NAIST-IS-DD0261032

Doctoral Dissertation

Fixed-Point ICA based Speech Signal Separation and Enhancement with Generalized Gaussian Model

Rajkishore Prasad

February 2, 2005

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Engineering.

Rajkishore Prasad

Thesis Committee:

Professor Kiyohiro Shikano	(Supervisor)
Professor Kenji Sugimoto	(Member)
Associate Professor Hiroshi Saruwatari	(Co-supervisor)

Dedicated to

My

Parents

A very hard nut to digest

कर्मण्येवाधिकारस्ते मा फलेषु कदाचन ।
मा कर्मफलहेतुर्भूर्मा ते सङ्गोऽस्त्वकर्मणि ।।

(In Sanskrit taken from The Geeta, a holly book containing teachings of Lord Krishna to his disciple Arjuna, a great warrior of Mahabharata.)

"You have right to work only, but never to the fruits thereof, you should not have the fruits of your action as your goal, not let be there any desire for inaction"

Fixed-Point ICA based Speech Signal Separation and Enhancement with Generalized Gaussian Model

Rajkishore Prasad

Abstract

Speech signal separation and enhancement under blind setup is one of the challenging areas of practical application. Excellent solutions to these problems are always required for the spoken communication between man and machine in the real world. The problem of speech separation arises in the presence of multiple speakers and that of enhancement pertains to reduce the effect of noise and other interfering signals. In the real world applications these two problems are often occurring simultaneously and their solutions are urgently required in the development of full-fledged conversational interface. The aims and scope of our work is also in the same context. Recently, Blind Signal Separation (BSS) based on the Independent Component Analysis (ICA) has emerged as a potential engineering solution for speech separation problem. Such algorithms work with the assumption of statistical independence of each sources and estimate original sources as the independent or least dependent components. This thesis also addresses development and application of ICA based algorithm for the blind separation of convoluted mixture of speech, observed by a two element linear microphone array, under the over-determined situation. The proposed ICA algorithm is based on the non-Gaussianization, by negentropy maximization, of the Time-Frequency Series of Speech (TFSS) signal. The functioning of ICA by non-Gaussianization is based on the heuristic idea of Central Limit Theorem (CLT) under which it happens that the mixed speech signals become more Gaussian than the individual signal and thus by reversing the process of non-Gaussianization individual signals can be estimated with arbitrary scale and permutation. Under such a framework a cost function is required to measure the

degree of non-Gaussianization and maximally non-Gaussian signals are taken as the independent components which are original sources. There are various measures such as kurtosis, entropy, negentropy for measuring non-Gaussianization but negentropy provides much better robustness to outlier and is widely used. However, direct measure of negentropy is cumbersome and it is approximated in terms of cumulants or non-linear functions. In this thesis various approximations of negentropy of TFSS by the higher order statistics of the non-linear, non-quadratic functions and their separation performances have also been investigated. The nature of nonlinear function used to approximate negentropy of the data depends on its statistical characteristics of the data. The detailed study on the probability density of TFSS has been presented to test the relative proximity of underlying distribution of TFSS with that of Gaussian distribution, Laplacian distribution and Generalized Gaussian Distribution (GGD). The results of different statistical tests such as moment test, Chi-square test, and Quantile-Quantile (QQ) plots have been found to favour closeness of distribution of TFSS with that of GGD. Accordingly, a GGD function based non-linear function has been proposed for negentropy approximation and its use in ICA algorithm. Also, it has been found that the proposed non-linear function gives less error in approximation of the negentropy than the conventional functions. The separation performances of conventional and proposed non-linear functions have also been studied with the fixed-point Frequency Domain Independent Component Analysis (FDICA) algorithm and have been found that GGD based non-linear function improves rate of convergence of the algorithm.

The problem of speech enhancement has also been addressed in the frequency domain. The speech enhancement in the frequency domain is done by manipulating spectral component of the noisy signal in accordance with noise suppression rule. This thesis also proposes the development of noise suppression rule based on the Maximum A Posteriori (MAP) estimation. The proposed MAP estimator uses flexible statistical models, based on GGD, for the TFSS of the speech as well as noise signals. Thus the noise suppression rule is adaptive with the statistics of the noise and the same can be used to reduce effect of different types of noise such as Gaussian and super-Gaussian or spiky signals. The noise suppression characteristic of the

estimator depends on the type of noise. In contrast to this, most of the conventional methods such as Wiener filter show same noise suppression characteristics to Gaussian and spiky noise signals. The statistics of the noise signal and clean speech signal are also estimated from noisy signal. First the statistics of noise signal is estimated from the noise only segments of the noisy signal and are used to estimate statistics of the clean signal from the higher order statistics of the noisy signal and noise signals. In order to demark noise-only portions of the noisy signal, a novel voce activity detector based on the organization measure of the spectral components has been proposed. Again, negentropy has been used as the measure of organization of the spectral components which is different for noise-only frames and noisy speech frames. The experimental results of enhancement of speech contaminated by different noise signals shows its superiority over the conventional Wiener filter. The flexibility in the noise suppression characteristics of the proposed MAP estimator is suitable for doing post processing of the speech signal separated by FDICA algorithms. The problem is difficult in the sense that the residual noise is also speech. The separated signals from an FDICA algorithm contain components of undesired sources in the residual form. Since these residual signals are speech like noise, it can be further reduced using proposed MAP estimator by using one separated component as the target speech while others as the source of the noise. However, for the proposed post-processing the knowledge of the level of residual noise present in the target speech is required and can be determined from the information about noise reduction done by the FDICA algorithm. However, this method is not blind as it requires original contribution of each source to each microphone. The experimental results show that the post processing by the MAPS estimator gives appreciable improvements in the noise reduction.

Keywords: BSS, ICA, Speech Signal Separation, Negentropy, speech Enhancement

*Doctoral Dissertation, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-0261032, November, 2004.

Acknowledgements

I feel lighter and loosing weight on brain while writing this page because I am going to distribute doerships and expose each of them who have really magnified my abilities in someway. I would like to express gratitude and thankfulness to Prof. Kiyohiro Shikano, who as an advisor has shown enough propitiousness to support my flexibility and freedom in work. This is one of the best gifts I enjoyed during my whole stay in his laboratory, popularly known as Shikano Lab. I am also deeply indebted to Prof. Hiroshi Saruwatari, my advisor, for always finding some time for discussion, mitigating my confusions and successfully introducing me to the fascinating worlds of microphone array processing, independent component analysis and blind signal separation. I like to give sincere thanks to Prof. Kenji Sugimoto, Head Systems Science Lab., NAIST, and a member of my thesis committee for his valuable suggestions and comments. I am also thankful to Prof. Akinobu Lee and Prof. Hiromichi Kawanami for suggestions and inspirations.

I express my thankfulness to Prof. J.-F. Cardoso , TSI(Traitement du signal et des images), department at E.N.S.T. for uploading my BSS demo at ICA Central(http://www.tsi.enst.fr/icacentral/algos.html).

I also like to appreciate and acknowledge encouraging discussion and communications with Hiroshi Sawada, NTT Communication Science Laboratories, Japan, on different research issues.

I like to thank all the elder, younger and current students and researchers of Shikano-Lab whose friendliness made my stay very easier and interesting at NAIST.

I am grateful to BRA Bihar University (my parent university), Muzaffarpur , India, where I am working as a Lecturer in the University Department of Electronics, for providing leave and supporting my ambitions and research.

I also like to pay sincere thanks to my teacher and ex-Head Department of Electronics of my parent university, Dr. Shyama Charan Prasad, Prof. A.K. Sinha, Deptt. of Electrical Engineering, Muzaffarpur for constant encouragements. I am also very

much thankful to my father-in-law Dr. Tapaswy Yadav, retired Prof. of Physics, uncle Prof. Asheshwar Yadav, Prof. of Physics and is currently Vice-Chancellor of my parent university, my elder brothers Prof. Balram Rai, Prof. of Physics, Prof. Rajeev Ranjan, Associate Prof. of Physics for every type of help and suggestions.

I like to pay my sincere thanks, of course word are unable to do so, to my parents who have always been behind me to carve my career and life.

Also I can not forget to express thanks to my *Indojin , Nihonjin and Gaikukojin* friends at NAIST who have always been helpful in making life easier and joyous. I am thankful to all those Japanese students who volunteered to give his/her time for listening test. I am also thankful to Virendra Singh, Computer Test Lab, NAIST, for comments on different parts of thesis.

Finally, extra special thanks goes to my wife Kumari Bharati for believing in me and constantly taking my care. This page remains uncompleted without accepting role of my lovely son *Chichikun* whose true laughs and voice, a mixture of broken Hindi, Japanese and English (now Japanese is at his tip of tongue, Hindi is in middle and English is in third preference) have always been diluting every day's tension. Since one year old he has been living with me and missed loving cares all other family members which I can not return.

I like to thank Ministry of HRD, Govt. of India, New Delhi for selecting and recommending my candidature for Japan Govt. research fellowship and to MONBUSHO, Govt. of Japan for providing financial assistance, without which it would have been impossible to drive up to here, as a doctoral fellowship.

Contents

List of Figures

Chapter 1

Introduction

"The knower, knowledge and object of knowledge; these three motivate action. Even so the doer, the organs and activity, these three are the constituent of action"

.....The Geeta,Peached by Lord Krishna , Chapter 18.

1.1. Background and Problem

Recently, researches on conversational interface [1] for intelligent machines such as a robot and computer have received much research attention because it renders facility to users to command and converse with machines in a very natural and easy way despite huge intrinsic sophistication in the system. The origin of research goals in science and engineering to develop such systems that can listen, understand and speak in a natural language are not new rather it is rooted in antiquity [2], however, in the last two decades related research topics have been moving from fringe area to focus along with due modifications. In fact the advancements in the artificial techniques for speech recognition and developments of efficient Automatic Speech Recognizer (ASR) software, a central module in conversational interface, have accelerated demand and development of voice activated system. Of course valuable contributions from other very closely related areas such as dialog management, computational linguistics, speech synthesis, etc., can not be inevitably undervalued in the integrated system.

The concept of spoken communication with a machine has been inspired by the speech communication mechanism in humans. However, the state of art in this technology does not model and implement the auditory system in toto. The establishment of vocal communication between human and machine seems easy and enjoyable but the mathematical modeling and physical implementation of the underlying myths have been proven to be one of the grand challenges for the

modern computing technologies. One of the most important causes for this is that the fundamentals of underlying processes of speech communication are still not crystal clear or undiscovered. However, efforts are underway but problems are numerous. Here only a surfacial touch to some of them will be given while focusing on the problem of blind separation of speech in hot pursuit.

As said above an ASR plays central role in the conversational interface, but only after fulfillment of many constraints its efficient use in the real world applications is possible. The most important thing is the quality of speech signal being feed. If the signal being feed is clean and undistorted like training data it is the best case but as the quality of test signal degrades, dramatic degradation in recognition accuracy of an ASR is well-known [3]. The journey of speech from mouth of a speaker to the speech pick-up device is not scot-free. The chances of getting it contaminated by background noise, or acoustic signals from other sources or the reflected and delayed version of itself are very common and are always possible in real world applications. It depends on the characteristics such as availability of different sound sources or nose signals and reverberation of acoustic environment shared by speaker and sensor. In Figure 1.1 some of the most frequent aberrations, a speech signal may suffer have been depicted. Any of such possible aberration in speech signal becomes problematic to speech recognizer because such signals produce new acoustic patterns to the system for which it is not trained and may lead to invalid recognition. Since every subsequent processing steps starting from speech signal pick-up and their performance depend upon the quality of speech signal, it is important to pick-up signal with the best possible quality. Since the impinging of undesired signal on the microphone can not be avoided completely there is always need of algorithms for cleaning the captured speech from noise signals, for minimizing the effect of reverberation and for separating the speech from other speech in the pre-processing stage of recognition. In real world applications chances of occurrence of these problems in alone are rare and their simultaneous appearance is very common situation which makes the problem complicated and challenging. As said above the idea of equipping a machine with vocal-activity is inspired by

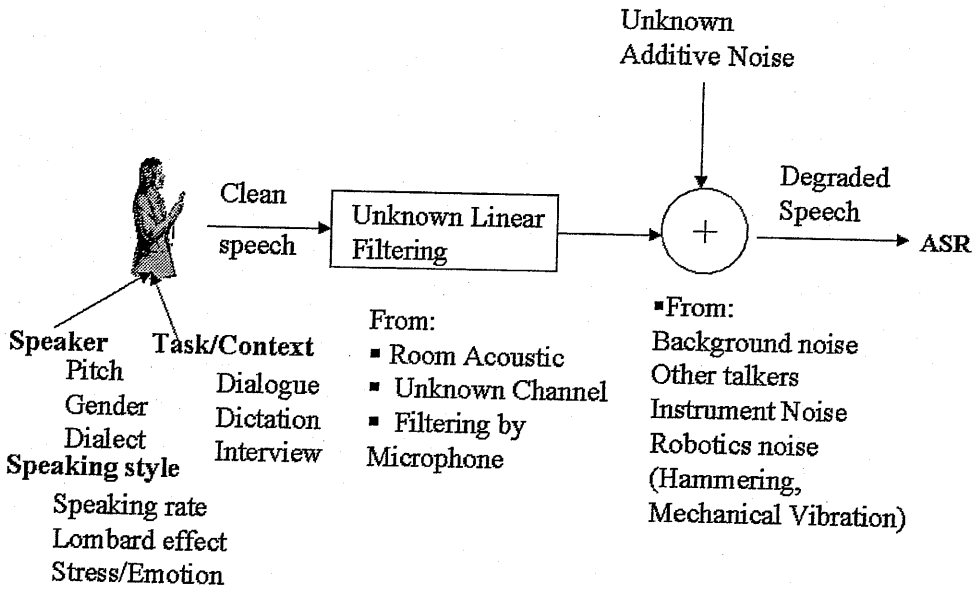


Figure 1.1. Degradation of speech in journey from speaker to ASR.

the natural system, it becomes important to implement special anthropomorphic capabilities required for ordinary conversation. The problem of speech signal separation arises in the multiple speaker environment where one is interested in hearing to a particular speaker for example hearing in crowd. Humans do it easily in everyday life. In the engineering sense humans are able in steering their hearing attention to a particular speaker! This anthropomorphic capability has been well-documented as Cocktail Party Effect in the scientific community [4]. This was recognized much before, however, still very little is known about underlying processing of simultaneous speech signals in human brain [5]. In the engineering sense problem is depicted in Figure 1.2, where there are many sources of acoustic signals and signals from all of them give very confused mixture to microphone. The signal recorded by microphone under such situation is garbage, if feed as it is, for an ASR. Thus in the artificial conversational interface imitation and implementation of human like capability of steering

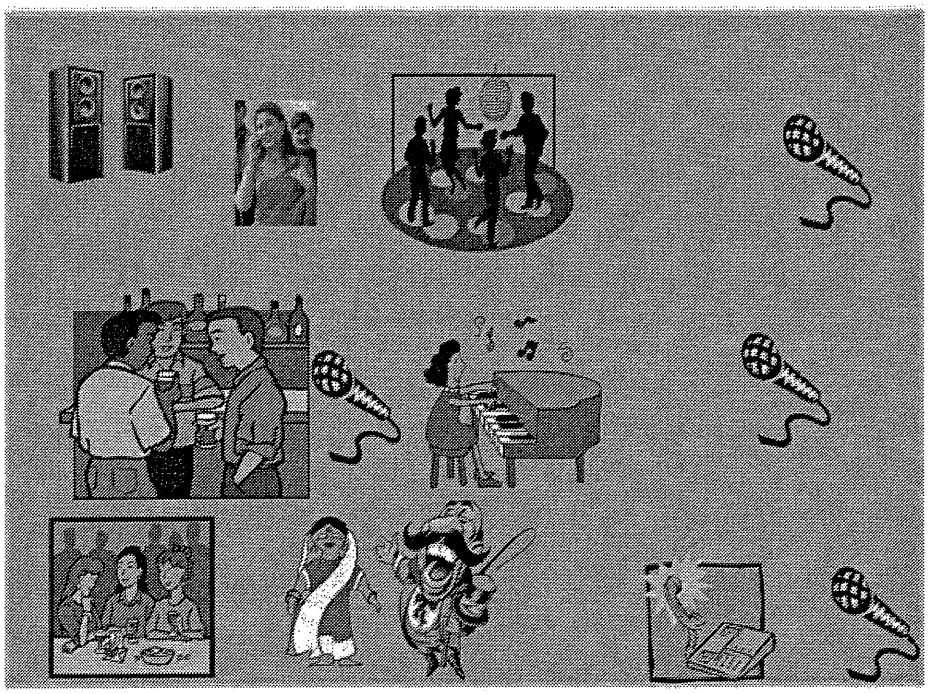


Figure 1.2. Cock-Tail Party situation. For humans it is not very difficult to pay their hearing attention to a particular speaker or sound. But the speech signal captured by microphone is confusing for ASR. The BSS problem is how to extract signal of interest from the signal observed at microphones without using any other information (in very strict sense).

hearing attention is essential for its usefulness in real world applications Whatever may be the signal representation in the human brain for the perception of individual speech signals from their hotchpotch [6][7], the engineering translation of Cock Tail Part Effect is the separation or extraction of individual signals from the cacophony of sounds. The problems becomes challenging and complicated because in the real situation one has only access to mixed signals observed by a microphone, or many microphones or by a microphone array and estimation of original sources under such conditions seems magical. However, there have been development of many approaches for the same and can be broadly categorized into two groups, namely, method based on single channel

input and those based on multichannel inputs. In the first category algorithms like tracking of formant structure [8], the organization technique for hierarchical perceptual sounds [9], methods based on Computational Auditory Scene Analysis (CASA)[10] have been proposed. In the second category various geometrical methods exploiting spatial and temporal information provided by a microphone array, have been proposed. In such methods Direction of Arrival (DOA) of the signal sources are estimated then separation system is computed by adjusting directivity pattern of the microphone array e.g. delay and sum (DS) beamformer and Adaptive Beamformer (ABF)[11][12][13][14]. The most important task in beamforming algorithms is DOA estimation and the separation performance deteriorate with inaccuracy in DOA estimation. In contrast to these source separation techniques, BSS algorithms do not need priori information like DOA. It is based on the higher order statistics of the signal which is used to segregate signals by restoring their statistical independence. It is called blind in the sense that there is no access to mixing process and estimation has to start from garbage (mixed signal). Thus it is process of estimation from nothing to something. In reality, the unique estimation of original signals is not possible without some prior knowledge; however, with certain indeterminacies such as scaling, permutation, and delay it is possible. It is also, practically, not problematic because in large number of applications, except for applications involving dynamical modeling, such arbitrariness in estimation are acceptable because useful information are available in the estimated waveform. In this formulation the Cock tail party effect totally fits. In general the BSS problem can be formulated as estimation, subject to aforesaid indeterminacies, of R original sources from their M mixed observations; from a MIMO system. However, it is not essential that the signal is coming from MIMO system in special cases it may be SIMO or SISO too. Among the solutions to this problem Independent Component Analysis (ICA) [17] based approaches are so much dominating, the use of term ICA seems a synonym for the BSS, however, it is one of the powerful tools for BSS. The ICA based BSS algorithms estimates original sources as the independent components of the mixed signal assuming all sources are

statistically independent[15][16]. This is the only prior assumption imposed on the sources, despite method is called blind. The details for different ICA algorithms will be discussed in the next chapter. ICA based algorithms for separation of speech signal have been developed both in time-domain and in frequency domain. However, separation of the convoluted mixture is easier in frequency domain because convolution is converted into multiplication and mixing in each frequency bin become instantaneous, but such ease is accompanied by other problems of permutation and scaling which must be fixed to get separated signal [18]. In this dissertation too, an ICA based BSS algorithm based on non-Gaussianization of the mixed signal has been studied for its application in separation of convoluted mixture of speech in frequency domain. The non-Gaussianization based ICA algorithm works under the assumption of Central Limit Theorem (CLT) which states that if N independent and identically distributed standardized random variables $X_1, X_2, \ldots\ldots X_N$ with arbitrary Probability Density Function (PDF) are combined to form another variable Z_N given by

$$Z_N = \sum_{1}^{N} X_N, \tag{1.1}$$

the distribution of Z_N converges to Gaussian distribution. The effect is reversible i.e. if the resulting random variable Z_N is non-Gaussianized, addend independent variables $X_1, X_2, \ldots\ldots X_N$ can be estimated. This gives one of the backbone techniques for the estimation of independent components. In the mixing process the mixed signals gain Gaussianity in similar way and can be separated by non-Gaussianization. Different algorithms, based on this working principle, have been developed using different measures such as kurtosis and negentropy for the non-Gaussianization. Since negentropy based measure provides better robustness to outliers in the data, its separation performance is better than that of the kurtosis based algorithms [19].

In this thesis we have used negentropy based measure for non-Gaussianity measure, as proposed in [27], and applied for the separation of convoluted

mixture of speech signal, captured by a two-element linear microphone array from two speakers. The application of fixed-point ICA for speech signal separation calls for many important points such as choice of non-linear function for negentropy approximation of Time Frequency Series of Speech (TFSS), solution for permutation and scaling problems, and removal of residual interfering signal after separation by ICA. The choice of non-linear function depends on the underlying PDF of the TFSS. Here we have made study on the statistical modeling of the TFSS and accordingly propose a new non-linear function, based on Generalized Gaussian function, for speech signal separation by fixed-point Frequency Domain ICA (FDICA) algorithm. The validity of basic idea of CLT has also been checked in case of speech mixing process. The combination of Null beamforming and fixed-point ICA has been studied to mitigate the effect of CLT non-compliance by TFSS. This thesis also proposes a novel method for denoising speech signal based on proposed statistical model. The study on application of proposed noise suppression rule in speech enhancement and in removal of interfering signal from separated signal has also been discussed.

1.2. Organization of Thesis

Rest of this dissertation is organized as follows:

In **Chapter-2**, the problem of blind signal separation has been defined under general framework. It also describes functioning and different approaches of ICA based BSS algorithms. Our main emphasis in this chapter is on the fundamental of non-Gaussianization, by negentropy maximization, based ICA algorithm for BSS.

Chapter-3 deals with the application of ICA by negentropy maximization, in frequency domain, for the separation of convoluted mixture of speech observed by a linear microphone array. It rationalizes the application of non-Gaussianization based ICA algorithm on the frequency sub-banded speech data in the light of CLT. It presents development of a deflationary learning rule to extract independent components from mixed signal by the negentropy

maximization. The separation performances of the algorithm under non-reverberant and reverberant conditions have been investigated. This chapter also explores how the non-compliance with the CLT by mixed speech signal affects the separation performance of the algorithm.

In **Chapter-4**, probabilistic modeling of TFSS has been described. Starting from Short-Time Fourier Transform (STFT) analysis, different statistical tests such as moment test, Chi-Square(χ^2) test , have been performed to check and compare closeness of the underlying PDF of the TFSS with Laplacian, Gaussian and Generalized Gaussian Distribution (GGD) functions. This chapter also deals with the parameter estimation of GDD using maximum-likelihood method. This chapter ends with the blind detection of CLT disobeying bins using GGD modeling and its application in combining null beamformer and fixed-point FDICA to mitigate the effect of CLT non-compliance on signal separation.

The work presented in **Chapter-5** have their bearings on Chapter-4. A GGD based non-linear function has been used to approximate negentropy of TFSS and has been compared with the approximation by other conventional non-linear function. Accordingly, the same has been used in the FDICA algorithm and its performances have been investigated with proper explanation for the experimental outcomes.

In **Chapter-6**, a general method for speech enhancement based on the GGD modeling of the speech and noise spectral components has been discussed. A MAP estimator, using GGD a priori, for this purpose has been derived. The experimental results for enhancement of speech, noised to different SNR levels by Gaussian and super-Gaussian noise signals, have been presented. The same technique has been applied to enhance output of FDICA and related results have been presented.

Chapter-7 contains summary of the thesis and related topics for future research. This is followed by references and my publications related to this thesis.

