ICA learning algorithm as reference signals [5]. These reference signals denoted by \( r(t) = [r_1(t), \ldots, r_L(t)]^T \), carry some information of the desire sources but not identical to corresponding desired signals. It finds one weight vector \( W \), so that the output signal,
$y(t) = w^T x(t)$ recovers desired source $s^*(t)$ by using $r(t)$ as the reference signal [4, 5]. Where $w$ is a column vector and $y(t)$ recovered source signal up to a scalar. Before running the algorithm the whitening is often used to transform the observe signals $x(t)$ to make algorithm simpler and faster.

$$z(t) = Vx(t)$$

Such that $\{z(t)z(t)^T\} = I$, where $V$ is the whitening matrix $A$ typical solution is given by

$$V = D^{-1/2}E^T$$

Where $D$ is the diagonal matrix of the Eigen values of the matrix $E \{x(t)x(t)^T\}$, $E$ is the matrix whose columns are the corresponding unit–norm eigenvector.

The cICA algorithm extracts the one independent source signal, a reliable and a flexible contrast function is the one based on negentropy [17], given by:

$$J(y) = \rho [E\{G(y)\} - E\{G(v)\}]^2$$

Here $\rho$ is a positive constant, $v$ is a Gaussian variable having zero mean and unit variance, $G(.)$ can be any non-quadratic function. The closeness between the ICA-R output $y$ and the reference signal $r$ is measured by $\varepsilon(y, r)$, which has a minimal value when $y = PDS^*$. A threshold $\xi$ is used to distinguish the desired source $s^*$ from other source signals such that $g(w) = \varepsilon(w^TX, r) - \xi \leq 0$ is satisfied only when $y = PDS^*$. Among all source signals, $g(w)$ is a feasible constraint to the contrast function, the problem of one unit ICA-R can be modeled in the framework of constrained independent component analysis [3], [13].

Maximize $J(y) = \rho [E\{G(y)\} - E\{G(v)\}]^2$

Subject to $g(w) = \varepsilon(y, r) - \xi \leq 0$, $h(w) = E\{y^2\}-1=0$

Where $h(w)$ is the equality constraint used to ensure the contrast function $J(y)$ and the weight vector $w$ are bounded. $\varepsilon(y, r)$ is the closeness between the output signal and reference signal $r$. The desired source signal is always closest to the reference signal $r$. In a Newton-like leaning algorithm is derived by finding the maximum of an augmented Lagrangian function corresponding [3], [5]. The updated weights are given by:

$$w_{k+1} = w_k - \eta \frac{R^{-1}L_{w_k}}{\delta(w_k)}$$
Where \( k \) is iteration index, \( \eta \) is the learning rate, \( R_{xx} \) is the covariance matrix of the input mixtures \( x \), also

\[
L_{w_k} = \rho E\{xG_y'(y)\} - 0.5 \mu E\{xg_y'(w_k)\} 
\]

(26)

\[
\delta(w_k) = \rho E\{xG_{y^2}(y)\} - 0.5 \mu E\{xg_{y^2}(w_k)\} 
\]

(27)

Where \( G_y'(y) \) and \( G_{y^2}(y) \) are the first and second derivatives of \( G(y) \) with respect to \( Y \), and \( g_y'(w_k) \) and \( g_{y^2}(w_k) \) are those of \( g(w_k) \). The optimum multipliers \( \mu \) and \( \lambda \) are found by iteratively updating them based on a gradient-ascent method. Here \( \gamma \) is the scalar penalty parameter. Further,

\[
\mu_{k+1} = \max\{0, \mu_k + \gamma g(w_k)\} 
\]

(28)

\[
\lambda_{k+1} = \lambda_k + \gamma h(w_k). 
\]

(29)

### 2.5 The Convolutive BSS Problem

In practical environments, instantaneous mixtures of audio signals are very rare to find. These problems are not realistic at all. For example, we have two sources \( s_1(t) \) and \( s_2(t) \), recorded by two microphones in a room. In real world recording sounds are delayed before to reach the microphones and they are convolved with room responses such as reflection from the room walls (reverberation). These are all considered in convolved mixtures, where the mixing filters are weighted delays [8], [9], [10], [19]. The sources have been mixed, convolved and delayed according to the equation,

\[
x_i(t) = \sum_{j=1}^{N} \sum_{k=0}^{M} s_j(t-k)a_{ij}(k) 
\]