Chapter 2

Blind Signal Separation

2.1. Introduction

Blind Signal Separation (BSS), a very hot topic of research among digital signal processing groups since a decade, is the general framework to estimate signal contribution of latent sources only from their observed mixtures without knowing the mixing process. In the BSS problem we are given with the observation $x(n)=[x_1(n),x_2(n)....x_M(n)]^T$ at M sensors produced by some unknown interaction function F among the R original sources $s(n)=[s_1(n),s_2(n).....s_R(n)]^T$ given as

$$x(n)=F[s(n)], \text{ where } n \text{ is time index.} \tag{2.1}$$

The task of BSS is to estimate the optimal $\hat{F}^{-1}$, the inverse of the interaction function, so that the underlying original sources can be optimally estimated, i.e,

$$\hat{s}(t)=[\hat{s}_1(n),\hat{s}_2(n).....\hat{s}_M(n)]^T=\hat{F}^{-1}[x(n)]. \tag{2.2}$$

The interaction function depends on the physical situation such as on the geometry of sources and sensors, the number of sources and sensors, and the source to sensor transfer function. Hereafter, we will refer to interaction function F as the mixing matrix and inverse interaction function $\hat{F}^{-1}$ as the demixing matrix. For the simplest condition F can generate linear instantaneous mixture. However, in this dissertation we will consider for the convolutive mixing system.

Because the method is blind and unsupervised in functioning [15], it has gained wide areas of applicability such as in speech processing, image processing, bio-informatics, cosmo-informatics [20], etc. BSS techniques have

emerged as one of the potential solutions for the extraction or segregation of hidden signals only from their observed mixtures. In the area of speech signal separation it provides one of the feasible solutions for the extraction of speech signal from the cacophony of the sounds.

2.2. ICA based BSS

The complete lack of a mixing process in the estimation of the original sources is compensated by pivoting computation on the assumption of the statistical independence of each latent source. However, the observed mixtures of signals are not statistically independent due to the unknown mixing process. The principle of statistical independence is brought into play by looking for either non-Gaussianity of or spectral dissimilarity among the sources [21]. The process of taking out hidden sources as the most independent components of the mixed data is called Independent Component Analysis (ICA) and there have been developments of numerous ICA-based BSS algorithms in the different areas of practical applications involving multisensor signal processing, such as, speech recognition and enhancement, biomedical signal analysis and classification, source localization and tracking by RADAR and SONAR equipments, cosmological image classification, and data mining [19][22][23][20]. The basic functioning of the ICA based BSS algorithms are shown in Figure 2.1. The observed mixed signals $x(n) = [x_1(n), x_2(n).....x_R(n)]^T = As(n)$, where A is the mixing system, are passed through a tentative initial demixing system W (randomly chosen or based on some heuristic guess and subject to further modification) and then the mutual independence among the estimated independent component signals $y(n)$ is evaluated by some cost function $J(W,y)$, usually based on the statistics of the signal and candidate demixing system. That in turn goes on modifying demixing system unless and until the cost function is not optimized for the maximum mutual independence among the separated ICs. So, paradigmatically, most of the known ICA-based BSS algorithms exhibit such functional similarities, but basic differences occur in the choice of the cost function, the domain of operation and the process of optimization.

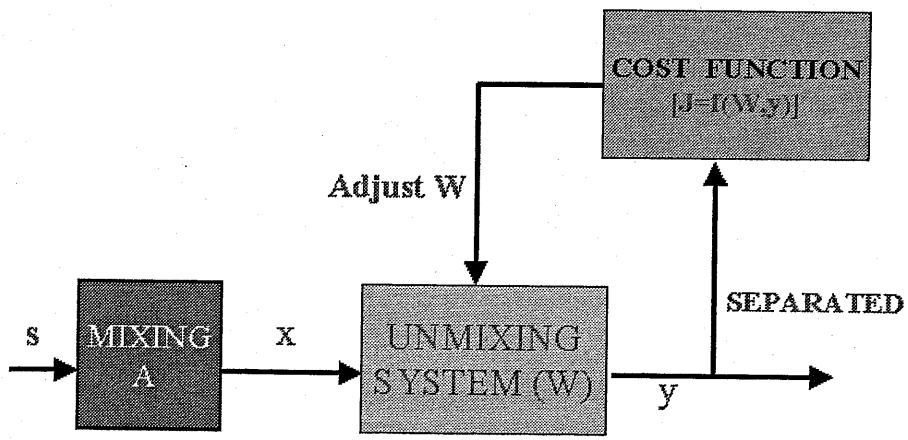


Figure 2.1 Block diagram showing basic working principle of the ICA based BSS algorithms.

The cost function may be based on the joint distribution or the marginal distribution of the signal. The most popular example of the first category is the Kullback-Liebler Divergence (KLD) metric, which measures deviation between the joint distribution of the signal and a pre-assumed source distribution. The second category of cost functions exploit only statistical properties of the marginal distribution and non-Gaussianity of the data. The most important examples of such cost functions are kurtosis and negentropy. The method of optimizations also differs, e.g., gradient based algorithm, evolutionary algorithms, fixed-point algorithms, etc. These cost function require prior knowledge of the source distribution which is not always feasible, however, some good approximations of their PDFs are used. The cost functions essentially measures degree of independence obtained in the process of optimization and there are many ways to measure the statistical independence which in turn has led to development of many ICA algorithms. Thus in an ICA based BSS algorithm ICA algorithm plays central role.An ICA process can be summarized as follows:

ICA Process= Objective function (Independence measure) +Optimization.

Here too I will give some elementary level description of well-known methods however ICA by non-Gaussianization and its application in speech signal separation is focal theme of this thesis.

2.2.1. Statistical Independence and Uncorrelatedness

As mentioned above the form of cost function depends upon the fact how independence is measured. Thus it will be not out of place to present some fundamentals of concept of statistical independence. Uncorrelatedness and independence are related terms but distinct in the statistical sense. Later is known as orthogonality or linear independence. Two random variables x_1 and x_2 are said to be decorrelated if

$$\text{cov}(x_1, x_2) = E\{x_1 x_2\} - E\{x_1\}E\{x_2\} = 0 \tag{2.3}$$

where $E\{x\}$ is the expected value of x and $\text{cov}(x_1, x_2)$ represents covariance of x_1 and x_2. In the case of correlated variables it becomes non-zero, may be positive or negative depending on the fact if x_1 is increasing with increase in x_2 or decreasing. It is symmetric relation as $\text{cov}(x_1, x_2) = \text{cov}(x_2, x_1)$.

In general decorrelation or uncorrelateness does not imply independence (except for Gaussian random variables).The set of N variables x_i are said to be statistically independent if

$$p(x_1, x_2 .... x_N) = \prod_{i=1}^{N} p(x_i) \tag{2.4}$$

where $p(x)$ represents PDF of x. The more tractable form of the above condition is expressed in terms of non-linear decorrelation of the involved random variables. Accordingly, two random variables x_1 and x_2 are independent if

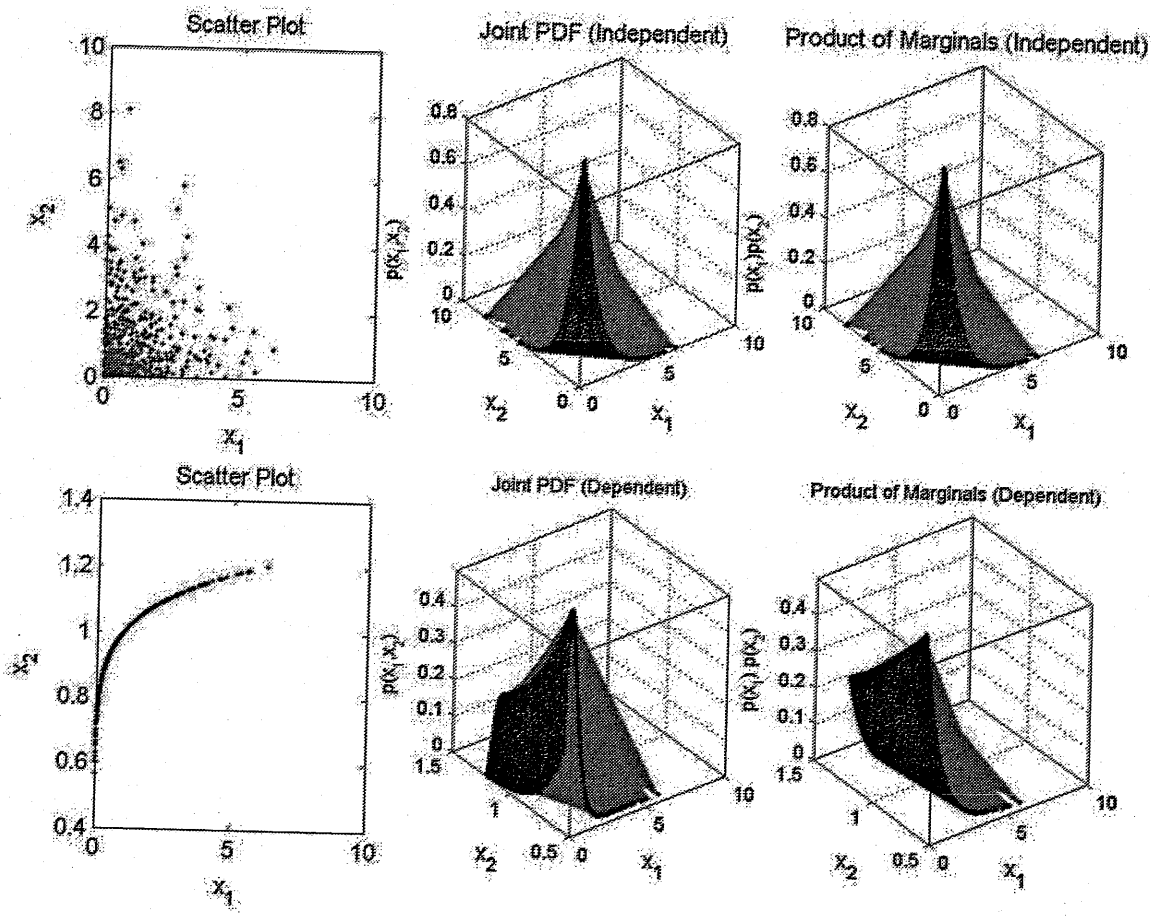


Figure 2.2 These figures illustrate assumption of statistical independence described in Eq.(2.4). The 1000 samples of random variables x_1 and x_2 from exponential distribution with marginal density functions $p(x_1)=\tau_1 e^{-\tau_1 x_1}$ and $p(x_2)=\tau_2 e^{-\tau_2 x_2}$ were used. The joint PDF $p(x_1,x_2)=\frac{\tau_1\tau_2}{1-\rho}\exp\left\{-\frac{(\tau_1 x_1+\tau_2 x_2)}{1-\rho}\right\} I_0\left\{\frac{2(\rho\tau_1\tau_2 x_1 x_2)^{0.5}}{1-\rho}\right\}$ is Downton's bivariate exponential PDF in which ρ represents correlation coefficient between them and $x_1,x_2,\tau_1,\tau_2>0;0\le\rho\le 1$ and $I_0(\omega)$ is modified Bessel function of ω of first kind. Plots in first row show scatter plot, joint PDF and product of marginal PDFs for independent x_1 and x_2 while plots in second row show the same when random variables x_1 and x_2 are dependent, as is obvious from their scatter plot.

Table 3.1 Measure of non-Gaussianity

Name	**Definition**
Differential Entropy	$H(x) = -\int p(x) \log p(x) dx$
Kurtosis	$K(x) = E\{x^4\} - 3(E\{x^2\})^2$
Negentropy	$J(x) = H(x_{Gauss}) - H(x)$
Mutual Information	$I(x) = J(x) - \sum_{i=1}^{n} J(x_i)$

$$E\{g(x_1)h(x_2)\} = E\{g(x_1)\}E\{h(x_2)\}. \tag{2.5}$$

where $g(.)$ and $f(.)$ are some non-linear transformation over x_1 and x_2. It can be imbued from Eq.(2.4) that extraction of independent components calls for a non-linear decorrelation of the variables such that variables are uncorrelated in the transformed space too. Thus the independence condition is stronger than the simple uncorrelatedness or linear independence of the variables.

The ICA algorithms separated independent components by looking for such independence in the mixed observation. There are several ways of measuring independence between set of random variables. One of the most straightforward methods is the use of Kullback-Leibler Divergence(KLD) as a measure of distance between two PDF $p(x)$ and $p(y)$ given by

$$KLD(x \| y) = \int p(x) \log \frac{p(x)}{p(y)} dx. \tag{2.6}$$

Use of definition of independence in Eq.(2.4) can lead to the KLD between the joint distribution of x and product of its independent constituents x_i given by

$$KLD(x \| y) = I(\boldsymbol{x}) = \int p(\boldsymbol{x}) \log \left(\frac{p(\boldsymbol{x})}{\prod_i p(x_i)} \right) d\boldsymbol{x}. \tag{2.7}$$

This is also known as mutual information $I(x)$ and is zero if the components x_i are mutually independent. This requires knowledge of $p(x)$ which is very hard to estimate and estimations are not fully reliable. However, several computational methods such as approximation by polynomials involving cumulants for the approximation of $p(x)$ have been developed and used in ICA algorithms [19]. In the area of speech signal such algorithms has been applied and developed but are computationally extensive and takes huge amount of iterations to learn separation matrix.

2.2.2. ICA by Non-Gaussianization

The other way of identifying hidden independent components in the data is to look for maximally non-Gaussian components [19]. This is based on the CLT which states that the mixing of two or more non-Gaussian signals pushes the distribution of mixed signal towards Gaussian distribution. Thus reverse of the same i.e. non-Gaussianization of the mixed signal can yield independent components. The different objective functions used to measure non-Gaussianity are shown in the Table 3.1.The non-Gaussianity measures like kurtosis and negentropy are based on the marginal distribution of the signal, however, negentropy is more robust to outlier than the kurtosis. In this thesis too negentropy will be used as an objective function.

A lot of algorithms using such cost functions have also been developed and is main concern of this paper. Examples of algorithms based on such cost functions and non-Gaussianization of the signals are fixed-point ICA by the kurtosis or negentropy maximization [19][23][24][25]. Such an algorithm was first developed and proposed in [26] for the separation of the instantaneous mixture. The key feature of this algorithm is that it converges faster than other algorithms, e.g., natural gradient-based algorithms, with almost the same separation quality. In [27] , the fixed-point algorithm of has been extended for complex-valued signals; however, this algorithm has no strategy for solving the problem of permutation and scaling arising in speech signal separation in the frequency domain. The fixed-point algorithm for audio source separation can be found in [28][29]. In [28], authors have proposed the application of the

fixed-point algorithm for speech signal separation with the time-frequency-model-based likelihood ratio jump scheme as a solution for permutation. In order to combine array signal processing techniques with fixed-point ICA by negentropy maximization, we proposed in [29] an algorithm for the audio source separation of convolutive mixtures using a directivity-pattern-based technique [30] to solve the permutation and scaling problem. Also, fixed-point-iteration-based ICA is very sensitive to the initial value from which iteration starts. The fixed-point FDICA algorithm for audio source separation works on the time-frequency series of speech (TFSS), and thus assumes obeyance of CLT from the TFSS in each frequency bin. However, in [31] it has been shown that TFSS of the mixed speech signal fails to follow CLT in every frequency bin and the separation performance of the algorithm too falls in such frequency bins. In general, any ICA algorithm based on the non-Gaussianization of the signal in the light of CLT can face a similar adverse situation and may fight to loose its performance in the same way because of non-compliance with CLT by the TFSS. Such disobedience of CLT by the TFSS pops up many hooked-up questions such as regarding suitability of negentropy based method for speech signal separation, why such failure occurs and how to get rid of it? These novel points will be discussed in coming chapters. The other important point in the marginal statistics based cost function is the statistical model used for marginal PDF of the data. In the context of speech signal statistical modeling of its spectral components will be explored with aim to use its best approximation in the non-Gaussianization based FDICA algorithm.

Chapter 3

Speech Signal Separation by non-Gaussianization based FDICA

3.1. Introduction

In the previous chapter how non-Gaussianization can be used in obtaining hidden independent components from the mixed data has been presented. This chapter applies the same technique for separation of speech signal observed by a linear microphone array. It is important to mention here that it is not necessary to use microphone array in multi-channel algorithm by BSS. It can be done by the distributed microphones, not with a fixed geometry such as linear, and circular, that can pick-up spatial variation of the signal. However, in our approach we have combined array processing technique with fixed-point FDICA to solve the permutation and scaling problems. Also, it has been investigated in [29] the fixed-point FDICA algorithms is sensitive to initial separation matrix and better separation can be obtained using some good initial guess for separation matrix such as null beamformer based initial separation matrix. In this regard use of microphone array is beneficial over randomly distributed many microphones. In this chapter the convolutive mixing model of the speech signal will be considered and separation will be done in the frequency domain using fixed-point ICA by negentropy maximization.

The origin of the BSS technique in audio signal separation can be traced back to the contributions of Cardoso [32] and Jutten [33] for practical signal

separation algorithms based on the aforementioned principle of statistical independence of the sources [34]. These algorithms are based on higher order statistics of the signals mutual independences measure among the independent components (IC). Recently, there have been development of many excellent algorithms, in the time domain and in the frequency domain or mutualistically combined in both while weighing their pros and cons, for audio source separation based on ICA [35] [36][37]. In fact, in the list of BSS methods for the audio source separation, ICA-based BSS algorithms have been dominating due to the emergence of several algorithms. However, due to their computational complexities and slow convergence there hardly exists any algorithm that can handle the general class of BSS problems for real world applications in real time [38]?

3.2. Speech Signal Mixing and Demixing

In the real recording environment, signals reaching each microphone are not only direct-path signals, but also delayed and attenuated versions of the source signals, which gives thought of existence of virtual or mirror sources, and noise signals. Therefore, in the real world mixing model is best approximated by the convolution of the source $s_i(n)$ to sensor transfer function and the source signal components reaching microphones. Accordingly, the speech signals picked up by a microphone array with M microphones are modeled as a linear convolutive mixture of R impinging source signals $s_i(n)$ such that the M-dimensional signal vector picked-up by the array is given by

$$x_j(n) = \sum_{i=1}^{R} \sum_{p=1}^{P} h_{ji}(p) s_i(n-p+1) + d_i(n); \quad (j = 1, 2, ...M), \tag{3.1}$$

where $s_i(t) = [s_1(t), s_2(t), ...s_R(t)]^T$ represents the original source signals, h_{ji} is the P-point impulse response between the source i and the microphone j, $d_i(n)$ is the noise signal, and n is the time index. The mixing model given here is for the arbitrary number of speakers and microphones, however, in this thesis we consider the case of two microphones and two sources, i.e., $M=R=2$, for

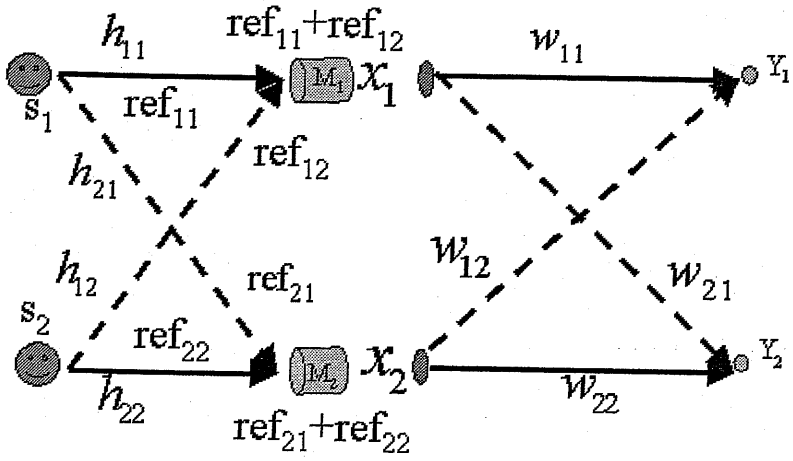


Figure 3.1 Convolutive mixing and demixing models for speech signal at the two element linear microphone array. The mixed signals $x_1(n)$ and $x_2(n)$ at microphones were obtained by adding the speech signals ref_{11}, ref_{12}, ref_{21}, and ref_{22} reaching each microphones from each source. The speech signals ref_{11}, ref_{12}, ref_{21}, and *ref22* reaching each microphone from each speaker are called as the reference signals. The right half (after microphones M_1 and M_2) of the figure shows demixing process, a reverse of the mixing process.

simplicity and convenience and no noise condition. For such a situation the signal mixing and demixing models are shown in **Figure 3.1**. Accordingly, the observed signals $x_1(n)$ and $x_2(n)$ at the microphones are given by

$$\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \otimes \begin{bmatrix} s_1(n) \\ s_2(n) \end{bmatrix} = \begin{bmatrix} ref_{11}+ref_{12} \\ ref_{21}+ref_{22} \end{bmatrix}, \tag{3.2}$$

where $ref_{11}=h_{11} \otimes s_1(n)$; $ref_{12}=h_{12} \otimes s_2(n)$; $ref_{21}=h_{21} \otimes s_1(n)$; $ref_{22}=h_{22} \otimes s_2(n)$ are called reference signals and '$\otimes$' represents convolution operation.