(30)

Where \( N \) is the number of the sources \( s_j \), \( a_{ij} \) is the \( M \) length mixing filter coefficients. These mixed source signals are using as input to the algorithm. Figure 2.7 shows the block diagram of the convolved mixtures for four sources. This is an FIR model, where each FIR filter (for fixed indices \( i \) and \( j \)) is defined by the coefficients \( a_{ij} \). With Convolutive mixtures the indeterminacies are more severe.
2.6 Empirical Mode Decomposition (EMD)

EMD is a nonlinear, non-stationary signal processing method. The main idea of EMD is
decompose a signal into a sum of oscillatory functions, called intrinsic mode functions
(IMFs). Any IMF satisfies two conditions 1) having same number of extrema and zero-
crossings or differ at most by one. 2) The mean value of envelope defined by the local
maxima and local minima is zero. With these two requirements, the meaningfully
instantaneous frequency of an IMF can be well defined. The first condition is similar to
narrow band requirements for a stationary Gaussian process. Second one is a local
requirement induced from global one, and necessary to ensure that the instantaneous
frequency will not have redundant fluctuations as induced by asymmetric waveforms.
The EMD algorithm to extract the residual $r(t)$ of a signal $x(t)$ is as follows [14], [18].

(1) Extract the local maxima and minima of the signal $x(t)$
(2) Interpolate the local maxima and local minima with cubic spline to form upper
    envelope and lower envelope of $x(t)$
(3) Calculate the point-by-point mean $m(t)$ from upper and lower envelopes.
(4) Extract the detail, $d(t)=x(t)-m(t)$. Check the properties of $d(t)$:
(5) If $d(t)$ do not follow above two conditions replace $x(t)$ with $d(t)$. Repeat the steps
    1 to 5 until it satisfies the stopping criteria.
(6) If \( d(t) \) meets the above two conditions, an IMF is derived and replace \( x(t) \) with the residual \( r(t)=x(t)-d(t) \). Repeat the above steps until \( r(t) \) hast at least two extrema, else the decomposition is finished.

End of the process \( x(t) \) can expressed as follows

\[
x(t) = \sum_{j=1}^{N} c_j(t) + r_N(t)
\]  

Where \( N \) is the number of IMFs \( r_N(t) \) denotes the final residual which can be interpreted as the DC component of the signal, and \( c_j(t) \) are nearly orthogonal to each other, all are zero mean. The signal is decomposed into \( N \) IMFs, each with distinct time scale as per results; the first IMF has the smallest time scale which corresponds to the fastest time variations signal. The decomposition process proceeds, time scale increase, and mean frequency mode decreases.

### 2.6.1 Speech extraction using ICA-R with EMD based Reference

In this approach the ICA-R algorithm extracts the target speech signal from the mixture of different speech signals by constructing the reference signal using Empirical Mode Decomposition (EMD) [14]. As mentioned above, the speech power spectrum is largely different from those of its environmental noises in that speech power is regularly distributed within several given frequency bins. Therefore, the approximate envelope of the power spectrum of the desired speech can be used as a reference signal, and can be well obtained by EMD. Here the speech signal is significantly enhanced by including prior information, i.e. the knowledge of the speech power spectra into the ICA separation process.

Here we calculate the IMF’s of the reference signal. As an example we got ten IMF components \( C_1, C_2...C_{10} \). The mean frequency gradually decreases from \( C_1 \) to \( C_{10} \). Thus by summing IMF’s with lower frequency, a rough envelope of the speech power spectrum can be formed. The obtained signal is passed as a reference signal to the ICA-R algorithm as described in section 3.2 to extract the desired signal from a set of observed input signals.
2.7 Ambiguities of ICA-R:

There are several important issues that should be noticed [7]. The major is choosing the value for the threshold $\xi$. When the threshold value is low the ICA-R algorithm will not converge and if the value is too large, the ICA-R may converge to other source signals, because of some source signals whose closeness measure may be less than the threshold. So the threshold value should be carefully selected.