3.3.1. Frequency domain model

As stated earlier that an FDICA algorithm separates signal independently

in each frequency bin. The signal is decomposed into frequency bins by time-frequency analysis. The notion of time-frequency analysis, which has been analogically developed from the concept of coherence states in quantum mechanics [39][40], grew in the field of signal processing for the analysis of speech signal. The technique of time-frequency processing of signal, especially suitable for the processing of non-stationary signals, captures both the static and dynamic aspects of spectral information in a single feature vector. Thus the time-frequency series not only gives spectral information but also reflects the time dependence of the spectral components. It has a wide range of applications such as acoustic analysis, radar tracking, and adaptive filtering. Since the speech signal is also non-stationary, more accurate analysis of it is possible by means of time-frequency processing. During the last 50 years there have been developments of many powerful time-frequency analysis methods for speech signal with comparable merits and demerits. Details of some of these methods can be found in [40][41][42][43]. The two most widely used methods; Short-time Fourier Transform (STFT) and Linear Predictive Analysis (LPC) make the implicit assumption that speech signals are stationary over a very short time interval called the analysis window size. This assumption leads to trade off between the achievable frequency and time resolution due to uncertainty principle. The other methods such as Cohen's class of generalized time frequency representation [40] and Winger distribution analysis [43] provide high time-frequency resolution without any trade off but are complicated by the inference terms [44]. We will consider in this study time frequency analysis of a speech signal by the STFT method for analysis in the joint domain of time and frequency. The whole process of STFT analysis is depicted in Figure 3.2.The arbitrary speech signal $x(n)$ is divided into M quasi-stationary frames by using overlapping analysis windows (hanning/hamming) $h(n)$ of the fixed length (say 20 ms), called as frame length such that

$$x_w(n,\lambda) = x(n)h(n-\lambda\varepsilon),\ \lambda=1,2,...M; \tag{3.3}$$

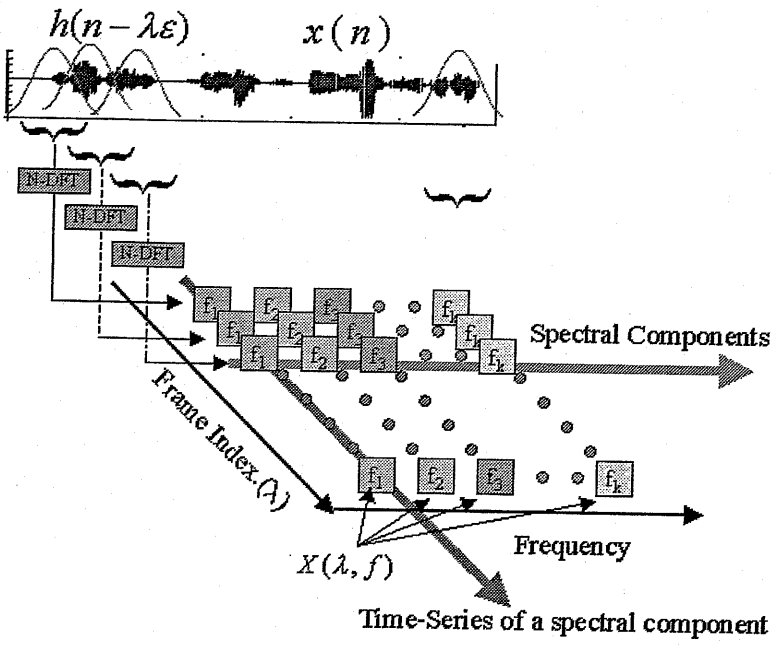


Figure 3.2 Process of the generation of time-series of speech spectral components by STFT analysis.

where $\lambda =$ is frame no. and ε is step size. Then N-point DFT of each such segment is taken to produce short-time spectrum $X(\lambda, f) = [X(\lambda, f_0), X(\lambda, f_1), ..., X(\lambda, f_N)]$ of N frequency components in which each is given by

$$X(\lambda, f_k) = \sum_{n=0}^{N-1} x_w(n, \lambda) e^{-j\frac{2\pi f_k n}{N}}, 0 \le k \le N. \tag{3.4}$$

The complex samples of the same frequency from each $X(\lambda, f)$ are chosen and stacked in time succession (in accordance with the frame no. which corresponds to time) to form a time series of spectral components or TFSS. Thus the time series of the kth frequency component (also known as frequency bin) is expressed as

$$Z(f_k) = [X_1(1, f_k), X_1(2, f_k), ...X_1(\lambda, f_k), ..., X_1(M, f_k)]^T, \tag{3.5}$$

The idea of using the short time spectra for the separation was first proposed in [18] because the impulse response between source and microphone are assumed to be stationary over short time due to which mixing process in each frequency bin does not remain convolutive rather it can be assumed to be instantaneous. Thus in the frequency domain formulation the mixing model in Eq.(3.1) can be expressed by taking its STFT as follows

$$X_j(\lambda, f) = \sum_{i=1}^{R} H_{ji}(\lambda, f) S_i(\lambda, f) + D_i(\lambda, f); \quad (j = 1, 2, ...M), \tag{3.6}$$

where symbols in capital represents STFT of the quantity represented by lower case letters. For the case of two speakers and two microphones mixing model under the clean condition (no noise) is given as

$$\begin{bmatrix} X_1(f) \\ X_2(f) \end{bmatrix} = H(f)S(f) = \begin{bmatrix} H_{11}(f) & H_{12}(f) \\ H_{21}(f) & H_{22}(f) \end{bmatrix} \begin{bmatrix} S_1(f) \\ S_2(f) \end{bmatrix}. \tag{3.7}$$

This equation revels that the mixed signal in any frequency bin is composition of the contribution from each sound source. Thus under the light of CLT the Gaussianity of either of mixed signal will exceed that of individual contribution signal in any frequency bin (However, later in this thesis it will be shown, of course that is also a part of contribution of the thesis, that this is not always true and is problematic for the ICA algorithm). This forms the basis for spectral separation of the speech signal by ICA methods based on non-Gaussianization. The FDICA separates the signal in each frequency bin independently, and this separation process is given by

$$\begin{bmatrix} \hat{S}_1(f) \\ \hat{S}_2(f) \end{bmatrix} = \begin{bmatrix} Y_1(f) \\ Y_2(f) \end{bmatrix} = W(f)X(f) = \begin{bmatrix} W_{11}(f) & W_{12}(f) \\ W_{21}(f) & W_{22}(f) \end{bmatrix} \begin{bmatrix} X_1(f) \\ X_2(f) \end{bmatrix}. \tag{3.8}$$

where $[Y_1(f), Y_2(f)]^T$ are ICs ; $S(f)=[\hat{S}_1(f)\ \ \hat{S}_1(f)]^T$ are estimated TFSS of the sources, and $W(f)=$ separation matrix in frequency bin f. Any one row of the separation matrix is called separation vector for a particular source.

3.3. BSS Algorithm for Spectral Separation

FDICA algorithm works on the TFSS of the mixed speech data to sieve out TFSS of the independent components in each frequency bin. Fixed-point ICA was first developed and proposed in [26] for the separation of the instantaneous mixture. The key feature of this algorithm is that it converges faster than other algorithms, like natural gradient-based algorithms, with almost same separation quality. However, the algorithm in [22][26] is not applicable to TFSS as these are complex valued. In [27][45], fixed-point ICA algorithm of [26] has been extended for the complex-valued signals, however, this algorithm has no strategy for solving the problem of permutation and scaling arising in FDICA for speech signal separation. The fixed-point ICA algorithm [22] is based on the heuristic assumption that when the non-Gaussian signals get mixed it becomes more Gaussian and thus its non-Gaussianization can yield independent components. The frequency domain mixing model for the speech signal in Eq.(3.7) revels that the TFSS in any frequency bin is superposition of spectral contributions of each source. Thus, in the light of CLT,TFSS of mixed speech signal in any frequency bin is more Gaussian than that of any independent source.

Obviously, non-Gaussianization of TFSS can give TFSS of independent sources from which original signals can be reconstructed. The process of non-Gaussianization consists of two-steps approaches, namely, pre-whitening or sphering and rotation of the observation vector as shown in Figure 3.3. Sphering is half of the ICA task and gives spatially decorrelated signals. The effect of mixing, whitening and rotation on the data is shown in the scatter plots of Figure 3.4. Whitening of the zero mean TFSS is done using Mahalanobis transform [46]. Accordingly, the whitened signal $X_W(f,t)$ in the fth frequency bin is obtained as follows:

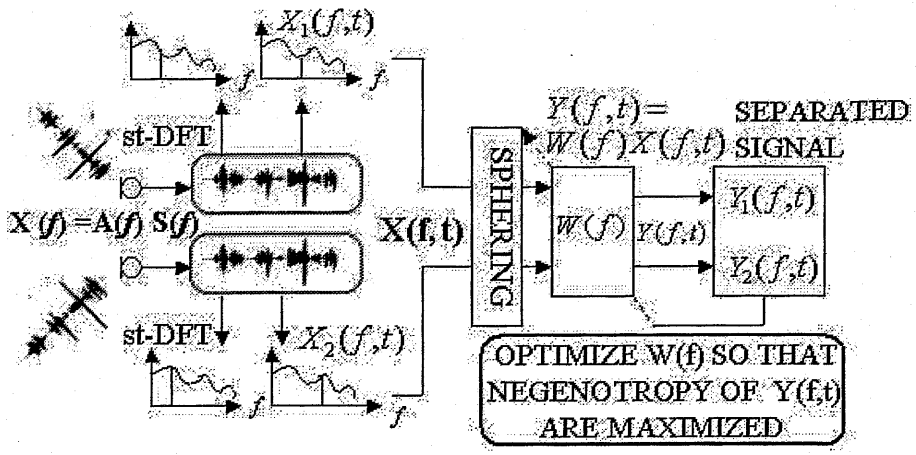


Figure 3.3 Functioning of the fixed-point FDICA for two input channels.

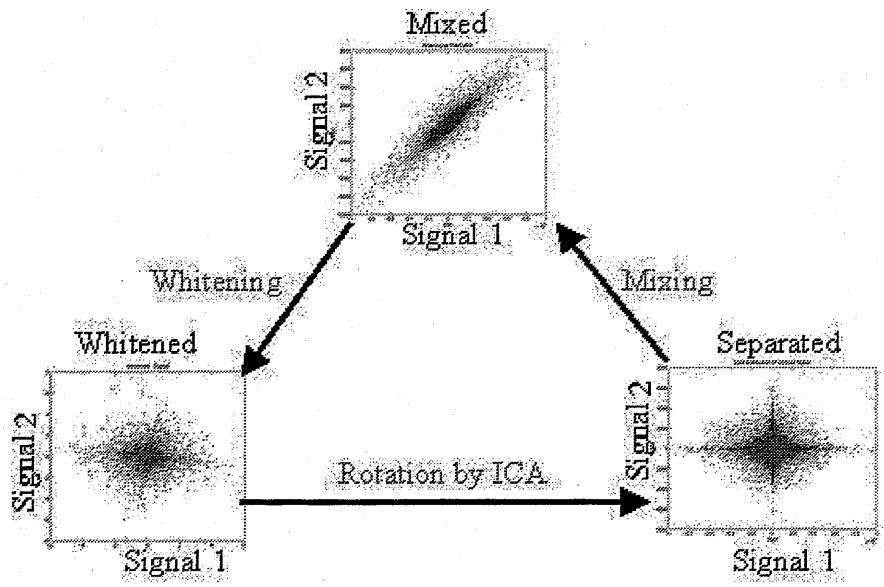


Figure 3.4 Scatter plots showing effects of mixing, whitening and ICA on the speech data distribution.

$$\mathbf{X}_w(f,t) = Q(f)\mathbf{X}(f,t), \tag{3.9}$$

where $Q(f) = \Lambda_x^{-0.5} V_x$ is called whitening matrix; $\Lambda_x = diag\{1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, ..., 1/\sqrt{\lambda_n}\}$ is the diagonal matrix with positive eigen values $\lambda_1 > \lambda_1 > ..... > \lambda_n$ of the covariance matrix of $X(f_p,t)$ and V_x is the orthogonal matrix consisting of eigenvectors. The task remaining after whitening involves rotating the whitened signal vector $X_w(f,t)$ by the separation matrix such that $\mathbf{Y}(f) = W(f)\mathbf{X}_w(f,t)$ equals independent components. The cost function used for measuring the non-Gaussianity is negentropy. The negentropy *J(Y)* of the TFSS of the candidate IC, $\mathbf{Y}(f,t)$ is given by (frequency index *f* and frame index *t* are dropped hereafter for clarity)

$$J(\mathbf{Y}) = H(\mathbf{Y}_{gauss}) - H(\mathbf{Y}) \tag{3.10}$$

where *H(.)* is the differential entropy of (.) and $\mathbf{Y}_{gauss}$ is the Gaussian random variable with the same covariance as of Y. This definition of negentropy ensures that it will be zero if $\mathbf{Y}(f,t)$ is Gaussian and will be increasing if $\mathbf{Y}(f,t)$ is tending towards non-Gaussianity. Thus negentropy based contrast function can be maximized to obtain optimally non-Gaussian component. Here we are placing derivation of such a deflationary learning rule in which one separation vector ***w*** (any one row of the separation matrix) at a time will be learned. The negentropy can be approximated in terms of non-quadratic non-linear function *G* as follows [19]:

$$J(y) = \sigma[E\{G(y) - E\{G(y_{gauss})\}]^2, \tag{3.11}$$

where σ is a positive constant. The performance of the fixed-point algorithm depends on the used non-quadratic non-linear function *G*. The choice of the non-linear function *G* depends on the PDF of the data. Some of the non-quadratic functions used for complex-valued signal separation are

$$G_1(Y) = \sqrt{a_1 + Y}\ ; a_1 = 0.01,$$
$$G_2(Y) = \log(a_2 + Y); a_2 = 0.01, \tag{3.12}$$

The most general form of non-linear function that can be used for speech data (assuming TFSS has super-Gaussian distribution) is $G_2(Y)$. Following findings in [19], we will also use non-quadratic function $G_2(Y)$, whose first and 2nd-order derivatives $g_2(Y)$ and $g_2'(Y)$, respectively, are given by

$$g_2(Y) = \frac{1}{\left(a_2 + Y^2\right)} \text{ and } g_2'(Y) = \frac{0.5}{\left(a_2 + Y^2\right)^2}. \tag{3.13}$$

The one unit algorithm for learning the separation matrix $W(f)$ is obtained by maximizing the negentropy based contrast function. The speech signal is also modeled as a spherically symmetric variable, and as pointed out in [19], for a spherically symmetric variable, modulus-based contrast function can be used to measure non-Gaussianity. Accordingly, we use the same contrast function as in [19] and is given by

$$J(\mathbf{Y}) = E\{G(|\mathbf{w}^H \mathbf{X}_w|^2)\} \tag{3.14}$$

where w is an M-dimensional complex vector such that

$$E\{|\mathbf{w}^H \mathbf{X}_w|^2\} = 1 \Rightarrow |\mathbf{w}| = 1. \tag{3.15}$$

This contrast function may have M local or global optimum solutions $\mathbf{w}_i$ (i=1,2, ..., M) for each source. Thus learning each w calls for the maximization of Eq.(3.14) under the constraint given in Eq.(3.15). The maxima of $J(Y)$ can be found by solving the Lagrangian function $L(\mathbf{w}, \mathbf{w}^H, \lambda)$ of the above, given as

$$L(\mathbf{w}, \mathbf{w}^H, \lambda) = E\{G(|\mathbf{w}^H \mathbf{X}_w|^2)\} \pm \lambda[E\{\mathbf{w}^H \mathbf{X}_w\} - 1\}, \tag{3.16}$$

where λ is Lagrangian multiplier. In order to locate maxima of the contrast function, the following simultaneous equations must be solved.

$$\frac{\partial L}{\partial \mathbf{w}} = 0; \frac{\partial L}{\partial \mathbf{w}^H} = 0; \text{ and } \frac{\partial L}{\partial \lambda} = 0 \tag{3.17}$$

These equations can be obtained from Eq.(3.16) as follows

$$\frac{\partial L}{\partial \mathbf{w}} = E\{g(|\mathbf{w}^H \mathbf{X}_w|^2)\mathbf{w}^H\} + \lambda \mathbf{w}^H = 0, \tag{3.18}$$

$$\frac{\partial L}{\partial \mathbf{w}^H} = E\{g(|\mathbf{w}^H \mathbf{X}_w|^2)\mathbf{X}_w^H \mathbf{w}\} + \lambda \mathbf{w} = 0, \tag{3.19}$$

$$\frac{\partial L}{\partial \lambda} = |\mathbf{w}|^2 - 1 = 0, \tag{3.20}$$

From here, we proceed further in the light of following two theorems [47]:

THEOREM 1: *If function f(z,z*) is analytic with respect to z and z*, all stationary points can be found by setting the derivative with respect to either z or z*.*

THEOREM 2: *If f(z,z*) is a function of the complex-valued variable z and its conjugate, then by treating z and z* independently, the quantity directing the maximum rate of change of f(z,z*) is* $\nabla z^* f(z)$

Accordingly, the final solution using Newton's iterative method is given by

$$\mathbf{w}_{new} = \mathbf{w} - \left[\frac{\partial L}{\partial \mathbf{w}^H}\right]\left[\frac{\partial L}{\partial \mathbf{w}}\left(\frac{\partial L}{\partial \mathbf{w}^H}\right)\right]^{-1}. \tag{3.21}$$

which can be further simplified into

$$\mathbf{w}_{new} = \mathbf{w}(E\{g(|\mathbf{w}^H\mathbf{X}_w|^2) + (|\mathbf{w}^H\mathbf{X}_w|^2)g'(|\mathbf{w}^H\mathbf{X}_w|^2)\}) - E\{g(|\mathbf{w}^H\mathbf{X}_w|^2)(\mathbf{X}_w^H\mathbf{w})\mathbf{X}_w\}. \quad (3.22)$$

The stopping criterion for iteration is defined as $\delta = (|\mathbf{w}_{old} - \mathbf{w}_{new}|)^2$, which becomes very small near the convergence. Since each update changes the norm of w, after each iteration separation vector w for each source is normalized as follows to maintain compliance with Eq.(3.15)

$$\mathbf{w}_{new} = \frac{\mathbf{w}_{new}}{|\mathbf{w}_{new}|} \quad (3.23)$$

As this is a deflationary algorithm, independent sources are extracted one by one in the decreasing order of negentropy from the mixed signal. Thus after each iteration, it is also essential to decorrelate w to prevent its convergence to the previously converged point. In order to achieve this, Gram-Schmidt sequential orthogonalization can be used, in which components of all previously obtained separation vectors falling in the direction of the current vector are subtracted. Accordingly, the orthogonalized separation vector $\mathbf{w}_i$ for the ith source after jth iteration is given by

$$\mathbf{w}_i = \mathbf{w}_i - \sum_{j=1}^{i-1}(\mathbf{w}_i^T\mathbf{w}_j)\mathbf{w}_j. \quad (3.24)$$

The update Eq.(3.22) is used to estimate separation vector w in each frequency bin from whitened TFSS of mixed signal for each source in the deflationary fashion and separation matrix $W(f)$ in any frequency bin f is given by

$$\mathbf{W}(f) = \begin{bmatrix} \mathbf{w}_1 \\ . \\ \mathbf{w}_R \end{bmatrix} = \begin{bmatrix} W_{11}(f) & .. & W_{1M}(f) \\ . & .. & . \\ W_{R1}(f) & .. & W_{RM}(f) \end{bmatrix} \quad (3.25)$$

Each row of this separation matrix uniquely corresponds to a separation vector w

for a particular source. Because this separation matrix has been obtained using whitened signals, its pre-multiplication with whitened signals in the frequency domain gives the TFSS $\mathbf{Y}(f,t)=[\mathbf{Y}_1(f,t),\mathbf{Y}_2(f,t),.....\mathbf{Y}_R(f,t)]^T$ of the separated signal, i.e.,

$$\hat{S}(f,t)=\mathbf{Y}(f,t)=\mathbf{W}(f)\mathbf{X}_W(f,t). \tag{3.26}$$

3.4. Permutation and Scaling Problem

In order to get separated signal correctly, the order of separation vectors (position of rows) in $\mathbf{W}(f)$ must be same in each frequency bin. The deflationary algorithm separates original sources in the decreasing order of negentropy. But the order of negentropy for the independent sources does not remain same, due to change in contents, in all frequency bins which in turn leads to the inter-exchange or flipping of rows of $\mathbf{W}(f)$ in an unknown order. This is called permutation problem. The other problem is related with different gain values in each frequency bin. However, for the faithful reconstruction of the signal it should be same. This is called scaling problem. If these two problems are not solved, Eq.(3.26) will give another mixed signals instead of separated components. There have been developments of several methods to resolve these two inherent problems [48]. However, we will use here Directivity Pattern (DP) based method using null beamformer [30] for the reason explained in the following section. The DP based method requires the DOA of each source to be known. In the totally blind setup, this cannot be known so it is estimated from the directivity pattern of the separation matrix. The DP $F_R(f,\theta)$ of the microphone array in the Rth source direction is given by [30]

$$F_R(f,\theta)=\sum_{k=1}^{M} W_{Rk}^{(JCA)}(f)\exp[j2\pi d_k \sin\theta/c], \tag{3.27}$$

where $W_{Rk}^{(JCA)}(f)$ is an element of the separation matrix obtained in Eq.(3.25), R=1,2. The DP of the separation matrix contains nulls in each source direction.

However, the positions of the nulls vary in each frequency bin for the same source direction. Hence by calculating the null directions in each frequency bin, the DOA of the Rth source can be estimated as

$$\hat{\theta}_R = \frac{2}{N}\sum_{p=1}^{N/2}\theta_R(f_p), \tag{3.28}$$

where $\theta_R(f_p)$ denotes the direction of null in the pth frequency bin. For the present case of two sources, these are given by

$$\begin{aligned}\theta_1(f_p) &= \min\left[\arg.\min_{\theta}|F_1(f_p,\theta)|,\ \arg.\min_{\theta}|F_1(f_p,\theta)|\right],\\ \theta_2(f_p) &= \max\left[\arg.\min_{\theta}|F_1(f_p,\theta)|,\ \arg.\min_{\theta}|F_1(f_p,\theta)|\right],\end{aligned} \tag{3.29}$$

where $\min[u,v]$ and $\max[u,v]$ are defined to choose minimum and maximum, respectively, from u and v. Then the separation matrix in each frequency bin is arranged in accordance with the directions of nulls, which sort-out the permutation problem. After estimating DOA, the gain value in each frequency bin is normalized in each source direction. The separation matrix normalized in this way will have unit gain in the target source direction and negative gain in the source direction to be jammed. However, it will have adequate gain in other directions which will be harmful in the reverberant conditions. Gain in the Rth source direction in the pth frequency bin is given by

$$\alpha_R(f_p) = \frac{1}{F_R(f_p,\hat{\theta}_R)} \tag{3.30}$$

where $\hat{\theta}_R$ is the estimated direction of the Rth source which can be obtained from Eq.(3.28) or by the histogram of the directivity pattern, as proposed in [29]. Thus, de-permuted and scaled separation matrix is given by

$$\mathbf{W}(f_p)=\begin{bmatrix}\alpha_1(f_p)0.. & 0\\ : \quad :: & 0\\ 0 \quad 0..\alpha_R(f_p)\end{bmatrix}\begin{bmatrix}W_{11}(f_p) & .. & W_{1R}(f_p)\\ . & .. & .\\ W_{M1}(f_p) & .. & W_{MR}(f_p)\end{bmatrix} \tag{3.31}$$

This scaled and depermuted matrix is used to separate the signals in each frequency bin. Then by using overlap-add technique [49] time-domain signal is reconstructed from the TFSS of each source. However, in order to use **W**(f) of Eq.(3.25) in the time domain to form an FIR filter, it is essential to de-whiten the separation filter as follows:

$$\mathbf{W}(f) = \mathbf{W}(f)(\mathbf{Q}(f))^{-1}. \tag{3.32}$$

Then using de-whitened **W**(f), an FIR filter of length P can be formulated to separate the signals directly in the time-domain as follows

$$y(n) = \sum_{r=0}^{P} w(r)x(n-r). \tag{3.33}$$

3.5. Algorithm initialization

The deflationary learning rule for **w** in Eq.(3.22) is sensitive to the initial value of separation vector **w**. It can be initialized by a random value or some heuristically chosen good guess values such as NBF-based initial value. NBF is a geometrical technique for the speech signal separation in which the separation filter depends on the DOA, frequency of the signal and the geometry of the used microphone array. NBF jams signals from the undesired directions by forming nulls in DP in that directions while setting look direction in the direction of desired signal source. Accordingly, DP in Eq.(3.27) for the NBF based separation matrix $W^{BF}(f)$ for the look direction $\hat{\theta}_1$ and null direction $\hat{\theta}_2$ should satisfy the following conditions

$$F_1(f,\hat{\theta}_1)=1 \text{ and } F_1(f,\hat{\theta}_2)=0 \tag{3.34}$$

These simultaneous equations can be solved to give the following solutions for the elements of separation matrix $W^{BF}(f)$

$$W_{11}^{BF}(f)=-\exp[-q_1\sin\hat{\theta}_2]\times\{-\exp[q_1(\sin\hat{\theta}_1-\sin\hat{\theta}_2)\times\exp[q_2(\sin\hat{\theta}_1-\sin\hat{\theta}_2)]\}^{-1}. \tag{3.35}$$

and

$$W_{12}^{BF}(f)=-\exp[-q_2\sin\hat{\theta}_2]\times\{-\exp[q_1(\sin\hat{\theta}_1-\sin\hat{\theta}_2)\times\exp[q_2(\sin\hat{\theta}_1-\sin\hat{\theta}_2)]\}^{-1}. \tag{3.36}$$

Similarly, for the look direction $\hat{\theta}_2$ and null direction $\hat{\theta}_1$ following conditions are satisfied by the elements of separation matrix $\mathbf{W}^{BF}(f)$

$$F_2(f,\hat{\theta}_1)=0 \text{ and } F_2(f,\hat{\theta}_2)=1. \tag{3.37}$$

On solving these, the following solutions are obtained

$$W_{21}^{BF}(f)=-\exp[-q_1\sin\hat{\theta}_1]\times\{-\exp[q_1(\sin\hat{\theta}_2-\sin\hat{\theta}_1)]-\exp[q_2(\sin\hat{\theta}_2-\sin\hat{\theta}_1)]\}^{-1}, \tag{3.38}$$

and

$$W_{22}^{BF}(f)=-\exp[-q_2\sin\hat{\theta}_1]\times\{-\exp[q_1(\sin\hat{\theta}_2-\sin\hat{\theta}_1)]-\exp[q_2(\sin\hat{\theta}_2-\sin\hat{\theta}_1)]\}^{-1} \tag{3.39}$$

where $q_1=j2\pi d_1 f/c$ and $q_2=j2\pi d_1 f/c$, c=velocity of sound in given environment. The NBF based separation matrix is approximately optimal and is derived for ideal far-field propagation of acoustic wave. However, under the reverberant condition, its separation performance degrades markedly.