The other issue is the design or choosing the reference signal which affects the threshold value. If the reference signal is very similar to the desired source signal, then the threshold value should be very small so that the algorithm can globally converge. Otherwise, the value should be large.
Chapter 3

Problem formulation

3.1 Speech extraction using ICA-R with Fourier Bessel

In this method we use filtered mixtures. Generally, in blind source separation problem, there are N unknown source signals \( s(t) = [s_1(t) \ldots \ldots s_M(t)]^T \). These signals are collected from N sensor \( x(t) = [x_1(t) \ldots \ldots x_M(t)]^T \). These signals are mixed, delayed and convolved according to

\[
x_i(t) = \sum_{j=1}^{N} \sum_{n=0}^{\infty} a_{ij}(n) s_j(t - D_{ij} - n) \quad (i = 1, \ldots, N)
\]  

(32)

Where \( s_j(t - D_{ij} - n) \) is the channel impulse response from \( j^{th} \) source to \( i^{th} \) sensor and \( D_{ij} \) is the delay between the \( j^{th} \) sources to \( i^{th} \) sensors and \( a_{ij} \) are the filter coefficients. Convolution in the time domain corresponds to instantaneous mixing in the frequency domain [8].

The goal of ICA-R is to compute the filter coefficients \( a_{ij} \) such that the desired speech signal is extracted from a set of observed signals by providing a suitable reference signal which is closely related to the reference signal.

3.2 Proposed method:

- We used convolved and instantaneous mixtures for mixing the input signals or sources. Here, number of sources and number of sensors are equal
- Calculated the Bessel coefficients of the mixed speech signals and reference signal by framing with 50% overlap.
- The calculated Bessel coefficients of the reference signal and set of mixed speech signals are passed as the inputs to the ICA-R algorithm for processing.
- The ICA-R algorithm extracts the Bessel coefficients of the desired signal from the set of Bessel coefficients of the observed signals.
- Re-synthesize the signal by performing inverse Bessel transformation and overlap-add method to get the desired speech signal.
In the ICA-R algorithm we use $G(y) = \log(\cosh(y))$, which is good general purpose function [12]. The closeness between the ICA-R output signal $y$ and the corresponding reference signal $r$ is defined as the mean square error and is given by:

$$\varepsilon(y, r) = E\{(y - r)^2\}$$

(33)

In ICA-R algorithm, $\mu, \gamma, \lambda$ are critical to the convergence of the algorithm of ICA-R [4, 5, 6, and 7]. It may be initialized with a small value to avoid the algorithm going to local optimum, and then is gradually increased to converge at the global maximum [3], [13].

### 3.2 Proposed Algorithm:

Given $N$ mixed signals $x(t) = [x_1(t) \ldots x_N(t)]^T$ of length $L$.

1. Initiate $W$ to random weight vector ($N$ by 1). set $n = 1$

2. Calculate the Bessel coefficients of the mixed signals and reference signals

$$C_m = \frac{2 \int_0^a t f(t) J_0(\lambda_m t) dt}{a^2 [J_1(\lambda_m a)]^2}$$

3. Calculate the output $s(n)$: $s(n) = Wu(n)$, $u(n)$ is the matrix containing the Bessel coefficients of the observed signals at each microphone.

4. Update $W$ according to the algorithm listed below

$$W^{(n+1)} = W^n - \eta \frac{R_x}{\delta(n)}$$

5. Update the learning rate $\eta$ and optimum multipliers $\mu$ and $\lambda$ from Eq.(28) and Eq.(29). Increment the iteration index, $n$ and go to 3 until up to algorithm converge as described in the section 2.4.

6. Re-synthesize the signal, $s(n)$ by performing the inverse Bessel transformation to get the desired output signal $y(n)$. 


Chapter 4

Implementation and Simulation

In this chapter an evaluation of the proposed method is performed. In this section we also compared our results with one of the previous existing methods to evaluate the performance of the proposed method. Here we have compared the ICA-R with Bessel and ICA-R using EMD in different scenarios like first with instantaneous mixing, second Convolutive mixing and third speech signals with additive noise.

4.1 Data Acquisition

We have considered four speech signals with zero mean as source signals each of different lengths. Zeros were appended to the four signals to equal the max of the length among the four signals. Now each source signal has 88320 samples and has sampling frequency of 16000 Hz. The four equal length signals were mixed randomly using a mixing matrix $A$ to obtain a set of mixed speech signals as shown in Figure 4.2. Our goal was to extract the target speech signal $s_3$ as shown in Figure 4.4(a) using the ICA-R algorithm by the proposed approach by providing the reference signal $r$, as shown in Figure 4.4(b).

As mentioned above the Bessel basis functions were calculated for the set of mixed signals and the reference signal by framing with 20 m.sec (320 samples) as frame size and 50% overlap and were passed as inputs to the ICA-R algorithm for processing.

4.2 SNR and Mean Square Error Analysis

The accuracy of the recovered signal $y(n)$ compared to the desired speech signal can be measured by the mean square error and the signal to noise ratio which is given by

$$SNR(dB) = 10\log_{10} \left( \frac{\sigma^2}{\text{mse}} \right)$$

(36)
4.3 Test Scenarios

4.3.1 Instantaneous Mixing

In this section we have taken four speech signals $s_1(t), s_2(t) ... s_4(t)$ with zero mean as shown in Figure 4.1 and a random Matrix $A$ of size $4 \times 4$ which represents a linear channel. An estimate of the observed signal set at microphones is calculated as $x(t) = As(t)$. In ICA-R using Bessel features, the Bessel coefficients of the observed signals in Figure 4.2 and the reference signal in Figure 4.3(b) are calculated and are passed as inputs to the ICA-R algorithm as described in section 3.2. In ICA-R using EMD the residual signal of the reference signal is extracted as described in section 2.6 and is passed as input to the ICA-R using EMD along with the observed signals in Figure 4.2.

The output of the ICA-R algorithm using EMD based reference and the Bessel Features are shown in Figure 4.3(c) and Figure 4.4(c). The Table 4.1 gives the Mean Square Error (MSE) and the SNR (dB) for the two methods.

![Figure 4.1: Four source signals (a) s1 (b) s2 (c) s3 (d) s4. The x-axis is time in m.sec and y-axis is amplitude](image-url)
Figure 4.2: Randomly mixed source signals (a) microphone1 (B) microphone2 (c) microphone3 (d) microphone4. The x-axis is time in m.sec and y-axis is amplitude.

Figure 4.3: ICA-R using EMD (a) Original/Desired speech signal. (b) Reference signal (different speech utterance of the target speech signal). (c) Extracted speech signal. The x-axis is time in m.sec and y-axis is amplitude.
Figure 4.4: ICA-R using Bessel Features (a) Original/Desired speech signal. (b) Reference signal (different speech utterance of the target signal). (c) Extracted speech signal. The x-axis is time in m.sec and y-axis is amplitude

Table 4.1: Comparison of ICA-R with EMD and ICA-R with Bessel Features using MSE and SNR as measures

<table>
<thead>
<tr>
<th>Performance Measures</th>
<th>Methods</th>
<th>EMD based Reference</th>
<th>Bessel Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>1.3400e-004</td>
<td>8.3400e-006</td>
<td></td>
</tr>
<tr>
<td>SNR(dB)</td>
<td>38.729</td>
<td>70.7883</td>
<td></td>
</tr>
</tbody>
</table>

4.3.2 Convolution Mixing

In real world applications, the physical properties of the propagation channel are often mathematically modeled as a convolution operation with the source signals and thus a linear time-invariant filter model may sometimes be more accurate, for example an IIR modeling of a room transfer function and its inverse filtering approach.
In this section we have taken four speech signals $s_1(t)$, $s_2(t)$...$s_4(t)$ show in Figure 4.5 with zero mean and a 16 FIR filters as a mixed phase mixing system, $A(z)$.

$$A = \begin{bmatrix}
A_{11} & A_{12} & A_{13} & A_{14} \\
A_{21} & A_{22} & A_{23} & A_{24} \\
A_{31} & A_{32} & A_{33} & A_{34} \\
A_{41} & A_{42} & A_{43} & A_{44}
\end{bmatrix}$$