3.6. TFSS and Central Limit Theorem (CLT) Compliance

The most important thing from here that can be concluded about the PDF of $X_i(f)$ is that it is convolution of PDF of $S_i(f)$. From Eq.(3.7) the signal received at ith microphone is given by

$$X_i(f) = H_{i1}(f)S_1(f) + H_{i2}(f)S_2(f) = Y_{i1}(f) + Y_{i2}(f). \tag{3.40}$$

where $Y_{i1}(f)$ and $Y_{i2}(f)$ represents respectively contribution of first and second source in frequency bin f at ith microphone. For simplicity in writing if the PDF of The PDF $f_{X_i(f)}(x_i)$ of $X_i(f)$ is given by convolution of PDF of $Y_{i1}(f)$ and $Y_{i2}(f)$ as follows

$$f_{X_i(f)}(x_i) = \int_{-\infty}^{\infty} f_{Y_{i1}(f)}(y_{i1}) f_{Y_{i2}(f)}(y_{i2} - y_{i1}) dy_{i1}. \tag{3.41}$$

This simple addition of contribution of each signal in each frequency bin pushes

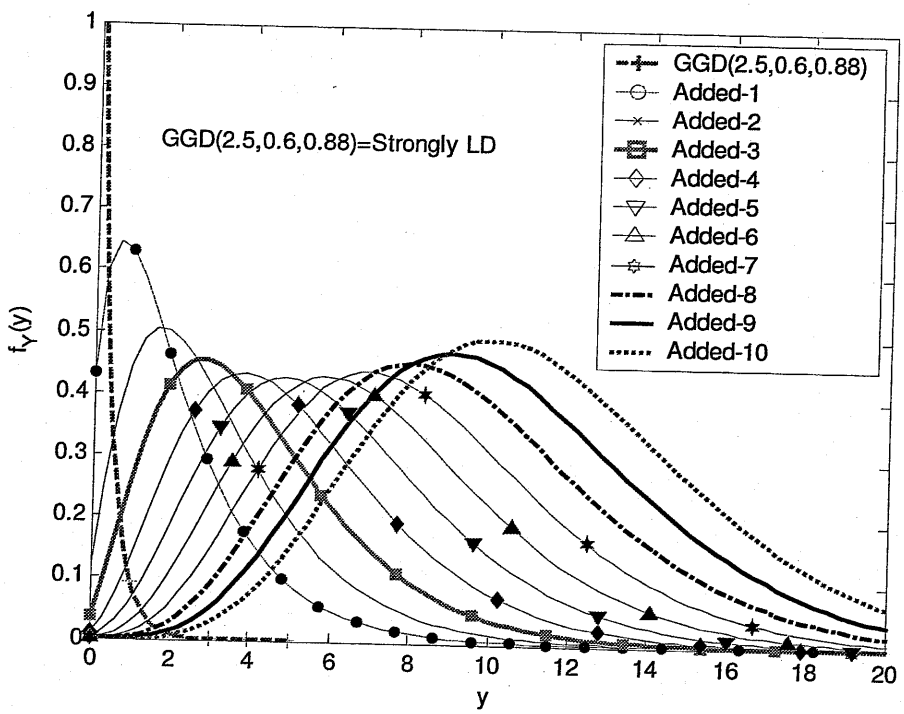


Figure 3.5 Showing effect of CLT by summing 10 strongly Laplacian distributions (shown by dashed line with GGD (mean, scale, shape)). New PDFs are obtained using Eq.(3.41).

the nature of distribution towards the Gaussian distribution under the light of CLT. It can also be imbued from here that if the signal contribution by any speaker in any frequency bin is insignificant the movement towards Gaussianity will be also insignificant and separation by non-Gaussianization will be poor. For example, **Figure 3.5** shows how the distribution of a strongly Laplacian distribution with unit variance changes when added to itself. As the number of addition increases the resulting distribution becomes more and more smooth.

The fixed point FDICA by negentropy maximizations extracts TFSS of independent sources by the non-Gaussianization. For the effective functioning of the fixed-point FDICA it is essential that the TFSS of the mixed speech signal should be more Gaussian than that of the independent components. It is evident from Eq.(3.7) that the TFSS of mixed signal in any frequency bin is a superposition of the spectral contributions of all mixing signals in the same frequency bin. This is the mathematical reason for the Gaussianization of the mixed signal. Thus the power, to separate ICs, comes in the algorithm due to the validity of the following logical fact

Gaussianity of the mixed speech signal> Gaussianity of the independent speech signals.

If the above fact is not followed, it will be against the very basic working principle of the algorithm and hamper the performances of algorithm as is shown in [31]. One of the easiest mathematical translations of the above logical touchstone can be done in terms of kurtosis. Accordingly, validity of CLT can be checked by computing and comparing the kurtosis of the TFSS of the mixed signal and reference signals in each frequency bin. The kurtosis of spectral component in each frequency bin, denoted hereafter as spectral kurtosis (SK), is given as the ratio of the fourth order central moment to the second order moment [50][51]. Accordingly *SK(f)* in frequency bin f is given by

$$SK(f)=\frac{C_4\{S^*,S^*,S^*,S^*\}}{[C_2\{S^*,S^*\}]^2}, \tag{3.42}$$

where $S^* \in \{\mathbf{X}(f,t), \mathbf{X}^H(f,t)\}$. This definition varies with the placement of conjugates [52] but following [53][67] and assuming spectral component of speech as complex conjugate random variable simplified expression for SK is given by

$$SK(f)=\frac{E\{|X(f)|^4\}-2E^2\{|X(f)|^2\}}{[E\{|X(f)|^2\}]^2}. \tag{3.43}$$

As in the fixed point algorithm, data are sphered so that Eq.(3.43) further simplifies to

$$SK(f)=E\{|X(f)|^4\}-2. \tag{3.44}$$

The aforementioned condition for Gaussianity of the mixed data can be satisfied by verifying the following conditions in terms of SK

$$\begin{aligned} SK_{m1}(f) &< \min\{SK_{ref_{11}}(f), SK_{ref_{12}}(f)\}, \\ SK_{m2}(f) &< \min\{SK_{ref_{21}}(f), SK_{ref_{22}}(f)\}, \end{aligned} \tag{3.45}$$

where $SK_{mi}=SK$ of mixed signal at the ith microphone.

Using the expressions for SK in Eq. (3.43) or(3.44), the validity of the CLT can be tested in each frequency bin. However, this method is not blind because it requires reference signals which are not available in the real applications.

3.7. Objective Evaluation Score

In order to evaluate the performance of the algorithm Noise Reduction Rate (NRR), Spectral NRR (SNRR), and Spectral Correlation Coefficient (SCRF) $\gamma(f)$ have been used. NRR is defined as ratio of speech signal power (computed from reference signal) to the noise power. SNRR is given as NRR in any frequency bin. SNRR for the ith source (here $M=R=2$) in the fth frequency bin is given by

$$SNRR_i(f) = 10\log_{10}\frac{E\{|W_{i1}(f)ref_{1i}(f)+W_{i2}(f)ref_{2i}(f)|^2\}}{E\{|Y_i(f)-W_{i1}(f)ref_{1i}(f)+W_{i2}(f)ref_{2i}(f)|^2\}}, \quad (3.46)$$

SCRF between ICs $Y_1(f)$ and $Y_2(f)$ in a frequency bin f is given by

$$\gamma(f)=\frac{\sum_{1}^{m}[\{Y_1(f)-\bar{Y}_1(f)\}\{Y_2(f)-\bar{Y}_2(f)\}]}{\sqrt{\sum_{1}^{m}|Y_1(f)-\bar{Y}_1(f)|^2}\sqrt{\sum_{1}^{m}|Y_2(f)-\bar{Y}_2(f)|^2}}. \quad (3.47)$$

3.8. Experiments and Results

The layout of experimental room is shown in Figure 3.6. The spacing between two microphone was kept at 4 cm. Voices of two male and two female speakers, at the distances of 1.15 meters and from the directions of –30° and 40° were used to generate 12 combinations of mixed signals x_1 and x_2 under the described convolutive mixing model. Mixed signals at each microphone were obtained by adding speech signals ref_{11}, ref_{12}, ref_{21}, ref_{22}. The speech signals ref_{11}, ref_{12}, ref_{21}, and ref_{22} reaching each microphone from each speaker are used as the reference signals. These speech signals were obtained by convolving seed speech with room impulse response, recorded under different acoustic conditions, which are characterized by a different Reverberation Time (RT), e.g., RT=0 ms, RT=150 ms and RT=300 ms. First of all STFT analysis of the mixed data is done to obtain TFSS. The STFT analysis conditions are shown in the Table 3.1. The TFSS data in each frequency bin are whitened in accordance with Eq.(3.9) before being fed into iterative ICA loop. As explained in the previous sections whitening is only half ICA, the whitened data are used to learn separation vector in accordance to Eq.(3.22). At first the algorithm is initialized using random values of separation vector **w** in each frequency bin. Algorithm learns separation vector in each frequency bin. The algorithm begins to converge after 20 iterations(less for RT=0 ms) for RT=300 ms and stops when the stopping criterion is satisfied. The convergence curves for RT=0 ms and RT=300 ms are shown in figures (a) and (b) respectively of Figure 3.8.

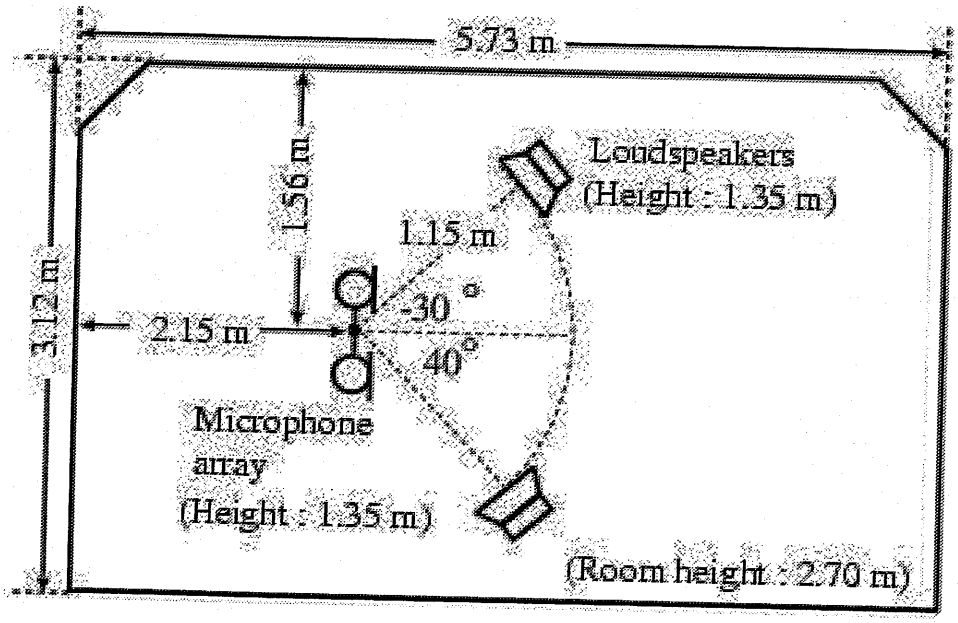


Figure 3.6 Layout of the experimental setup.

The stopping criterion δ was fixed at 0.001.Using directivity-pattern-based methods, DOAs of the sources are estimated. The DOAs of the 1st source s_1 and 2nd source s_2, estimated using Eq.(3.28), are presented in Table 3.2 along with true DOAs. The histograms of Direction of Nulls (DON) formed by the separation matrix are shown in Figure 3.7. It is evident from there that in all frequency bins DON are not in the same direction. In some frequency bins it is swapped with the DOA of other sources indicating that separation matrix is permuted, however, maximum no. of nulls are occurring in a particular source direction, shown as white bar in Figure 3.7, and hence this can also be used as the DOA information.Using DOA information, the separation matrix is scaled using Eq.(3.31). The DP of the separation matrix before and after de-permutation and scaling are shown in Figure 3.9. That figure shows how the directional nulls of the separation matrix get blurred with increasing RT resulting in poor separation. After solving the permutation and scaling problem the DP of separation matrix shows unity gain in the look direction and nulls in the direction of source to be rejected.

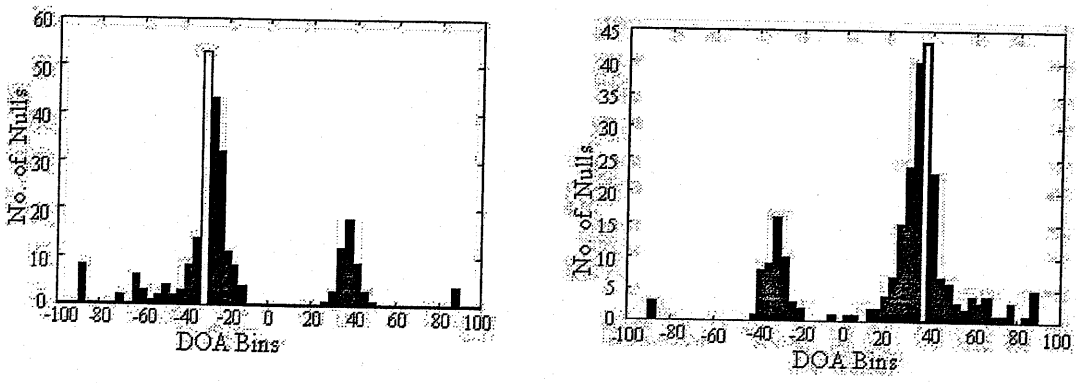
Table 3.1. Signal analysis conditions

Sampling freq.	8000 Hz
Frame Length	20 ms
Step Size ε	10 ms
Window	Hanning
FFT length	512
δ	0.001

In order to evaluate the performance of the algorithm with NBF based initialization, the initial value of **w** is generated for every frequency bin using the estimated DOA and Eq. (3.35)-(3.39). Using these initial values in each frequency bin, ICA is performed. The NRR results under both initializations are shown in Figure 3.10. There occurs severe degradation in the separation performance with the increasing reverberation time in both cases. It is also evident from Figure 3.10 that the NRR improvements for the non-reverberant case are almost same for the both types of initializations. However, for reverberant conditions, NBF- based guess value shows better performance in the NRR as well as in the convergence speed, see Figure 3.11, over random initialization. In order to study the effect of over-iteration on the separation performance, NRRs for the different number of iterations for both the NBF based initialization and random value based initialization were observed under different RTs. The average NRR versus

Table 3.2 DOA Estimation result

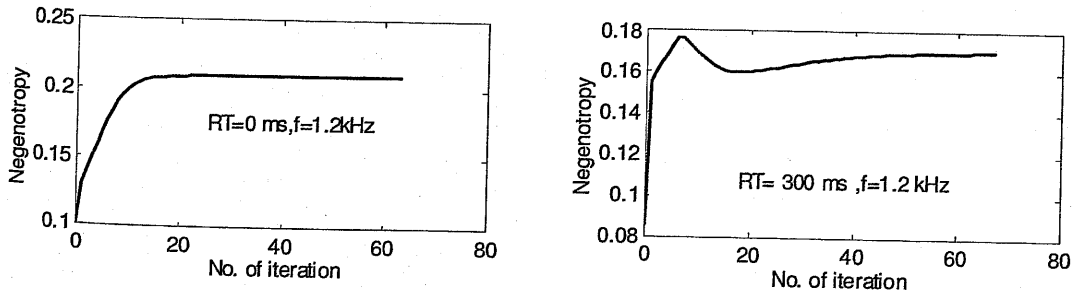
RT→	RT= 0 ms		RT=150ms		RT=300 ms	
Sources→	S_1	S_2	S_1	S_2	S_1	S_2
Est.DOA	-31.1	40.0	-32.2	39.0	-28.1	42.1
True DOA	-30	40	-30	40	-30	40



(a) **(b)**

Figure 3.7 Estimated DOAs from directivity patterns of the separation matrix. These figures show histogram of DON in DP of the separation matrix for male-female speaker combination. Permutation can be observed as the DON formed by separation vector are available in both source direction, however, maximum no. of nulls (shown as white bar) are available in a particular source direction

number of iterations for RT=150 ms and RT=300 ms are shown in Figure 3.12. The maximum iteration limit was set at 1000. It is evident from that figure that NRR performance is slightly changed by over-learning and NBF based initialization results in better performance than that of the random value based



(a) **(b)**

Figure 3.8 Convergence of the algorithm for the source combination male and female, f=1.2kHz,(a) RT=0 ms (b), RT=300 ms.

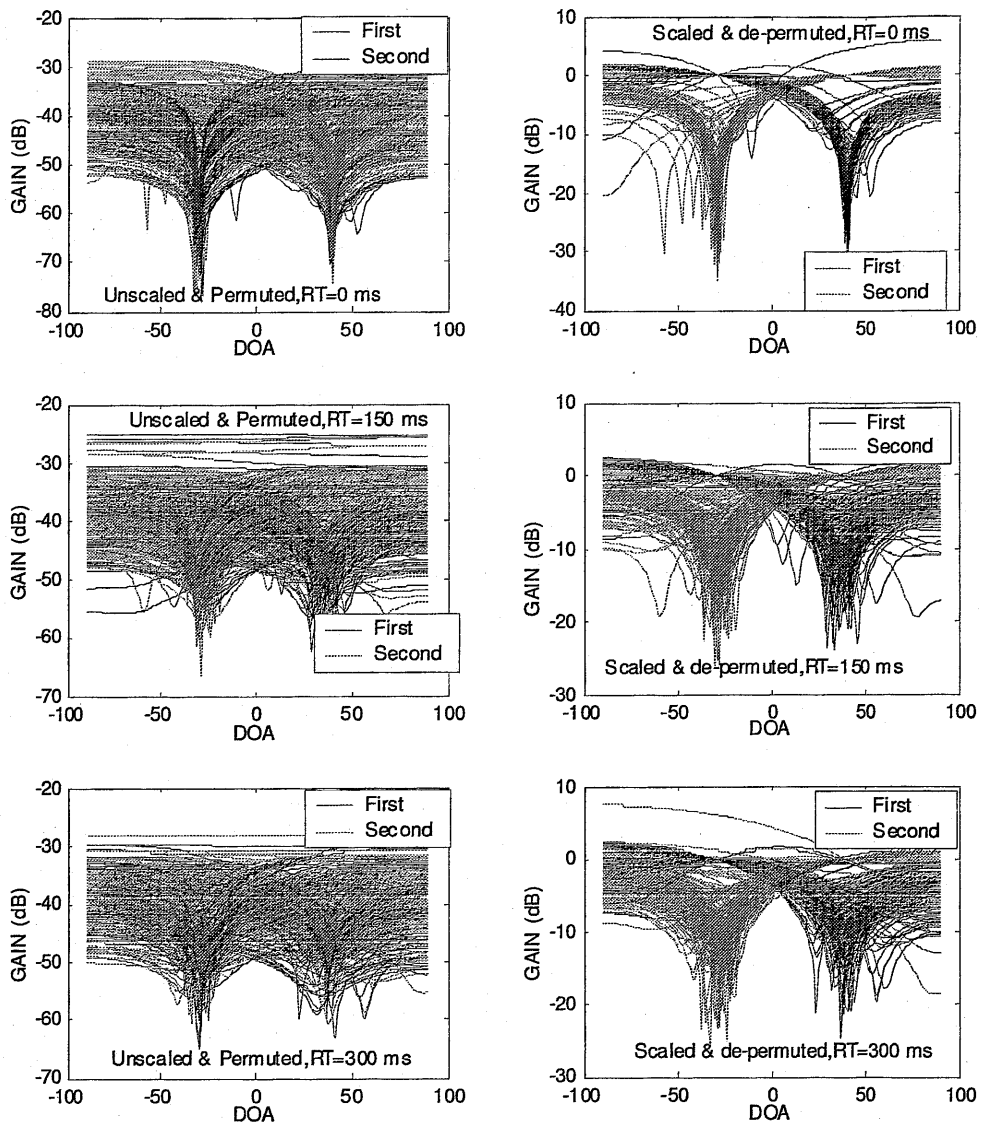


Figure 3.9 DP of the ICA based separation matrix obtained under different reverberation time. The left-hand side is unscaled and permuted and right-hand side figures represent DP for the scaled and permuted separation matrix. Under no reverberation nulls are sharp and clear resulting in good separation. For moderate or high reverberation directional nulls are blurring which results in poor separation.

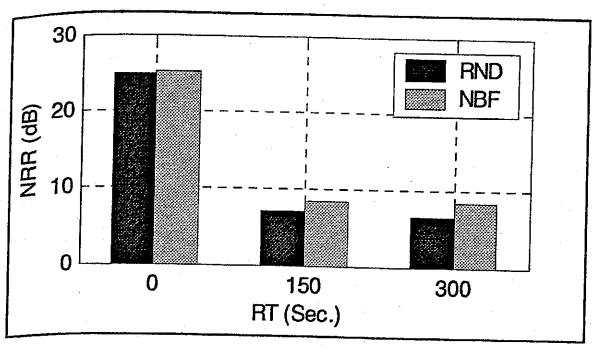


Figure 3.10 NRR improvement using NBF based and random initial value for w in different acoustic environment.

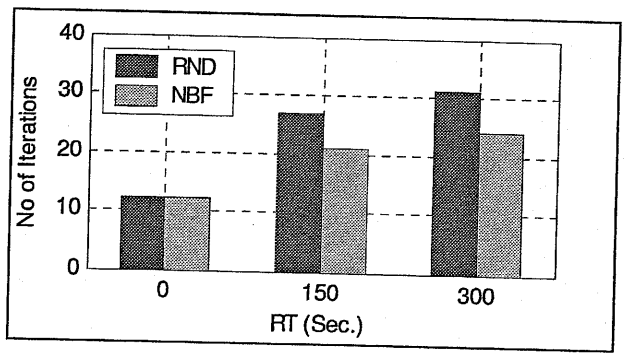


Figure 3.11 Average no of iteration consumed in extracting both sources under NBF and random (RND) value based initialization for different RT.

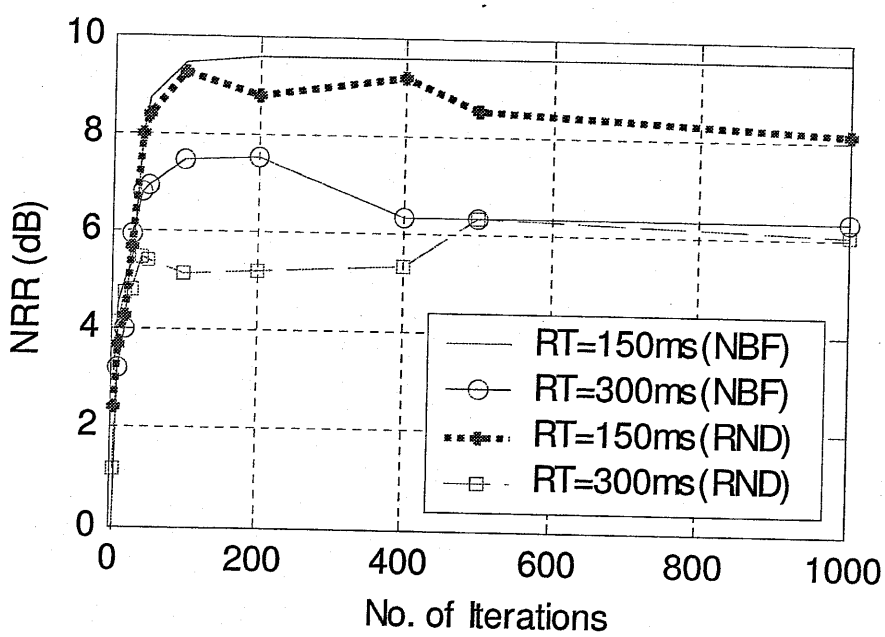


Figure 3.12 Effect of over-iteration on the NRR performance.

initialization. In order to see the performance of the algorithms in each frequency bins spectral NRR, defined in Eq.(3.46), and correlation coefficients between the separated components, as defined in Eq.(3.47), were studies. Since the algorithm separates the signals independently in each frequency bins, the separation performance in each frequency bin is important. It has been found that the separation quality in each frequency bin is not same, however, TFSS in each frequency bin is assumed to be independent. SNRR for the male female speaker combinations for RT=0 ms, RT=150 ms, and for RT=300ms are shown in Figure 3.13, Figure 3.14, and Figure 3.15. It is evident from these figures that the separation performance in different frequency bins is unexpectedly uneven. In Figure 3.16 the spectral correlation coefficient between the separated components is shown which also indicates different degree of separation in each frequency bin. Since the TFSS in each frequency bin is assumed to be independent such unevenness in performance is unexpected. Factor responsible for this may be the difference in nature of data, as while the other experimental conditions are same, in different frequency bins. Among the other statistics of the TFSS Gaussianity of the TFSS of mixed data is important for the proper working of algorithm. This is discussed next along with more experimental results. One of the possible causes of such behaviors in SNRR has been discussed in next sub-section.

In order to study the effect of different DFT size and frame shift sizes, further experiments were performed with random and NBF based initialization. The analysis frame size was fixed at 20 ms, which contains 160 samples of data at a sampling frequency of 8000 Hz, and the frame shift size has been varied from 10% to 80% of the analysis frame size. The results of achieved NRR and consumed computation power (number of iterations consumed for fixed δ) are shown in Figure 3.17. The obvious benefit of the NBF based initialization over random value based initialization is rapid convergence. This is natural because Newton-Raphson method is well-known for its sensitivity to initial value in finding the solution and NBF provides one of the ideal or highly optimized separation matrixes under no reverberation.

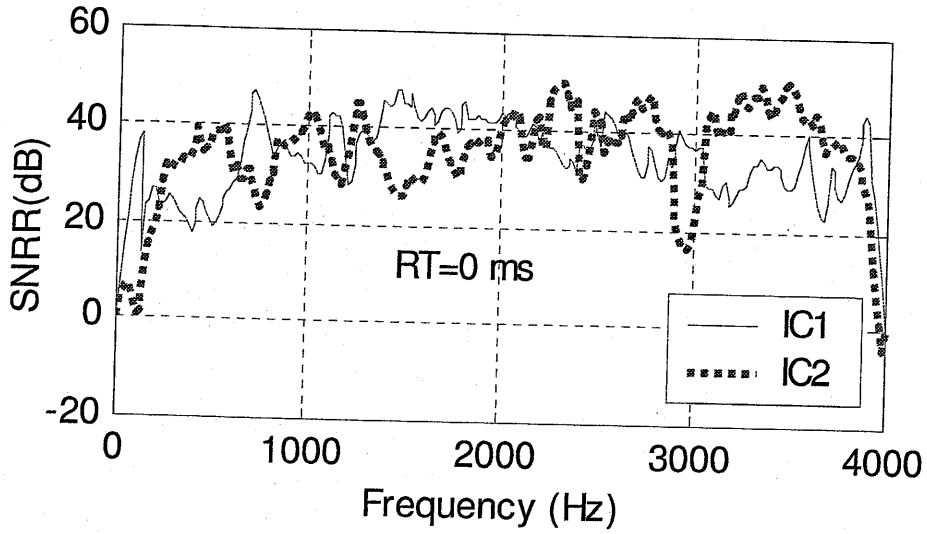


Figure 3.13 SNRR for RT=0 ms for male female speaker combination.

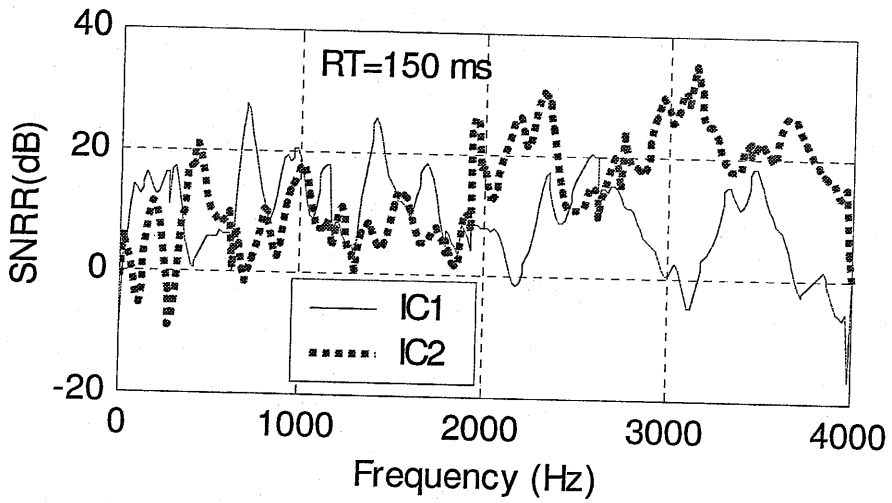


Figure 3.14 SNRR for RT=150 ms for male female speaker combination.

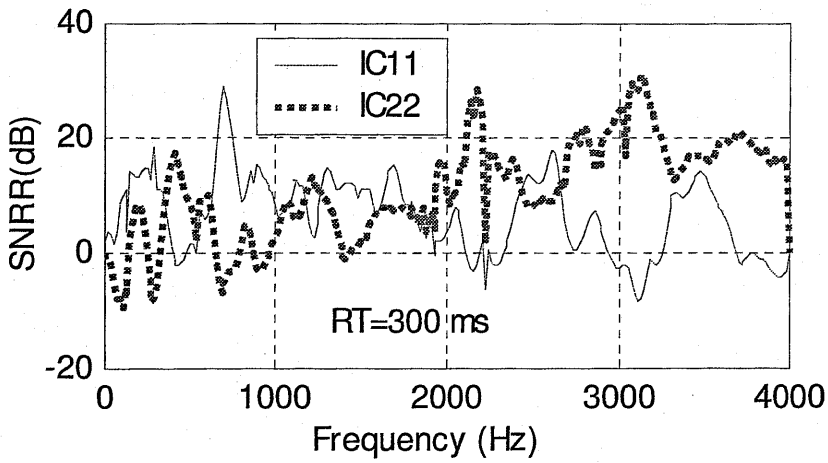


Figure 3.15 SNRR for RT=300 ms for male female speaker combination.

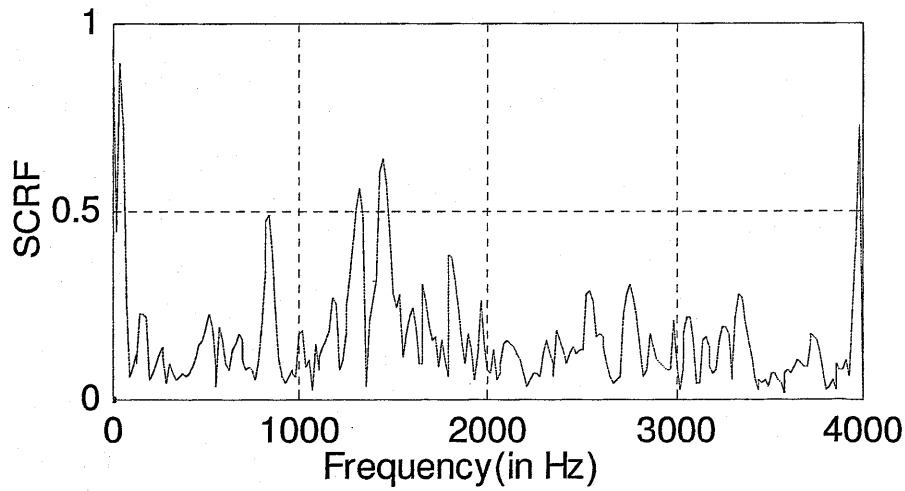


Figure 3.16 SCRF for RT=300 ms for male female speaker combination.

Experimental Results of CLT Compliance Test:

The validity of CLT in TFSS of any frequency bin can be checked by verifying the relation given in Eq.(3.45) for the CLT compliance test. That test was performed for the speech data for the six combinations of mixed data for different DFT sizes and RTs. The related results are shown in Figure 3.18. It is interesting to note that the TFSS does not follow the CLT in every frequency bin. The percentage of CLT disobeying TFSS is almost independent of the DFT size

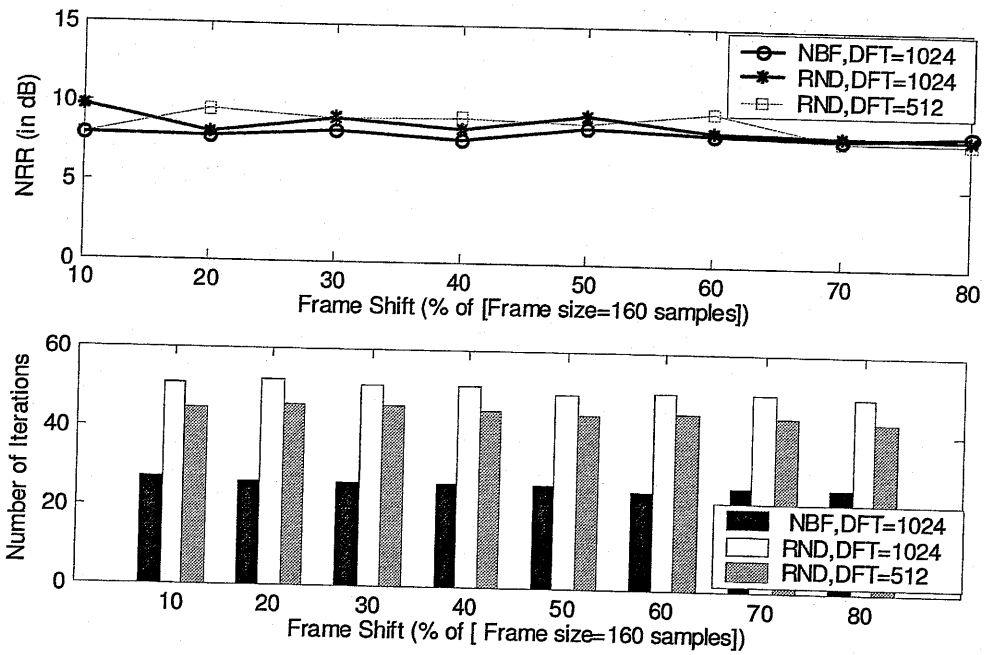


Figure 3.17 NRR and number of iterations consumed by the FDICA algorithm for different values of the DFT size, frame size and frame shift size. RND (NBF) indicates random initial value (NBF based initial value) for w was used. (Result is averaged for six speaker combinations).

and there are no significant changes with the change in reverberation time. However, for the higher values of RT a significant difference in the percentage of CLT-failing sub-bands has been found, as shown in Figure 3.19 and Figure 3.20, for both microphones. This is indicative of the fact that the room acoustics is also influential in the disobedience of CLT by the TFSS. As the DFT size increases, the number of CLT-disobeying bins does increase, however, they remain clustered. This is shown in Figure 3.21, and that is happening due to an increase in the frequency resolution for higher DFT sizes. In order to explain this interesting phenomenon we take into consideration the contribution of each signal source in the mixing process, as it is evident from Eq.(3.7) that TFSS in each frequency bin is a superimposition of spectral contribution from each mixing source and this is the cause of Gaussianization. For this the spectral content of the mixed signal and reference signals were examined in the CLT-disobeying frequency bin and in the nearest CLT-obeying frequency bins. In order to measure the spectral contribution, plots of the magnitude of the spectral contribution from each of the reference signals and the mixed signal were examined, and one of such plots is shown in the Figure 3.22. In that figure, the temporal contribution of each source in a CLT non-complying frequency

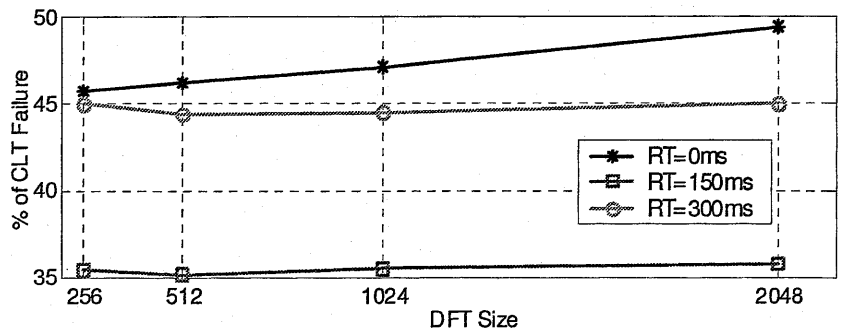


Figure 3.18 CLT-disobeying bins for different DFT size and reverberation time at Mic1. Shown values are averaged from 6 mixed speech data.

sub-band and the nearest CLT complying frequency bin are shown. It is evident from this figure that in the shown CLT-failing frequency bin, the contribution from the first speaker is not available at all instances, however, in the CLT-obeying frequency bin its temporal contribution is relatively better. It is also evident that in the CLT obeying bins both sources make a rich contribution but in the CLT-disobeying bin either one make a very rare contribution or no contribution, which in accordance with Eq.(3.7) results in a mixed signal with content from either source. The resulting TFSS thus in reality contains a signal from single source and thus fails to comply with CLT. It is, therefore, concluded that the sparseness in the spectrum has an important role in relation to the CLT non-compliance. It is also important to note that only spectral sparseness cannot be considered to be the sole cause of CLT disobedience. The role of other causes such as room acoustics, natural pauses (this also results in spectral sparseness in the temporal queue of TFSS) cannot be denied. Since TFSS is generated by the STFT analysis it can be inferred that unless there are no long pauses in the speech, it cannot contribute a large number of dumb samples to the TFSS in any frequency bin. In the presence of moderate reverberation, the pause period may be modified by the reflected speech. Such reflected speech increases correlation only among the samples of TFSS, and the spectral content of the signal remains the same even under high reverberation, but if there is any role of pauses in the CLT failure it will be modified by the reverberation. However, such possibilities are still unexplored and are left for further study. In order to show the effect of the CLT disobedience by the TFSS on the separation performance, spectral NRR and SCRF were observed for different source combinations. Such results for one of the source combinations are shown in Figure 3.23 to Figure 3.26. It is evident from these figures that in the CLT-disobeying frequency bins SNRR is low and SCRF is high. This occurs because TFSS in such frequency bins do not comply with CLT. It is interesting to note that there have been development of ICA algorithm which exploits the temporal absence and existence of signal from different speakers for the blind source separation [55] in the anechoic environment, however, spectral sparseness is problematic for FDICA based on

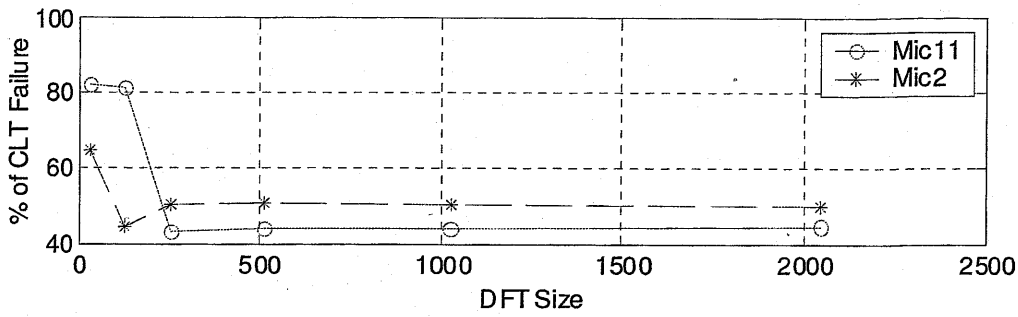


Figure 3.19 CLT failure at both microphones for RT=0 ms.

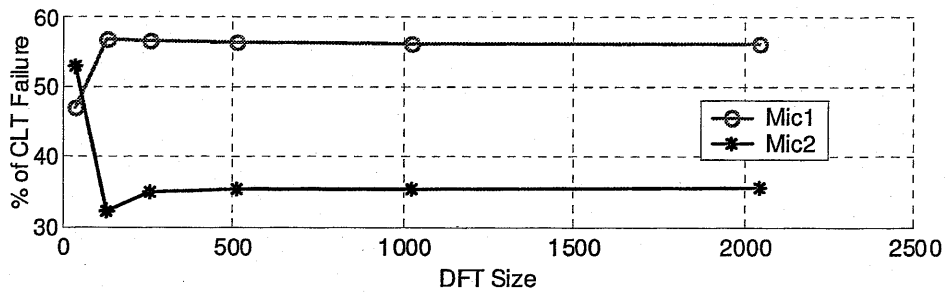


Figure 3.20 CLT failure at both microphones for RT=300 ms.

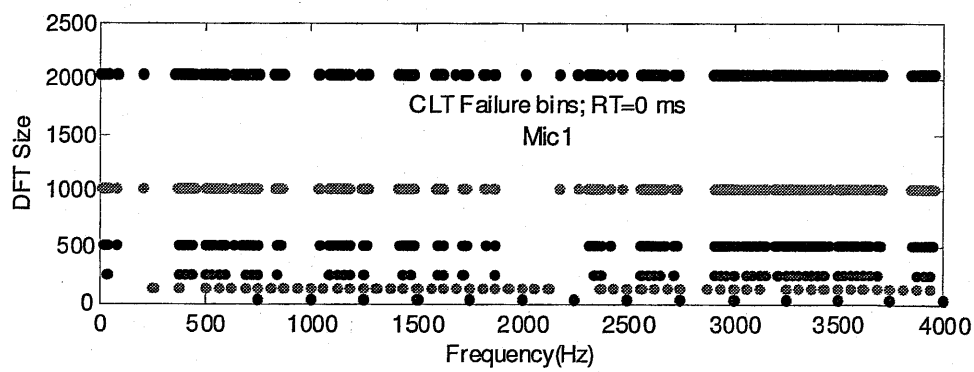


Figure 3.21 Clustering of CLT-disobeying TFSS for different DFT sizes for speech signal picked up by Mic.1.

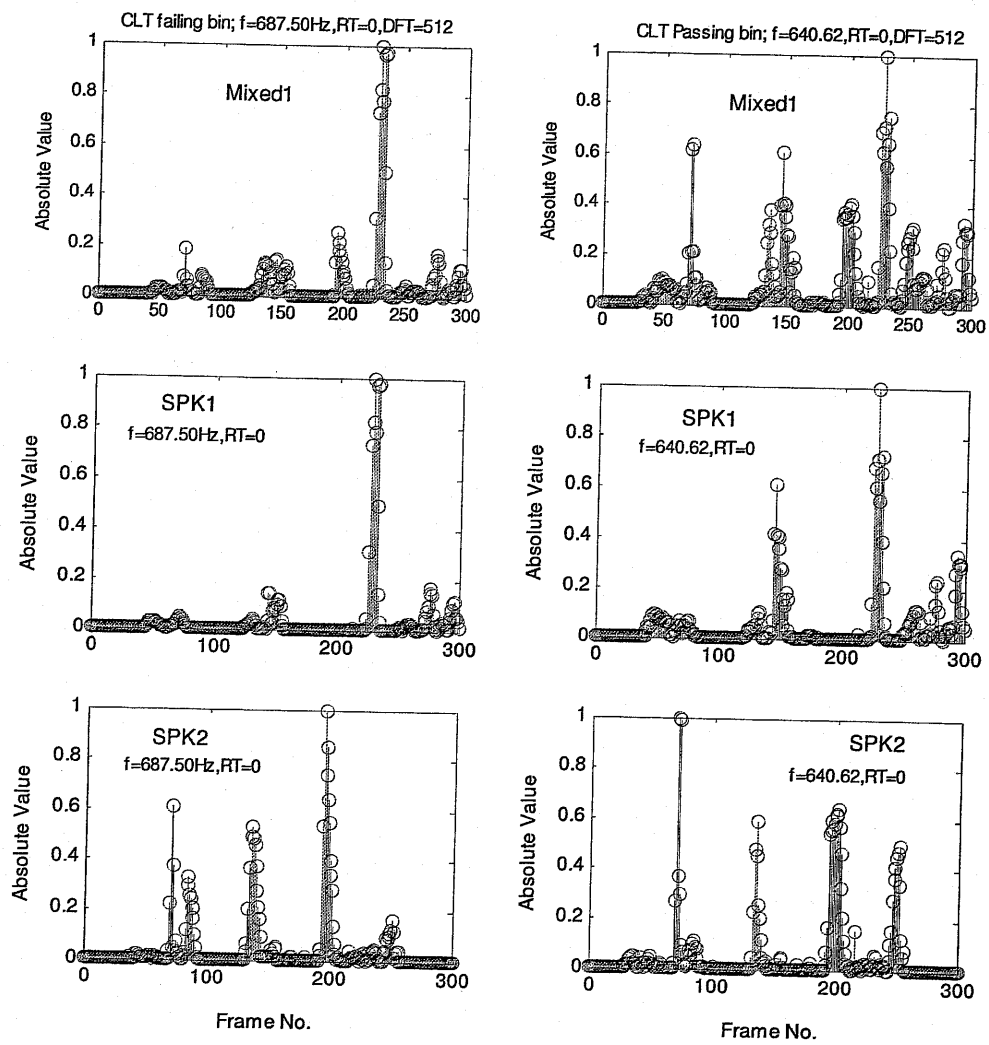


Figure 3.22 Role of spectral sparseness in CLT-disobedience. Plots in the left side column represent CLT failing bin at f=687.50 Hz and those of in right side represent CLT-complying bin at f=640.62 Hz. DFT size=512. SPK1 and SPK2 represent plots for spectral contributions from the first and second speakers, respectively.

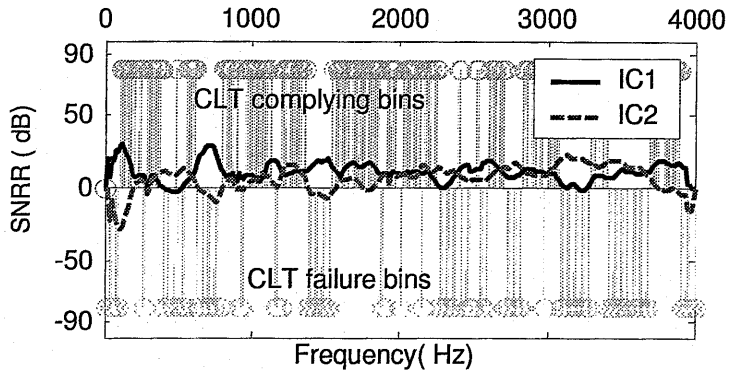


Figure 3.23 SNRR with CLT-disobeying frequency bins at Mic.1 for RT=300ms. (Speakers male and female).

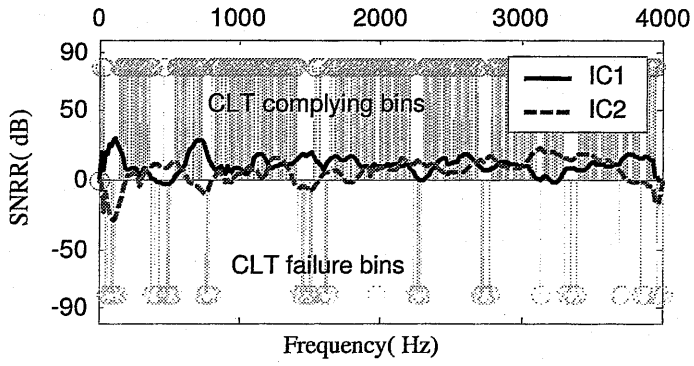


Figure 3.24 SNRR with CLT-disobeying frequency bins at Mic. 2 for RT=300ms. (Speakers male and female).