Where

$$A_{11} = 0.025 + 0.23z^{-1} + 0.48z^{-2} + 0.23z^{-3} + 0.025z^{-4}$$
$$A_{12} = -0.04 + 0.16z^{-1} + 0.81z^{-2} + 0.16z^{-3} - 0.04z^{-4}$$
$$A_{13} = -0.04 - 0.16z^{-1} + 0.81z^{-2} - 0.16z^{-3} - 0.04z^{-4}$$
$$A_{14} = 0.025 - 0.23z^{-1} + 0.48z^{-2} - 0.23z^{-3} + 0.025z^{-4}$$
$$A_{21} = 0.033 + 0.24z^{-1} + 0.45z^{-2} + 0.24z^{-3} + 0.033z^{-4}$$
$$A_{22} = 0.003 + 0.25z^{-1} + 0.64z^{-2} + 0.25z^{-3} + 0.003z^{-4}$$
$$A_{23} = -0.05 - 0.08z^{-1} + 0.87z^{-2} - 0.08z^{-3} - 0.05z^{-4}$$
$$A_{24} = 0.025 - 0.23z^{-1} + 0.48z^{-2} - 0.23z^{-3} + 0.025z^{-4}$$
$$A_{31} = 0.023 + 0.24z^{-1} + 0.47z^{-2} + 0.24z^{-3} + 0.023z^{-4}$$
$$A_{32} = 0.002 + 0.25z^{-1} + 0.64z^{-2} + 0.25z^{-3} + 0.002z^{-4}$$
$$A_{33} = -0.056 + 0.09z^{-1} + 0.86z^{-2} + 0.09z^{-3} - 0.056z^{-4}$$
$$A_{34} = 0.006 - 0.22z^{-1} + 0.48z^{-2} - 0.22z^{-3} + 0.006z^{-4}$$
$$A_{41} = 0.013 + 0.22z^{-1} + 0.53z^{-2} + 0.22z^{-3} + 0.013z^{-4}$$
$$A_{42} = -0.06 + 0.007z^{-1} + 0.87z^{-2} + 0.007z^{-3} - 0.06z^{-4}$$
$$A_{43} = -0.002 - 0.24z^{-1} + 0.66z^{-2} - 0.24z^{-3} - 0.002z^{-4}$$
$$A_{44} = 0.033 - 0.24z^{-1} + 0.46z^{-2} - 0.24z^{-3} + 0.033z^{-4}$$
An estimate of the observed signal set at microphones is calculated as $x(t) = As(t)$ shown in Figure 2.7. In ICA-R using Bessel features, the Bessel coefficients of the observed signals in Figure 4.6 and the reference signal in Figure 4.7(b) are calculated and are passed as inputs to the ICA-R as described in section 3.2. In ICA-R using EMD the residual signal of the reference signal is extracted as described in section 2.6 and is passed as input to the ICA-R using EMD along with the observed signals in Figure 4.6.

Figure 4.5: Four source signals (a) s1 (b) s2 (c) s3 (d) s4. The x-axis is time in m.sec and y-axis is amplitude

The output of the ICA-R algorithm using EMD based reference and the Bessel Features are shown in Figure 4.7(c) and Figure 4.8(c). The Table 4.2 gives the Mean Square Error (MSE) and the SNR (dB) for the two methods.
Figure 4.6: Observed signals using convolutive mixing (a) microphone1 (b) microphone2 (c) microphone3 (d) microphone4. The x-axis is time in m.sec and y-axis is amplitude.

Figure 4.7: ICA-R using EMD (a) Original/Desired speech signal. (b) Reference signal (different speech utterance of the target speech signal). (c) Extracted speech signal. The x-axis is time in m.sec and y-axis is amplitude.
Figure 4.8: ICA-R using Bessel Features (a) Original/Desired speech signal. (b) Reference signal (different speech utterance of the target speech signal). (c) Extracted speech signal. The x-axis is time in m.sec and y-axis is amplitude.