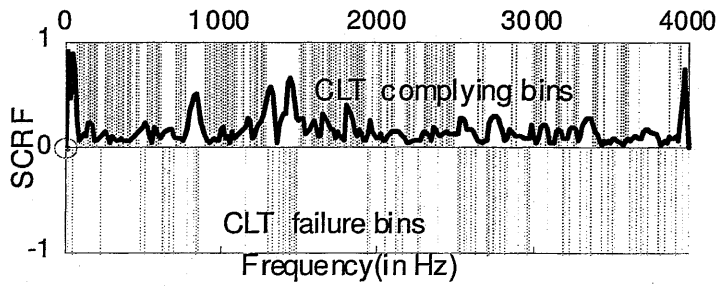


Figure 3.25 SCRF between separated ICs w.r.t CLT test for mixed signal at Mic.1 (RT=300ms).

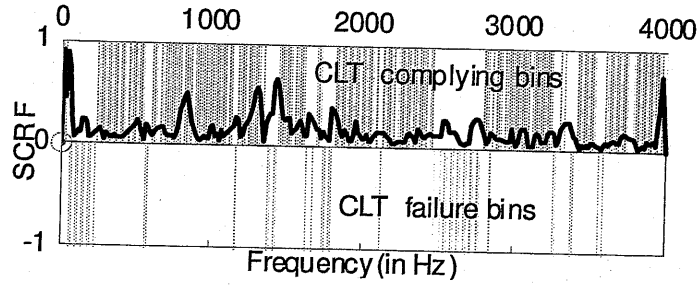


Figure 3.26 SCRF between separated ICs w.r.t CLT test for mixed signal at Mic.1 (RT=300ms, (Speakers both male, DFT size=512).

non-Gaussianization and to the best of our knowledge its use in audio signal separation in a realistic environment has not been reported yet. Almost similar results have been found for the other CLT-obeying and disobeying frequency bins. Obviously, CLT compliance is of vital importance for ICA algorithm working under the assumption of Gaussianization of the mixed data under the CLT principle. As the cause, sparseness of spectrum of speech signal, of CLT failure by speech is inherent so its happening cannot be stopped. The only way is to use the algorithms independent from such constraints, or combine some other methods such as NBF having no such problem in the CLT-failing frequency bins. However, this requires the blind detection of CLT obeying and disobeying frequency bins which has been presented in the next Chapter along with combination of NBF and FDICA algorithm.

Chapter 4

Probabilistic Modeling of TFSS and Its Application in BSS

4.1. Introduction

The statistical study of TFSS is important from many angles of thought for FDICA algorithms in general. The proposed fixed-point FDICA by negentropy maximization has its functional bearing on the compliance of CLT by the mixing process. This implies that there should always be gain in Gaussianity of the mixed signal over that of the unmixed signal. However, if the density function of the TFSS of individual speaker belongs to stable distribution the mixing process will not result in gain in Gaussianity because such PDFs are closed under any linear operations [19]. It has been pointed out in the last Chapter that the performance of the described FDICA algorithm is unexpectedly uneven in all frequency bins despite use of same non-linear functions. This is indicative of the fact that statistical characteristics of the TFSS in each frequency bin are different and are influential in separation process. It has also been pointed out that the choice of non-linear function G for the approximation of negentropy of the data depends on the PDF of the data. The performance of the fixed-point algorithm depends also on the used non-quadratic non-linear function G. It is desirable that the function G should provide robustness toward outlier values in the data as well as better approximation of true negentropy. Better robustness to outliers can be ensured by choosing G with slow variation with respect to change in data and at the same time very close approximation of negentropy can be expected if statistical characteristics of G inherit PDF of the data. The statistically efficient and optimal G that can accommodate maximum information about HOS of the data is chosen as the function that can minimize trace of the asymptotic variance of separation vector $\mathbf{w}$ and can be approximated by [19]

$$G = c_1 \log p(Y) \tag{4.1}$$

where c_1 is an arbitrary constant and $p(Y)$ represents PDF of Y. Thus investigation of statistical nature of TFSS is essential which has been presented in this chapter along with some experimental results.

4.2. Probability Density of TFSS

A statistical model of speech is not only needed in BSS but also in many statistical signal processing applications of practical importance, e.g., speech coding, speech recognition, speech enhancement, voice activity detection and so forth. The performance of such speech processing systems depends largely on the used statistical model of speech. Therefore, it is a matter of the utmost importance to identify and use the most exact or the best approximate statistical model of the speech in the domain of operation. A speech signal is a non-stationary random signal. Unfortunately, like its inherent natural non-stationarity, its statistical modeling by different researchers has also been inconsistent. The artificial speech recognition group has mostly modeled the speech signal by Gaussian Distribution (GD) or a mixture of GD or in rarely seen application by a generalized Laplacian [56]. In other applications it has been modeled by Laplacian Distribution (LD) or Gamma Distribution (γ-D). One of the natural reasons for adopting speech PDF as the Gaussian is the conceptual simplification in developing algorithms. The modeling of the speech probability distribution was started in 1950 by Davenport [57]. In that paper Davenport reported that the speech data in the time domain has a γ-D. In contrast to today's well-accepted LD model for speech, there has also been a research report that LD presents very poor and simpler approximation of the speech probability distribution [58]. In [59], the probability density of a very short segment of band limited speech has been modeled by multivariate GD with slowly time varying power. There have been many differences among the research reports due to the fact that the statistics of speech depends on its content (voice, silence, noise, etc.). Inside the speech one can find quasi-stationary (voiced and fricatives) parts as well as extremely non-stationary (explosion phase of stop consonants) part

that makes overall nature of speech signal non-stationary. The other equally important issue has been the duration of the used speech signal because for short time and long time speech data the PDFs differ. More recent reports on the time domain modeling of speech can be found in [60]. In that studies the authors tested different statistical hypotheses on speech PDF. The statistical modeling of long time and short time segments of speech has been extensively studied and the authors have emphasized that the speech signal PDF can be best approximated by the LD while negating the accuracy of approximation of the same by GD. Experimental conclusions in [60], for different time lengths of speech, suggest different PDFs such as speech signals have PDF similar to LD within time frame longer than 5 ms, the GD is favored for time frame shorter than 2.5 ms, while longer frames (>0.5 s) of speech shows γ-D or GGD with the shape parameter 0.44. These findings are important in understanding the PDF of speech segments of different time lengths in any linear transform domain such as DFT domain.

The statistical modeling of the spectral component of speech has also been a controversy since last five decades. One of the earlier efforts to model the spectral component of speech can be found in [61][62]. In these studies the authors have modeled gain normalized cosine transform coefficients, which are very much similar to the real part of the Fourier transform, of speech by the Gaussian PDF. In another research reported in [63], the authors found that the amplitude of the gain normalized Fourier Transform follows γ-D. However, in that study long time segment of the speech was used so it cannot be accurate to assign the same PDF to the DFT of small (20 ms - 40 ms) speech segments. Interestingly, in different applications researchers have assumed different PDF to the spectral components of speech, e.g., researchers have used the Gaussian PDF as speech spectral PDF in order to develop speech enhancement algorithms [64][65] and a voice-activity detection algorithm [66]. In such applications the basic reason behind the adoption of the GD for the DFT coefficients of speech have been the implication of the CLT because the DFT coefficients are weighted sum of the random data samples. It is important to note that in all of these efforts to model the DFT coefficients statistically, attention has been focused on the

DFT coefficients (as spectral components) of either the short or long segments of speech rather than on the distribution of the time-series of particular frequency components.

The TFSS $Z(f)$ in Eq.(3.5) by STFT analysis in any arbitrary frequency bin, except the DC and Nyquist frequency, represents a Complex Random Variable (CRV).These DFT coefficients are not mutually decorrelated. The normalized correlation coefficient between different DFT coefficients depends upon the frame length used. It approaches to zero as frame length tends to ∞ [64]. However, in the STFT method of frequency sub-banding, the frame-length of speech segment is deliberately confined between 10ms to 40 ms in order to introduce, artificially, the concept of stationarity or quasi-stationarity. Therefore, the completely decorrelated DFT coefficients cannot be produced for the time-series of speech spectral components. Thus a certain degree of correlation among DFT coefficients will exist even if we accept the assumption of complete decorrelation. Under the Cramer representation, the spectral components $X(f)$ (frame index λ is dropped for convenience) of the windowed signal $x_w(n)$ can be represented by

$$x_w(n) = \int_{-1/2}^{1/2} e^{2\pi jfn} dX(f). \tag{4.2}$$

It can be found in [54] that for the stationary $x_w(n)$, the spectral components $X(f)$ are circular. Thus the DFT coefficients of each quasi-stationary segment may also be supposed to form a Complex Circular Random Variable (CCRV). Each sample of the time-frequency series $\mathbf{Z}(f)$, of a particular frequency, is taken from the M set of such variables and is the ultimate representative, unique in the case of no overlapping, of time successive quasi-stationary speech segments. Since cross-segment stationarity in speech signal is not possible, at first glance the existence of concept of circularity for $\mathbf{Z}(f)$ seems illogical. However, due to overlapping and arbitrariness in the analysis window position, it seems natural to expect circularity in multidimensional complex random variable $\mathbf{Z}(f)$. A

multidimensional complex random variable $\mathbf{Z}(f)$ is said to be circular if its probability density is independent of complex rotation, i.e.

$$p(Z(f)) = p(Z(f)e^{j\phi}), \tag{4.3}$$

where $p(.)$ = probability density of (.) and ϕ = angle of complex rotation. The probability density function of the CRV $\mathbf{Z}(f) = a + ib$ depends on the PDFs of the real parts a and of imaginary parts b. Under the polar representation we have for $\mathbf{Z}(f)$

$$\mathbf{Z}(f) \dashv |\mathbf{Z}(f)| \angle \theta = \rho \angle \theta \tag{4.4}$$

where $\rho = \sqrt{a^2 + b^2}$ = Magnitude, $a = \rho \sin\theta$; $b = \rho \cos\theta, \theta = \arctan(-\frac{b}{a})$ = Phase. Thus, the PDF of the CRV can also be expressed in terms of the PDF of polar magnitude ρ and the phase θ. In the polar coordinate system the circularity condition in Eq.(4.3) of the multi-dimensional CRV can be expressed as[67]

$$p(\rho, \theta) = p(\rho, \theta - \varphi), \tag{4.5}$$

where φ=angle of complex rotation. This implies that for circularity of $\mathbf{Z}(f)$ its PDF should be independent of phase θ which in turn means that θ must have a uniform distribution. We discuss this issue further in the experiment section where it will be shown that the PDF of phase θ is uniformly distributed. Obviously, the PDF of the time-series of spectral components in each frequency bin can be determined by looking into the PDFs of their real part, imaginary part or polar amplitude and phase. There have been developments of several ICA algorithms for BSS that use PDF of the real part and imaginary part separately to optimize cost functions [18]. Similar algorithms based on the distribution of the magnitude of $\mathbf{Z}(f)$ in the polar co-ordinate can be found in [68]. However, all such algorithms approximate PDF by the same LD in all frequency bins. This is one of the serious causes of mismatch between the real PDF and the model used.

The PDF of $\mathbf{Z}(f)$ in each frequency bin depends upon the content of speech and may or may not be same in every frequency bin. However, for a few neighboring frequency bins there will be maximum similarity in the PDFs.

It was mentioned in the beginning of this chapter that the statistical modeling of the DFT coefficient of the windowed speech signal has been described as a normal distribution. A number of valid reasons, besides computational and conceptual simplicity, behind this can be categorized as follows. First, the DFT is asymptotically Gaussian in accordance with the following two theorems:

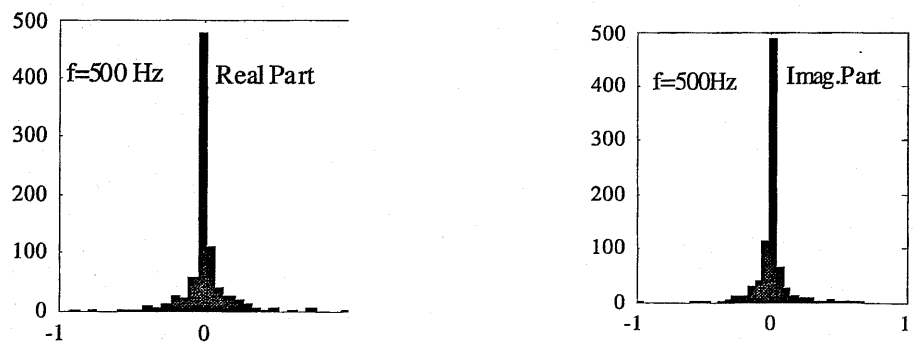
Theorem 1: *The joint distribution of any finite set of elements belonging to N-point DFT of a block of length L from a stationary sequence converges to normal as L approaches infinity if the elements of the sequence are strongly mixing (i.e. far separated frequency components are weakly dependent) and obey the (Lyapunov) condition that for any $\delta>0$, the $(\delta+2)$th moment is finite.*

Theorem 2: *The joint distribution of any finite set of elements belonging to the DFT of a block of length N from a sequence of independent, identically distributed random variables of finite variances converges to normal as N approaches infinity.*

The proofs of these theorems can be found in [42]. Secondly, the CLT supports Gaussianity of the DFT coefficient, as DFT coefficients are weighted sum of the random samples. This is evident from Eq.(3.4). Thirdly, a quasi-stationary segment of speech in the time domain is assumed to be Gaussian. DFT is a linear transformation upon such segment, so the PDF of DFT coefficients are also Gaussian. However, it is known fact that PDF of a small time segment of speech varies with its time length and is not necessarily Gaussian [60]. It is natural to consider that the PDF of the time-series of a spectral component of a quasi-stationary speech segment is related with the distribution of DFT of each segment and may be derived from it. Under such a framework the problem of PDF determination of $\mathbf{Z}(f)$ may be formulated as the

determination of the relation of PDF of $\mathbf{Z}(f)$ with the PDF of M sets of DFT coefficients, provided they are known, as obtained for each pseudo-stationary segment. However, such a derivation of relatedness among PDFs seems difficult and may be complicated. We adopt here a statistical hypothesis testing approach as well as unknown parameter estimation of the candidate theoretical distributions in order to find the suitable match for the PDF of $\mathbf{Z}(f)$.

In order to check the PDF of $\mathbf{Z}(f)$ we choose GD, LD and GGD with estimated parameters as the candidate theoretical PDF for the null-hypotheses for the PDF of $\mathbf{Z}(f)$. The choice of these theoretical density functions is not arbitrary but is made on the basis of following two reasons. First, different researchers have used such PDFs to approximate the PDF of speech spectral components. The second reason, which is more convincing, is the shapes of the histograms of the $\mathbf{Z}(f)$ in different frequency bins are spiky as shown in Figure 4.1 for f=500 Hz. These histograms are very spiky, like Laplacian distribution, with variation in the peakedness in the different frequency bins. To follow this variation in peakedness we have used the GGD. We present here very brief mathematical descriptions, which will be helpful in further discussion. Gaussian PDF $f_G(z;\mu,\sigma)$ of random variable z and Cumulative distribution function (CDF)



(a) (b)

Figure 4.1 Histograms of (a) real parts, and (b) imaginary parts of time-series of DFT of male speech at Mic1.

$F_G(z)$ are given by

$$f_G(z;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}, \tag{4.6}$$

where μ=mean and σ=variance and CDF is given by error function (*erfc*) as follows

$$F_G(z)=\frac{1}{2}erfc\left(\frac{z-\mu}{\sqrt{2\sigma^2}}\right). \tag{4.7}$$

The PDF $f_L(z;\mu,\alpha)$ and CDF $F_L(z)$ of the LD are given by

$$f_L(z;\mu,\alpha)=\frac{\alpha}{2}e^{-\alpha|z-\mu|}, \alpha>0 \ \& \ -\infty<z<\infty, \tag{4.8}$$

and

$$F_L(z)=\begin{cases}\frac{1}{2}e^{\alpha z} \ for \quad z\le 0\\ 1-\frac{1}{2}e^{-\alpha z} \ for \quad z>0.\end{cases} \tag{4.9}$$

The generalized Gaussian distribution is a flexible parametric distribution family that incorporates various distribution shapes such as uniform, normal, Laplacian, and even more highly peaked distributions with exponentially decaying heavy tails. Accordingly it can be used to model data with distributions symmetric at mean with varying degree of peakedness. This distribution was introduced for the first time in [69] in the development of the Bayesian inferential process. However, that incorporates peakedness up to Laplacian only. More detailed descriptions for the wide range of distribution shapes can be found in [70].

The GGD family with the location parameter μ, scale parameter α and shape

parameter β is given by

$$f_{GG}(z;\alpha,\beta)=\frac{\beta}{2\alpha\Gamma(1/\beta)}\exp\left(-[|z-\mu|/\alpha]^{\beta}\right)=\frac{b\beta}{2\Gamma(1/\beta)}\exp(-[b|z-\mu|]^{\beta})$$
$$=A\exp(-[b|(z-\mu)|]^{\beta}) \tag{4.10}$$

and

$$b=\frac{1}{\alpha}=\frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}}, \tag{4.11}$$

where $\Gamma(x)=\int_0^{\infty}e^{-t}t^{x-1}dt$= Gamma distribution, and $A=\frac{b\beta}{2\Gamma(1/\beta)}$, $-\infty\le z\le\infty$, $\alpha>0$, and $\beta>0$.

The CDF of GGD function is given by $F_{GG}(z;\alpha,\beta,\mu)=\int_{-\infty}^{z}Ae^{-\left(\frac{|z-\mu|}{\alpha}\right)^{\beta}}dz$.

This integral can be solved numerically or can be solved in a few steps in terms of incomplete gamma function γ_{inc} as follows

$$F_{GGD}(Z)=\begin{cases}0.5-0.5\gamma_{inc}\left(\frac{(|z-\mu|)^{\beta}}{\alpha^{\beta}},\frac{1}{\beta}\right), & z-\mu<0\\ 0.5+0.5\gamma_{inc}\left(\frac{(|z-\mu|)^{\beta}}{\alpha^{\beta}},\frac{1}{\beta}\right), & z-\mu>0\\ 0.5, & z-\mu>0\end{cases}, \tag{4.12}$$

where $\gamma_{inc}(x,\delta)=\frac{1}{\Gamma(\delta)}\int_0^{x}t^{\delta-1}e^{-t}dt;\delta>0$. The shape parameter β determines the shape of the distribution. When $-1<\beta<0$, the distributions are short tailed and well-peaked compared to normal; when $\beta>0$, it shows the opposite characteristic. Distribution graphs are symmetrical about the mode at $x=\mu$, and exponential power curves approaches x-axis asymptotically at both extremes [71]. For $\beta=1$ the distribution is Laplacian, for $\beta=2$ the distribution is Gaussian and distribution tends to become uniform as $\beta\to\infty$. The distribution shapes for different values of β are

shown in Figure 4.2.

4.3. GGD Parameter Estimation

In order to fit the GGD distribution in the time-series of the speech spectral component, GGD parameters μ, α and β in each frequency bin are estimated from the speech data. For this purpose, Maximum Likelihood (ML) estimation will be used. However, the exact determination of GGD parameters by solving the likelihood equation is cumbersome as these parameters are interdependent. The location parameter μ can be estimated from the mean or median of the data in each frequency bin. The scaling parameter α depends on the variance of the data and shape parameter β. In the time domain it is assumed that long speech data has zero mean and unit variance, however, the same is not true for the quasi-stationary segments of speech that are too small. When a quasi-stationary segment $x_w(n)$ undergoes N-point DFT, its variance is rearranged over the spectral components [72][73] such that

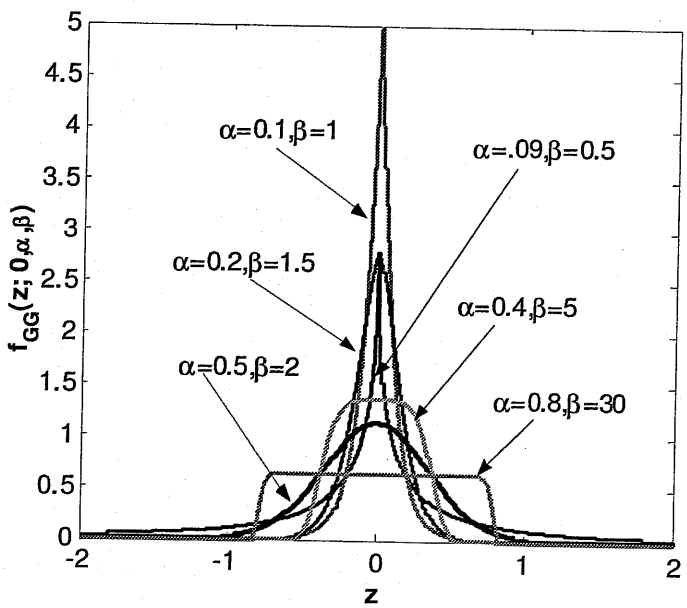


Figure 4.2 GGD distribution for different values of α and β.

$$\mu_x = Z(f_0)/N, \tag{4.13}$$

$$\sigma_x^2 = \frac{1}{N^2}\sum_{k=1}^{N-1} |X(f_k)|^2 . \tag{4.14}$$

Obviously, each complex sample of the time-series of a spectral component has different variance history and it is difficult to relate with the above equation. In order to test the suitability of mean $\bar{z}$ and median $\tilde{z}$ as an estimator for the location parameter for the different values of β we define efficiency of the estimator as the ratio of their variances i.e.

$$\eta(\beta) = \frac{\text{var}[\bar{z}]}{\text{var}[\tilde{z}]}. \tag{4.15}$$

All the odd order moments of GGD vanish and distribution is characterized by even order moments. The rth even order moment of GGD is given by

$$E[z^r] = \int_{-\infty}^{\infty} z^r A e^{-|bz|^\beta} dz = 2A\int_0^{\infty} z^r e^{-(bz)^\beta} dz. \tag{4.16}$$

Now, with setting $(bz)^\beta = y$, the above integrand can be solved to

$$E[z^r] = \frac{2A}{b^{r+1}\beta}\int_0^{\infty} y^{\frac{(r+1)}{\beta}-1} e^{-y} dy = \frac{1}{b^r}\frac{\Gamma\left(\frac{r+1}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}. \tag{4.17}$$

Since E[z]=0; the var($\bar{z}$)=var(Z)=E[z^2]. Putting r=2 in Eq.(4.17) we get,

$$E[z^2] = \frac{1}{b^2}\frac{\Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}. \tag{4.18}$$

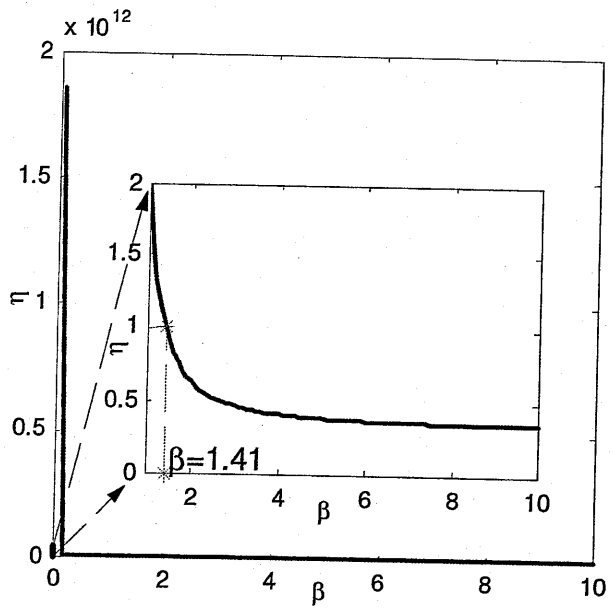


Figure 4.3 Theoretical value of η for estimators of μ.

The variance of median $\tilde{z}$ is related to $f_{GG}(z;\alpha,\beta)$ as follows [74]

$$\mathrm{var}(\tilde{z}) = \frac{1}{4f_{GG}^2(0;\alpha,\beta)} = \frac{1}{4A^2} = \frac{b^2\beta^2}{4\Gamma^2(1/\beta)}. \tag{4.19}$$

From Eq.(4.15), (4.18) and (4.19), the value of η is given as

$$\eta(\beta) \approx \frac{\Gamma(3/\beta)}{\Gamma(1/\beta)\Gamma^2(1+1/\beta)}. \tag{4.20}$$

This function is plotted in Figure 4.3 for different values of β. The inner figure shows a scaled-up y-axis for the higher values of β. It is evident from this graph that the mean is good estimator of location parameter for β>1.41 (shown by '*' in the inner figure). For the lower value of β ($\beta<1.41$), the median is better estimate of the location parameter. Since neither var[$\bar{z}$] nor var[$\tilde{z}$] is a

monotonic function of β, because they depends on gamma function of β, the suitability of both estimators worsen as $\beta \to 0$. Under this condition, $\eta \to \infty$. Therefore, for extremely low value of β neither can provide good estimation of the location parameter. However, as $\beta \to \infty$, it can be shown that $\eta \to 1/3$ and mean provides a good estimation of the location parameter. Thus the estimation of the location parameter depends upon the β. However, in our experiment we will use the mean as the estimator of the location parameter as the value of β for almost all the frequency bins is not extremely low.

Now Maximum Likelihood (ML) approach is described for which has also been used by other researchers to measure the scaling and shape parameters of GGD [75][76]. The ML estimator can provide a very accurate estimation of the GGD parameters provided that β is not too small [77]. The ML function for centered samples $z = [z, z_2, \ldots\ldots z_M]$ in the frequency bin f can be given as

$$L(z;\alpha,\beta) = \log \prod_{i=1}^{M} (z_i;\alpha,\beta). \tag{4.21}$$

In accordance with [75], the likelihood equations having a unique root in probability give maximum likelihood estimator and can be obtained by partial differentiation of the above function with respect to unknown parameters α and β as follows

$$\frac{\partial L(z;\alpha,\beta)}{\partial \alpha} = -\frac{L}{\alpha} + \sum_{i=1}^{L} \frac{\beta |z_i|^{\beta} \alpha^{\beta}}{\alpha} = 0 \tag{4.22}$$

and the partial derivative w.r.t. β is given by

$$\frac{\partial L(x;\alpha,\beta)}{\partial \beta} = -\frac{L}{\beta} + \frac{L\psi(1/\beta)}{\beta^2} - \sum_{i=1}^{L} \left(\frac{|x_i|}{\alpha}\right)^{\beta} \log\left(\frac{|x_i|}{\alpha}\right) = 0, \tag{4.23}$$

Where $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$ represents diagamma function. For $\beta > 0$ from Eq. (4.22) we obtain an estimate of parameter α as follows

$$\alpha=\left[\frac{\beta}{L}\sum_{i=1}^{M}|z_i|^{\beta}\right]^{1/\beta} \tag{4.24}$$

Using this estimate in Eq.(4.23), we get the following transcendental equation for β (Since $\frac{L}{\beta}\neq 0$)

$$\frac{L}{\beta}+\frac{M\psi(1/\beta)}{\beta^2}-\sum_{i=1}^{M}\left(\frac{|z_i|^{\beta}}{\frac{\beta}{L}\sum_{i=1}^{M}|z_i|^{\beta}}\right)\log\left[\frac{|z_i|}{\left(\frac{\beta}{L}\sum_{i=1}^{M}|z_i|^{\beta}\right)^{1/\beta}}\right]=0 \tag{4.25}$$

$$\Rightarrow\frac{L}{\beta}\left[1+\frac{\psi(1/\beta)}{\beta}-\frac{\sum_{i=1}^{M}|z_i|^{\beta}}{\sum_{i=1}^{M}|z_i|^{\beta}}\left\{\log|z_i|-\frac{1}{\beta}\log\left(\frac{\beta}{L}\sum_{i=1}^{M}|z_i|^{\beta}\right)\right\}\right]=0$$

$$\Rightarrow\frac{L}{\beta}\left[1+\frac{\psi(1/\beta)}{\beta}-\frac{\sum_{i=1}^{M}|z_i|^{\beta}\log|z_i|}{\sum|z_i|^{\beta}}+\frac{\log\left(\beta/L\sum_{i=1}^{M}|z_i|^{\beta}\right)}{\beta}\right]=0.$$

Since $\frac{L}{\beta}\neq 0$

$$1+\frac{\psi(1/\beta)}{\beta}-\frac{\sum_{i=1}^{M}|z_i|^{\beta}\log|z_i|}{\sum|z_i|^{\beta}}+\frac{\log\left(\beta/L\sum_{i=1}^{M}|z_i|^{\beta}\right)}{\beta}=0. \tag{4.26}$$

Eq.(4.26) can be further expressed as

$$g(\beta)=0, \tag{4.27}$$

where $g(\beta)$ represents RHS of the Eq.(4.26). The roots of Eq.(4.26) give the value of the shape parameter. This equation can be solved numerically. The solution using the Newton-Raphson iterative method is given by

$$\beta_{k+1} = \beta_k - \frac{g(\beta)}{g'(\beta)}, \tag{4.28}$$

in which $g'(\beta)$ represents first order derivative of $g(\beta)$ w.r.t β and is given by

$$g'(\beta) = -\frac{\psi(1/\beta)}{\beta^2} - \frac{\psi'(1/\beta)}{\beta^3} + \frac{1}{\beta^2} - \frac{\sum_{i=1}^{M} |z_i|^\beta (\log |x_i|)^2}{\sum_{i=1}^{M} |z_i|^\beta} + \frac{\left(\sum_{i=1}^{M} |z_i|^\beta \log |z_i|\right)^2}{\left(\sum |z_i|^\beta\right)^2}$$

$$+ \frac{\sum_{i=1}^{M} |z_i|^\beta (\log |z_i|)}{\beta \sum_{i=1}^{M} |z_i|^\beta} - \frac{\log\left(\frac{\beta}{L}\sum_{i=1}^{M} |z_i|^\beta\right)}{\sum_{i=1}^{M} |z_i|^\beta}. \tag{4.29}$$

The solution obtained by Eq.(4.28) is sensitive to initial value of β. The good initial value can be obtained from the Generalized Gaussian Ratio (GGR), denoted by $\gamma(\beta)$ and is defined as the ratio of mean of the absolute value to the standard deviation of the data[60]as follows

$$\gamma = \frac{E[|z|]}{\sigma_z^2}. \tag{4.30}$$

In the context of GGD it can be shown that the mean of absolute value is given as

$$E[|z|] = \int_{-\infty}^{\infty} |z| f_{GG}(z) dz = \int_{-\infty}^{\infty} |z| A e^{-[b|z|]^\beta} dz. \tag{4.31}$$

Since this integration is on the absolute value of the variable over the given limit. It can be given by

$$E[|z|] = 2\int_{0}^{\infty} A z e^{-[bz]^\beta} dz. \tag{4.32}$$

Now substituting $(bz)^\beta = y$, above integral can be solved to

$$E[|z|]=\frac{2A}{b^2\beta}\int_0^{\infty} y^{(\frac{2}{\beta}-1)}e^{-y}dy=\frac{2A}{b^2\beta}\Gamma(2/\beta),\quad for\ \ 0<\beta. \tag{4.33}$$

Using the values of A and b from Eq. (4.10) and (4.11) , respectively, Eq.(4.33) can be simplified to

$$E^2[|z|]=\frac{\sigma_z^2\Gamma(1/\beta)}{\Gamma(3/\beta)}\frac{\Gamma^2(2/\beta)}{\Gamma^2(1/\beta)}, \tag{4.34}$$

$$\frac{E[|z|]}{\sqrt{\sigma_z^2}}=\gamma(\beta)=\frac{\Gamma(2/\beta)}{\sqrt{\Gamma(3/\beta)\ \Gamma(1/\beta)}}, \tag{4.35}$$

$$E[|z|]=(1/M)\sum_{i=1}^{M}|z_i| \text{ and } \sigma_z^2=(1/M)\sum_{i=1}^{M}z_i^2. \tag{4.36}$$

Now the denominator and numerator of the LHS of Eq.(4.35) can be computed from the given data and good initial value $\beta_{initial}$ of the shape parameter, given by

$$\beta_{initial}=\gamma^{-1}\left(\frac{E[|z|]}{\sqrt{\sigma_z^2}}\right). \tag{4.37}$$

Using this as initial value of β in the iterative Eq.(4.28), the final value is obtained in a few iterations.

4.4. Other Statistical Tests

4.4.1. Moment Test

The moment test is based on the fact that the value of the GRR function γ is distinctive and unique for each theoretical distribution defined by β. This test statistics for different theoretical distributions, obtainable for different values of

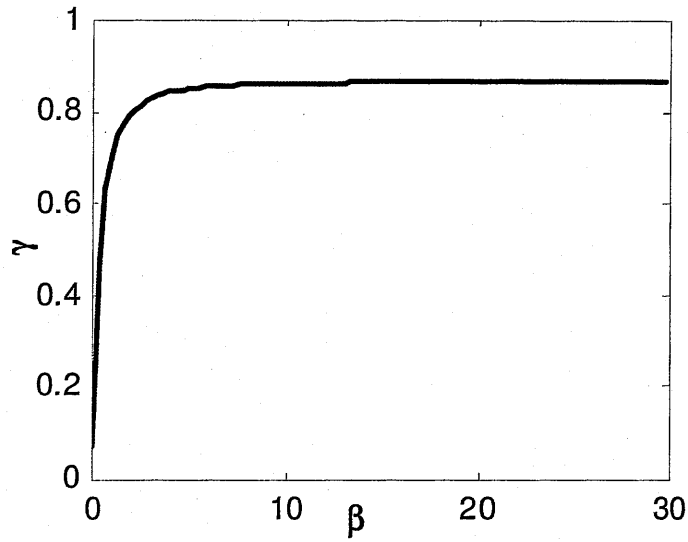


Figure 4.4 Plot of values of Genenaralized Gaussian Ratio function γ versus β.

$$\gamma(\beta)=\begin{cases} 0.7 & \textit{Laplacian}; \quad \beta=1 \\ 0.8 & \textit{Gaussian}; \quad \beta=2 \\ >0.8 & \textit{Uniform}; \quad \beta>2. \end{cases} \tag{4.38}$$

$\beta>0$ from the GGD, are given below. These test statistics can be used to infer nearness and relatedness of the given data distribution by comparing the standard values in Eq.(4.38) with the corresponding values calculated from the data using Eq.(4.35) and (4.36). The theoretical variation in the value of γ as a function of β is shown in Figure 4.4.The value of γ becomes almost constant as $\beta\rightarrow\infty$.

4.4.2. Quantile-Quantile (QQ) Plot:

The acronym QQ-plot stands for the quantile-quantile plot which is used to check similarity of the unknown data distribution with standard PDF or to check whether the two data sets can be approximated by the same PDF [78]. Supposing the probability $p\in(0,1)$, the pth order quantile $Q(\mathrm{p})$ of a distribution of a random variable $Z(f)=[z_1,z_2,z_3,\ldots z_M]$ with distribution function $F(\mathrm{z})$ refers to

that value z_p of Z for which

$$Q(p) = z_p = \Pr(Z \leq z_p) \leq p, \tag{4.39}$$

and

$$F(z_p) = \Pr(Z \leq z_p) = p, \tag{4.40}$$

Where Pr(.) represents probability of (.). It is evident that the quantiles are the value of Z where its CDF crosses probabilities p. Equations (4.39) and (4.40) can be combined to give

$$F(Q(p)) = p \Rightarrow Q(p) = F^{-1}(p). \tag{4.41}$$

Thus the Quantile (Percentile) Point Function (QPF or PPF) is computed as inverse of CDF which implies that

$$Q(p) = \{z_p \in R; p \leq F(z_p)\}. \tag{4.42}$$

In the QQ-plot QPF of the theoretical PDF is plotted against the sorted value (order statistics) of the observed data. The observed data in each frequency bin is sorted in ascending order such that $\{z_{(1)} \leq z_{(2)} \leq z_{(3)} \ldots\ldots \leq z_{(M)}\}$. The ith order statistics $z_{(i)}$ is the $(i-0.5)/M$ quantile, i.e.,

$$z_{(i)} = Q\left(\frac{i-0.5}{M}\right). \tag{4.43}$$

As an indication of the good fit of the theoretical distribution to the given data the plotted values fall onto a straight line. Also QQ-plot between the data of the two frequency bins can be plotted to see the similarity in their distribution. Supposing probability $p \in (0,1)$, the QPFs $Q_G(p)$, $Q_L(p)$, and $Q_{GG}(p)$ of the GD, LD and GGD, respectively, are given as the inverse of their CDF functions as follows:

$$Q_G(p) = \sqrt{2}\sigma \, erfinv(2p)+\mu. \tag{4.44}$$

$$Q_L(p) = \begin{cases} \log 2p, & p < 0.5 \\ 0, & p = 0.5 \\ -\log 2(1-p), & p > 0.5. \end{cases} \tag{4.45}$$

and

$$Q_{GG}(p) = \begin{cases} -G^{\frac{1}{\beta}} + \mu, & 0 \le p < 0.5 \\ G^{\frac{1}{\beta}} + \mu, & 1 \ge p > 0.5 \end{cases} \tag{4.46}$$

where $G=\Gamma^{-1}(p_1, 1/\beta, \beta)$ is the gamma function of p_1 with parameters $1/\beta$ and β and

$$p_1 = \begin{cases} 1-2p, & 0 \le p < 0.5 \\ 1+2p, & 1 \ge p > 0.5. \end{cases} \tag{4.47}$$

The QPF for the Laplacian and Gaussian distribution can also be computed from Eq.(4.46) by putting β=1 and 2 respectively.

4.4.3. Chi-Square Goodness of Fit Test

The χ2-test [79] does not require any parameter estimation. It is used to compare data distribution with the theoretical PDF. For this test data is divided into B data bins and difference between observed and expected frequency (no. of occurances) in each data-bin is obtained to calculate Chi-square score given by

$$\chi^2 = \sum_{i=1}^{B} \frac{(O_i - E_i)^2}{E_i}, \tag{4.48}$$

where O_i =observed frequency in ith bin and E_i =expected frequency in the ith bin.The observed frequency is computed by direct counting of the number of samples falling in the each bin while the expected frequency is computed as follows

$$E_i = M[F(B_U) - F(B_L)], \tag{4.49}$$

where M=total no. of samples; F= CDF of hypothesis distribution and $B_U(B_L)$ = upper (lower) limits for bin *i*.

The $\chi 2$-test is sensitive to bin width. The bin width is selected such that the expected frequency becomes more than 5 in each data-bin. The most widely accepted bin width is calculated as

$$B_U - B_L = 0.3\sigma, \tag{4.50}$$

where σ=Std. deviation. The lower value of $\chi 2$-score provides better similarity between hypothesized PDF and the PDF of the given data. The $\chi 2$ –test has been used to check the goodness of fit of null-hypotheses for the real part, imaginary part and polar amplitude of the time series of speech spectral components $\mathbf{Z}(f)$ in each frequency bin.

Null-hypotheses

- Z(f) follows the Gaussian distribution
- Z(f) follows the Laplacian distribution
- Z(f) follows the Generalized Gaussian distribution with parameters estimated from Z(f).

4.5. Experiments and Results

Experimental Setup

In the experiment, we used a two-element linear microphone array with inter-element spacing of 4 cm for the simulated speech data generation. The direction of arrival (DOA) of two speech signal sources (male and female) were fixed at -30^0 and 40^0, assuming center of the microphone as reference and broadside on positions as the 0° DOA. The distances of the speakers were set at of 1.15 m from the center of the array. The whole experimental setup can be seen in Figure 3.6. Two types of sentences spoken by a male and a female speaker, of

time length 32.5 sec (produced by concatenation), from the ASJ continuous speech corpus for the research [80], were selected to serve as dry sources S_1 and S_2 for the generation of the mixed signals. Mixed signals at each microphone are obtained adding together the speech signals arriving from each source. The contribution of each source at each microphone is obtained by convolving the seed speech samples with the room impulse response between the involved source and microphone, recorded in a real room with reverberation time RT=300 ms. In this way the set of reference signals $[ref_{11}, ref_{12}, ref_{21}, ref_{22}]$, as indicated in Figure 3.1, as the contribution of each source at each microphone were obtained. The marginal distribution of these reference signals can approximate the distribution of the original sources S_1 and S_2. In order to do further study these reference signals are subjected to STFT to generate TFSS. The STFT signal analysis conditions were kept same as in Table 3.1. The total number of samples obtained in each frequency bin is sufficient to give good statistics of the data.

Results of Estimation of GGD Parameters:

The GGD parameters, namely, location parameter μ, scaling parameter α and the shape parameter β were calculated using the ML method, as discussed previously, in each frequency bin for each reference signal. The mean of the data has been used in each frequency bin to estimate the location parameter. As the computation of scaling parameter in Eq.(4.24) needs shape parameter, first shape parameter was computed using Eq.(4.28). The initial value of the shape parameter computation in Eq.(4.37) requires inversion of the GGR function for different values of β. It is computed from pre-computed look-up table. In the case if exact value is not available in the table the nearest value is extrapolated or interpolated. It is the shape parameter that determines the shape of the PDF. The shape parameters calculated for the male and female reference signals, ref_{21} *and* ref_{22}, respectively, at the second microphone are shown in Figure 4.5 and Figure 4.6. For each speaker, β is shown for the real part, imaginary part, polar magnitude and phase of the signal. The value of β for very low frequency

(<200 Hz) bins for both speakers have outlier values. However, these components are non-speech signals. The statistics of the shape parameters in different frequency bins can be better conceived from the histograms shown in Figure 4.7 and Figure 4.8 for the two speakers at the second microphone. The shape parameters for both the real and imaginary part are less than 0.5 in almost frequency bins, which corresponds to a strong Laplacian distribution as defined by GGD. In a very small number of frequency bins β is very near to 1. Also, the shape parameter is different for each frequency bin, which shows that the distribution of each frequency bin is different. However, the shape parameters for the neighboring frequency bins are almost same. Thus the assumption of LD for the imaginary part and the real part of the $\mathbf{Z}(f)$ in any frequency bin looks loose and inappropriate. However, different ICA algorithms have been developed with such assumption. The shape parameter for the polar magnitude in almost all frequency bins, except very low frequency bins, is very near to unity, which corresponds to the Laplacian distribution. However, it is not obvious why the probability distribution of the polar magnitude is nearer to LD than to that of the real or imaginary parts of the same signal, however, on the on the basis of CLT it can be urged that polar magnitudes are summation of squared real and imaginary parts, which are more spiky, so the polar magnitude should be less spiky. This fact may be one of the causes of the better performance of polar co-ordinate based non-linear function, as proposed in [68], over the Cartesian co-ordinate based non-linear function for the FDICA.

The shape parameter for the phase data is greater than 15 for the both speaker in almost all frequency bins. This value of β corresponds to a uniform distribution as defined by GGD. This result agrees with the intuitive fact that the phase of the samples in the DFT coefficients depends on the analysis window position, which is arbitrary. Therefore, the phase of the DFT coefficients of each quasi-stationary segment has uniform distribution. The TFSS in each frequency bin contains samples chosen from the DFT coefficients of speech segments, so its phase distribution is also uniform. The uniformity in the phase distribution ensures its neutrality for an arbitrary complex rotation which in turn means, in

accordance with Eq.(4.3), that $Z(f)$ in each frequency bin is the CCRV. It is also important to note that no major difference between the values of β for the $Z(f)$ for male and female speech was found. Almost similar results were found for the reference signals ref_{11} *and* ref_{12} at the first microphones also. The fitting of the GGD PDF with the estimated parameters in the histogram of $Z(f)$ at f=300Hz is shown in Figure 4.9 for the male speaker. These figures also show fitting of the Laplacian PDF. For the polar representation GGD and LD have almost same fittings however, for the real part, imaginary part or phase GGD fitting is much better than that of LD or GD.

Results of Moment Test:

The moment test for the reference speech signals of each speaker at each microphone was done. The choice of reference signal gives clean signal captured by both microphones from different speakers. The result of moment test for speech from a male speaker received at the second microphone is shown in Figure 4.10 . The moment ratio or the GGR function γ is obtained in each frequency bin using Eq.(4.35). The GGR functions for the four theoretical distributions namely GD, LD, GGD and uniform are drawn as dashed vertical lines. The results of moment test also favor the GGD for the real part and imaginary part. For the phase, again, the uniform distribution is favored. It is interesting to note that the moment test results in most of the frequency bins for polar magnitude do not favour LD. ML estimation of β for the polar magnitude supports nearness with LD while the statistics of γ favors a strong super-Gaussianity as defined by the GGD. The cause of this significant difference between the results is the difference between the values of $\beta_{initial}$, estimated by Eq.(4.37) and that of β_{ML} estimated by polishing $\beta_{initial}$ using Eq.(4.28) under the ML approach. These values of β are also shown in Figure 4.5and Figure 4.6. These two values of shape parameter do not differ by handsome amount for real and imaginary parts. These values are significantly different for the polar magnitude. β_{ML} has been found to be higher in every frequency bin for the polar magnitude for both the male and female speakers.

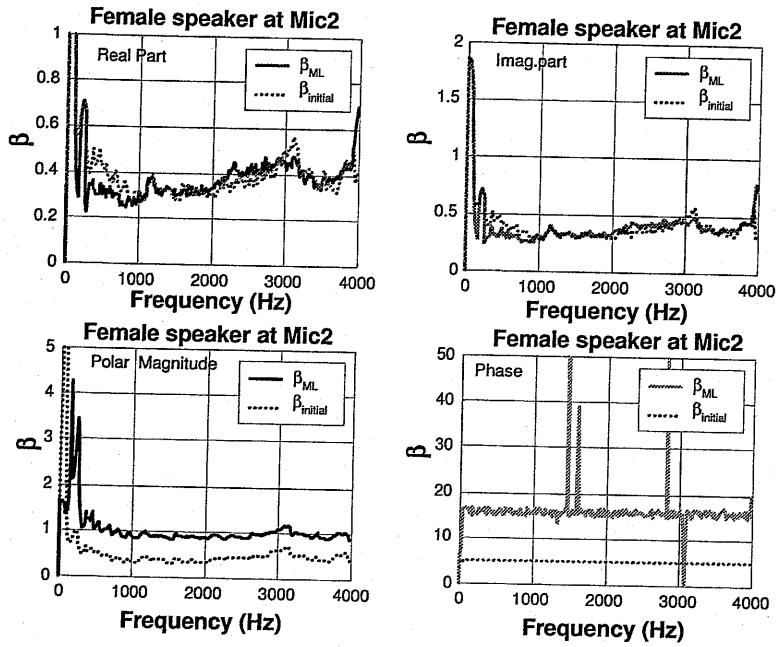


Figure 4.5 Shape parameter β for the time-series of spectral components of the female speech at the second microphone.

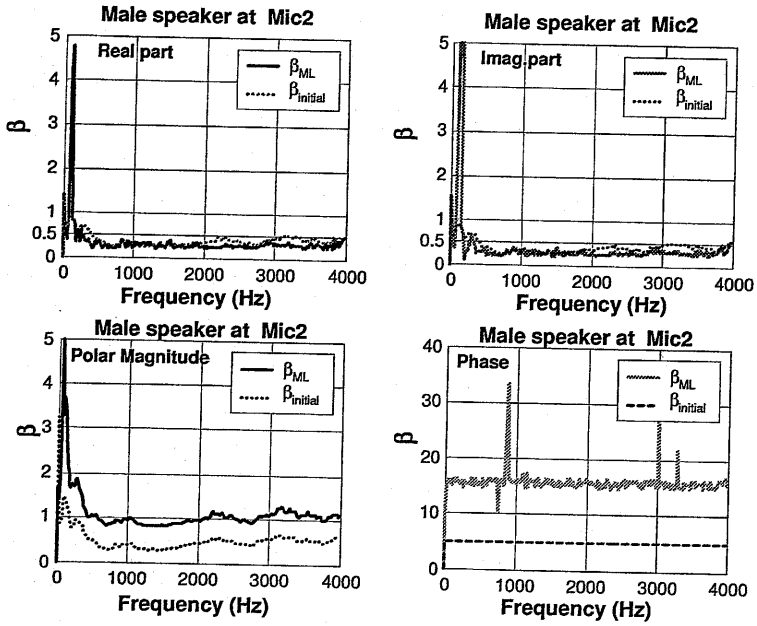


Figure 4.6 Shape parameter β for the time-series of spectral components of the male speech at the second microphone.