Table 4.2: Comparison of ICA-R with EMD and ICA-R with Bessel Features using MSE and SNR as measures

<table>
<thead>
<tr>
<th>Performance Measures</th>
<th>Methods</th>
<th>EMD based Reference</th>
<th>Bessel Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td></td>
<td>0.9200e-004</td>
<td>1.2200e-005</td>
</tr>
<tr>
<td>SNR(dB)</td>
<td></td>
<td>28.624</td>
<td>64.862</td>
</tr>
</tbody>
</table>

4.3.3 Speech Signals along with Noise Signals

In this section we have taken two speech signals $s_1(t)$, $s_2(t)$ and a noise signal $n_1(t)$, $n_2(t)$ with zero mean ($S=[s_1,n_1,s_2,n_2]$) as shown in Figure 4.9 and a 16 FIR filters as a mixed phase mixing system, $A(z)$. The Filter matrix and the filter coefficients are shown in the section 4.3.2.
Figure 4.9: Four source signals (a) $s_1$ (b) $n_1$ (c) $s_2$ (d) $n_2$. The x-axis is time in m.sec and y-axis is amplitude.

An estimate of the observed signal set at microphones is calculated as $x(t) = As(t)$ shown in Figure 2.7. The Bessel coefficients of the observed signals in Figure 4.10 and the reference signal in Figure 4.12(b) are calculated and are passed as inputs to the ICA-R using Bessel algorithm. In ICA-R using EMD the residual signal of the reference signal is extracted as described in section 2.6 and is passed as input to the ICA-R using EMD along with the observed signals in Figure 4.10.

The output of the ICA-R algorithm using EMD based reference and the Bessel Features are shown in Figure 4.11(c) and Figure 4.12(c). The Table 4.3 gives the Mean Square Error (MSE) and the SNR (dB) for the two methods.
Figure 4.10: Observed signals using convolutive mixing (a) microphone1 (b) microphone2 (c) microphone3 (d) microphone4. The x-axis is time in m.sec and y-axis is amplitude.

Figure 4.11: ICA-R using EMD (a) Original/Desired speech signal. (b) Reference signal (different speech utterance of the target speech signal). (c) Extracted speech signal. The x-axis is time in m.sec and y-axis is amplitude.
Figure 4.12: ICA-R using Bessel Features (a) Original/Desired speech signal. (b) Reference signal (different speech utterance of the target speech signal). (c) Extracted speech signal. The x-axis is time in m.sec and y-axis is amplitude.

Table 4.3: Comparison of ICA-R with EMD and ICA-R with Bessel Features using MSE and SNR as measures

<table>
<thead>
<tr>
<th>Performance Measures</th>
<th>Methods</th>
<th>EMD based Reference</th>
<th>Bessel Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.8240 -004</td>
<td>7.9200e-006</td>
<td></td>
</tr>
<tr>
<td>SNR(dB)</td>
<td>26.241</td>
<td>73.541</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.13: Comparison of ICA-R using EMD and Bessel Features with respect to SNR (dB).

Figure 4.14: Comparison of ICA-R using EMD and Bessel Features with respect to Mean Square Error (MSE).

From the Figure 4.13 and Figure 4.14 we can say that the ICA-R using Bessel Features is much more effective than ICA-R using EMD.
Chapter 5

Summary and conclusions

In this thesis, we proposed an approach for extracting a desired speech signal from a mixed source signal using the ICA-R algorithm and Bessel features. Here the desired speech signal and the reference signal are two different speech utterances of a same speaker.

In the current existing literature most of the methods deal with designing the reference signal with prior information of the target speech signal. The crucial problem is the design of reference signal in advance which is close to the desired signal when the desired source signal is very week in mixed signals and also when there is no prior information about the desired source signal.

In the proposed method we do not require any prior information about the desired speech signal that has to be extracted. The ICA-R algorithm is extended to use Bessel coefficients of the observed signals and the reference signal for processing as they are more efficient in representing speech-like waveform.

From the simulation results and the performance analysis in chapter 4, comparing the proposed method with one of the previous existing methods shows that the proposed method is more effective. This shows that the computation done at the feature level i.e. the Bessel coefficients of the signals yields better results than on the sample values.

This is very useful for many applications such as speaker verification, speaker identification and so on even when the desired signal is very week in mixed signals [8]. An example of real world application where the proposed method is useful is the voice recognition and speaker verification login even in the presence of external disturbances and also in voice security applications.
References