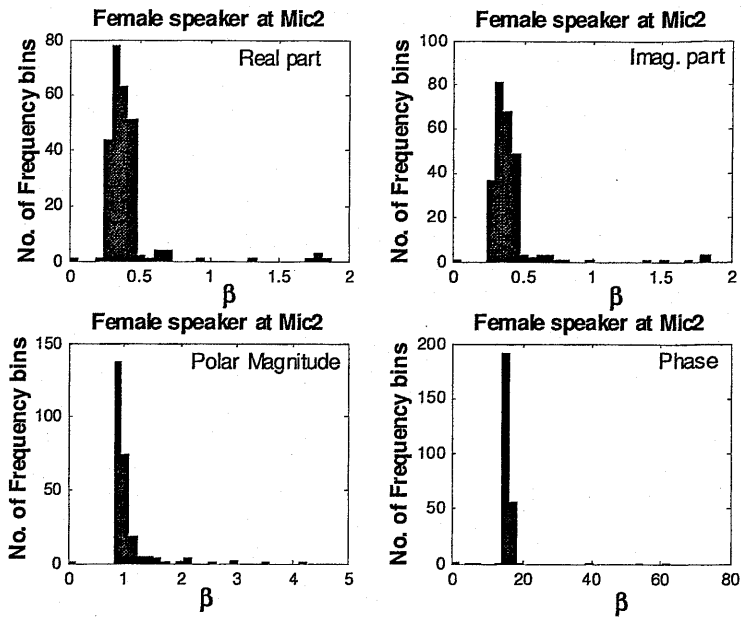


Figure 4.7 Histogram of shape parameter of different frequency bins for the female speaker at the second microphone.

Ss

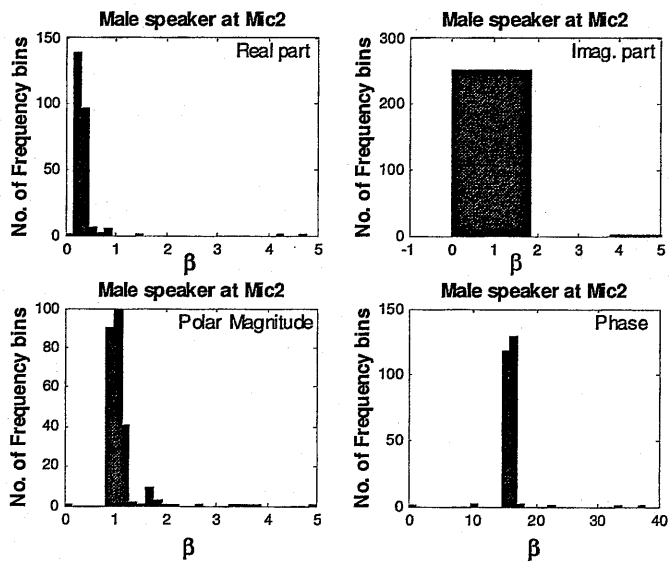


Figure 4.8 Histogram of shape parameter of different frequency bin for the male speaker at the second microphone.

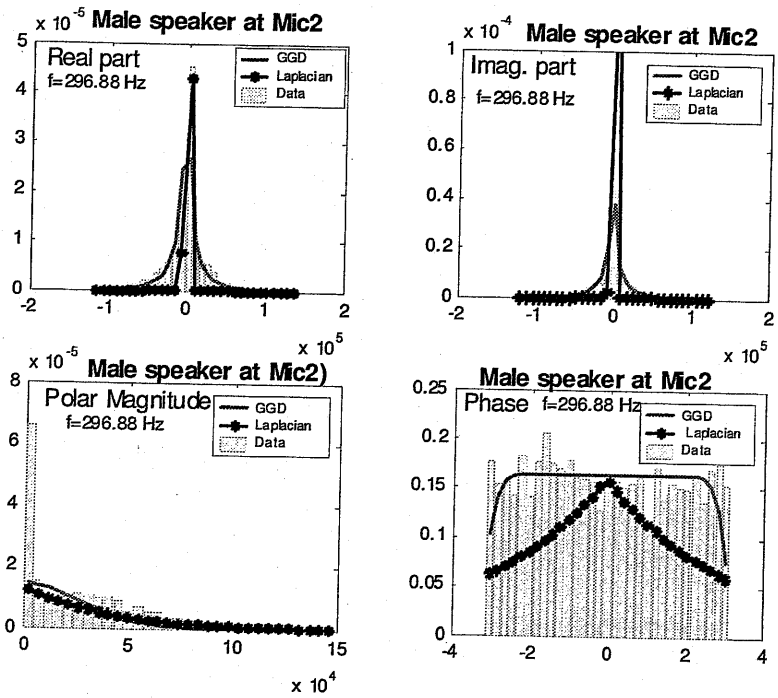
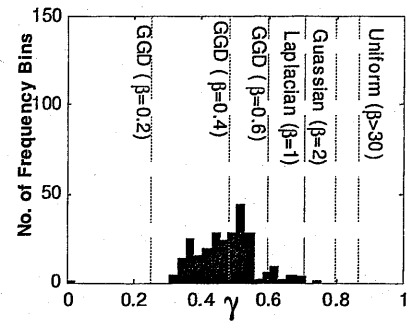


Figure 4.9 Comparison of GGD (with estimated parameters) and Laplacian (variance decided by the estimated α) PDF fitting in the histogram of Z(f), f=296.88 Hz, of the male speech at the second microphone.

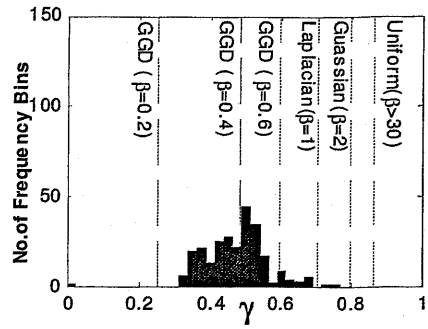
For phase $\beta_{initial}$ may be greater than the shown value because it is the highest value of β included in the look-up table and have not been extrapolated. Therefore, relying on the ML estimates of the shape parameter for polar magnitudes it can be concluded that polar magnitude is nearer to LD than the real or imaginary part. It is natural as the polar magnitude is summation of two less Gaussian variables(real and imaginary parts) which makes polar magnitude to move towards Gaussian ($\beta \rightarrow 2$) under the implication of CLT. Results for other reference signals were also found to give similar evidences.

Results of QQ-plots:

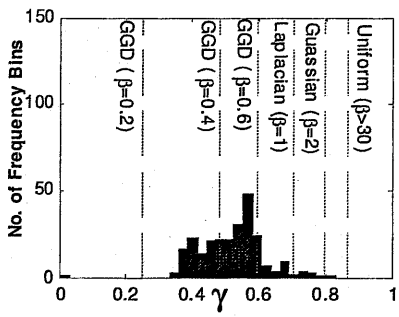
The QQ-plots for the **Z**(*f*) of the speech signals of the male speaker at the second microphone are shown in the Figure 4.11. This figure contains QQ-plots



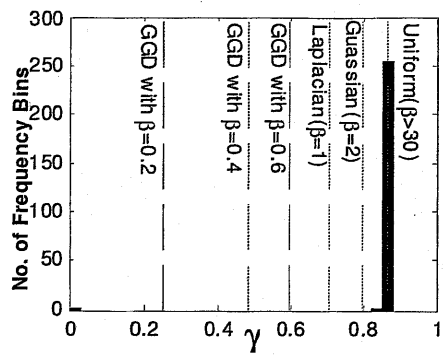
(a) Real part, Male speaker at Mic2.



(b) Imaginary part, Male speaker at Mic2.



(c) Polar Magnitude, Male speaker at Mic2.



(d) Phase, Male speaker at Mic2.

Figure 4.10 Histograms of the moment test for the real part, imaginary part, polar magnitude and phase of the speech from male speaker at Mic2.

for the real part, imaginary part, and polar magnitude in the frequency bin of 703.12 Hz. QQ-plots are drawn for the quantiles of $\mathbf{Z}(f)$, which are computed using Eq.(4.43), and quantiles of theoretical distributions GD, LD and GGD with the estimated parameters, computed by their respective QPF from Eq.(4.44) -(4.47).It can be found from plots for the real and imaginary parts that GGD provides a superior linearity (regression line is shown as dashed line along the plot) over that of GD or LD. For the polar amplitude GGD and LD both provide a

comparable linearity in the plots for both speakers. QQ-plots for other reference signal also have same trend. In Figure 4.12, the QQ-plots for *Z(f)* for two different frequency bins show that data in different frequency bins do not necessarily have the same distribution. It also agrees with the value of the shape parameter, which is not same for all frequency bins. Therefore, it is strange to consider fix PDF for all frequency bins. However, same PDF has been assigned to every frequency bin in majority of FDICA algorithms.

Results of Chi square Test:

The Chi-square test was performed separately on the real, imaginary and polar magnitude of the Z(f), in every frequency bin, for the speech signals of the two speakers at the second microphone. The results are shown in **Figure 4.13** for the speech from a male speaker at the second microphone. The $\chi 2$ –scores for the GGD, in every case and in every frequency bin, have been found to be less than those for the GD or LD. However, the nearness between the $\chi 2$ –scores for GGD and LD for the polar magnitude of *Z(f)* is more than that for the real or imaginary parts. This characteristic of score complies with the previous results from the QQ-plots and the moment tests.

4.6. GGD Model based Blind Detection of CLT Disobeying TFSS

The important requisition for CLT compliance by the TFSS of the speech data is that the TFSS of each independent speech source should not belong to a stable statistical distribution, because such distributions are closed under linear combination [54][81] and fortunately it is strongly LD and can be better approximated by the GGD which is parameterized by the mean, scale and shape parameter β. The value of shape parameter β decides shape of the distribution. GGD represents Gaussian PDF for $\beta=2$, Laplacian PDF for $\beta=1$, and highly parsimonious PDF for $0<\beta<1$. Since the CLT obeyance or disobeyance is logically related to the Gaussianization of the mixed signal, the change in β and SK of the TFSS can be used to detect CLT obeyance or disobeyance. The shape parameter β and SK can be computed from data.

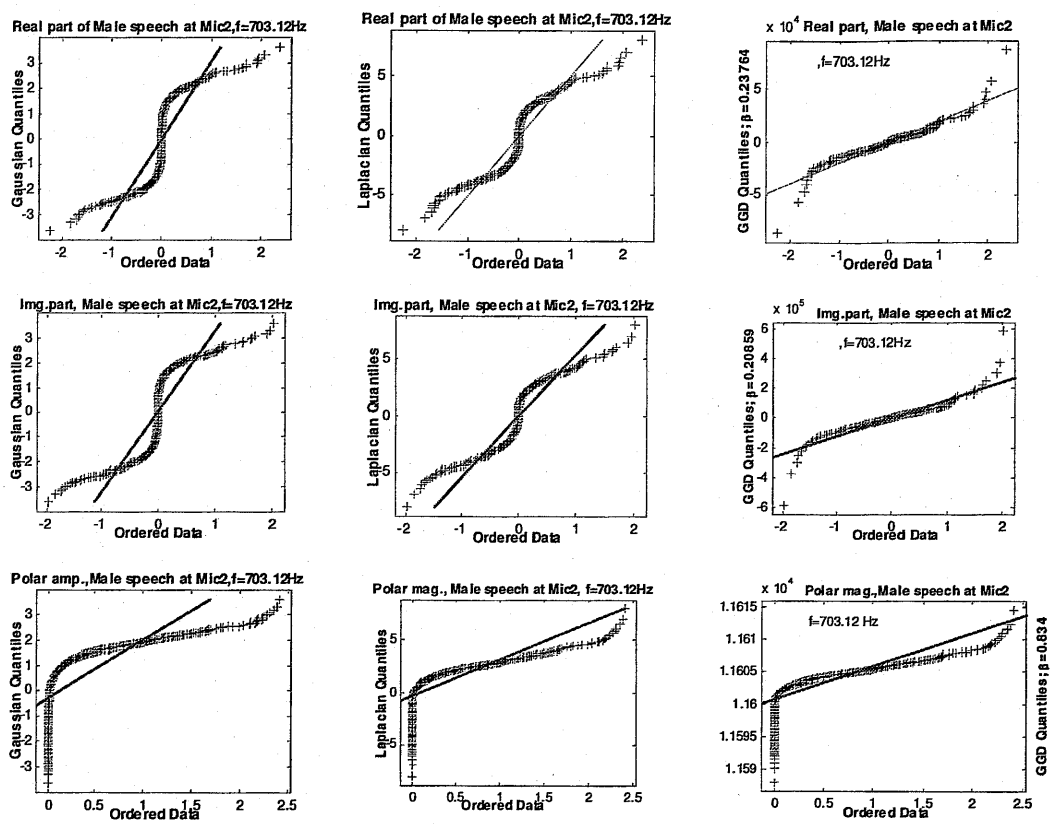


Figure 4.11 QQ-plots for the quantiles of the real part, imaginary part and polar magnitude of Z(f) a f=700 Hz.

The threshold value can be determined by looking into change in Gaussianity of the mixed signal. The relation of β parameter and kurtosis $K(\beta)$ of GGD is given in Eq.(4.51). This relation is monotonic function of the shape parameter such that kurtosis is high for spiky signal (lower value of shape parameter) and is low for higher shape parameter. Kurtosis becomes zero for Gaussian signal. The variation of kurtosis with the value of β is shown in the Figure 4.14. The change

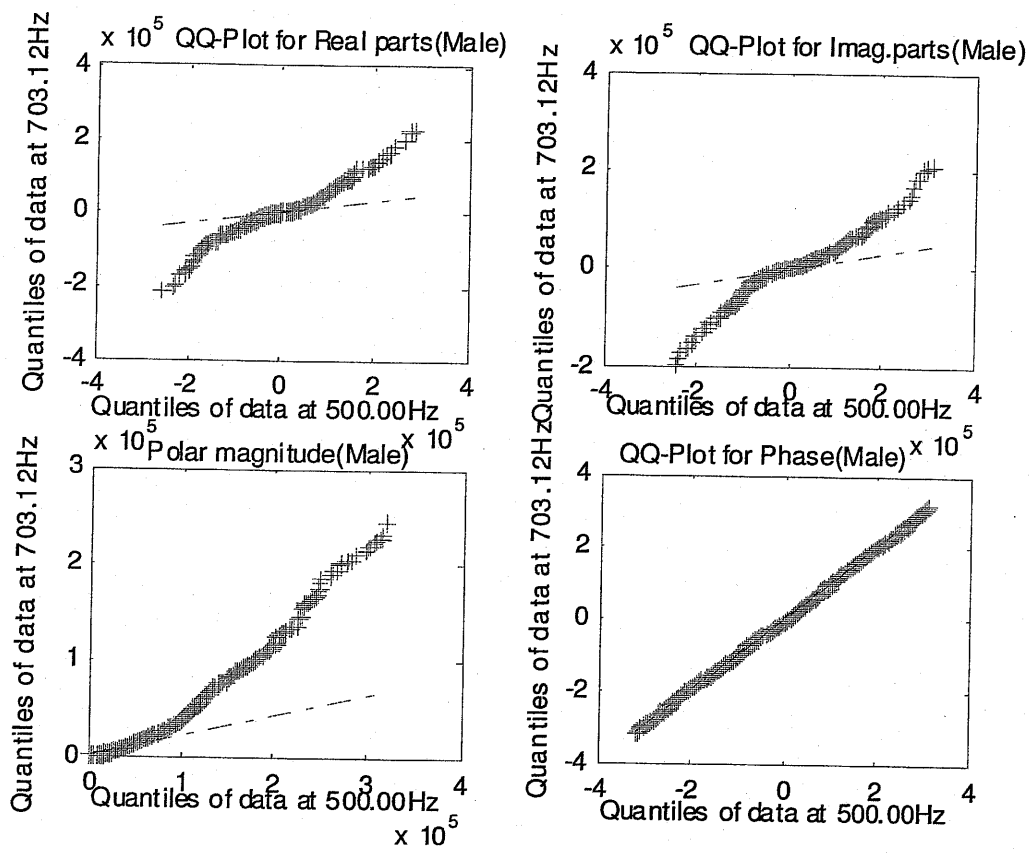


Figure 4.12 QQ-plot of $z(f)$ in two different frequency bins.

in value of kurtosis with β is very steep for super-Gaussian distribution.

$$K(\beta)=\left[\Gamma\left(\frac{5}{\beta}\right)\Gamma\left(\frac{1}{\beta}\right)\right]\left[\Gamma\left(\frac{3}{\beta}\right)^2\right]^{-1}, \tag{4.51}$$

where $\Gamma(x)=\int_0^\infty e^{-t}t^{x-1}dt$=Gamma distribution .

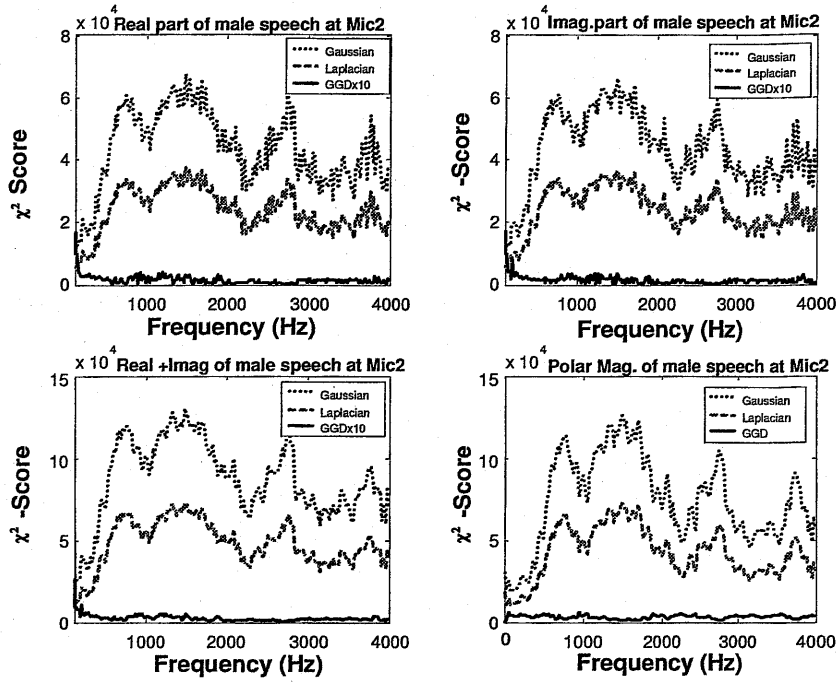


Figure 4.13 $\chi 2$ –score for the real, imaginary and polar magnitude of the male speech at the second Microphone ($\chi 2$ –score for GGD for real, imaginary and (real +imaginary) part is scaled up by 10).

Thus if a TFSS of the mixed signal is fully Gaussian its SK will correspond to $SK_G = K(2) = 3$ in Eq.(4.51), and if it is not mixed signal, the speech will be at least Laplacian or strongly Laplacian for which SK corresponds to $SK_L = K(2) = 6$. For the strongly Laplacian case, which is more accurate as shown in[82], kurtosis will be higher than 6. The SK of TFSS can be directly computed using Eq.(3.43). Thus if SK of TFSS, calculated from Eq.(3.43), lies above SK_L it will represent a Laplacian or strongly Laplacian signal and related TFSS will fail to comply CLT, however, if SK is below SK_L it means signal has gained some Gaussianity due to mixing with other speech signals and so it will comply with CLT. Thus change in

kurtosis can be related to the change in the shape parameter β and some threshold value of it can be used to detect CLT obeying and disobeying sub-bands. The acoustic channel too Gaussianizes speech signal, so the Gaussinity of true speech is less than that of received by the microphones. However, mixing of two-speech signal is bigger effect than the Gaussianization by the channel. Thus the threshold corresponding the $0.5 \leq \beta \leq 1$ can work well [83].

4.7. Results of Combining Null-Beamformer and ICA

As the proposed cause, spectral sparseness of speech signal, of CLT non-compliance by speech mixing, is inherent weakness of speaker, its happening cannot be stopped. The only way is to use the algorithms robust to it or independent from such constraints, or combine some other methods such as NBF having no such problem in the CLT-failing frequency bins. However, this requires the blind detection of CLT obeying and disobeying frequency bins. As discussed in section 4.6, a threshold value of SK or β can be determined for the blind detection of CLT disobeying bins. The relation between CLT disobeying bins and SK can be observed in the right-hand side of Figure 4.14. The left-hand side represents the variation of kurtosis of GGD with β and the right-hand side shows CLT failing bins (gray colored vertical lines in the background) and a plot of SK computed using Eq.(3.43). It is evident that SK is high for the CLT-disobeying bins and is relatively low for the CLT-obeying bins. The dashed horizontal lines across the plots in Figure 4.14 show different threshold values for the different values of β. As the signal is Gaussianized, the value of β shifts towards 2 while for single unmixed speech it is around 1. The blind detection result and true detection result are shown in Figure 4.15. The term true detection represents the result obtained by the verification of the of conditions stated in Eq.(3.45) which needs reference signals from each speaker. However, in the real application, these reference signals are unavailable. The plots in Figure 4.15 show effect of different value of β on the detection accuracy of the blind method. The plot with legend (*Blind-true*) represents the number of uncommon frequency bins by the blind and true detection method. This has minimum value for the threshold around $\beta = 0.6$. Evidently, the blind method falsely detects some bins as

the CLT failing, while giving clean chit to a number of frequency bin, which actually fail. However, for the threshold around $\beta = 0.6$, 70-80 % of bins can be correctly detected. As it is evident from plots for kurtosis and β in the same figure that the slight change in β produces large change in SK, the slight change in the threshold thus can significantly affect the detection accuracy.

The advance information about CLT non-compliance in any bin can be used to stop separation by ICA in such frequency bins and some other alternative method can be used. An experiment to examine such sub-band based combination for NBF and fixed-point FDICA was carried out. The combination strategy for the ICA filter and the NBF filter is complex due to occurrence of CLT-failure in different or same frequency bins at both microphones. Thus there are several ways to combine NBF with ICA. However, in our experiment we replaced the ICA filter by that of NBF if CLT failure in any frequency bin is occurring at either microphone. The separation performance, averaged for four sources, is shown in the Figure 4.16Error! Reference source not found.. It is evident from the figure that the combination shows a significant improvement in the NRR for RT=0, and fails to improve for RT=150 ms and RT=300 ms. The reason for this can be explained with the help of Figure 4.17 and Figure 4.18. These figures show, the spectral NRR under RT=0 ms, RT=150 ms and RT=300 ms, respectively, for ICA only, NBF only and their combinations. It is evident from these figures that the NBF has a better spectral performance under the non-reverberant condition. The performance of NBF degrades as the reverberation time is increased. Under the high reverberation condition, the spectral performance of the NBF is not better than that of the ICA. The spectral performance of both NBF and ICA follow the similar (not exactly same) trend and the overall performance of NBF is worse than that of the ICA. Thus if NBF has a poorer performance than ICA and if the separation filters are exchanged in CLT-failing bins, their combination cannot give any improvement instead it may further degrade the performance. Thus the replacement of the ICA filter by the NBF filter results in poor or unimproved performance. However, in some cases it does improve and for in other cases its performance was found to be worse than

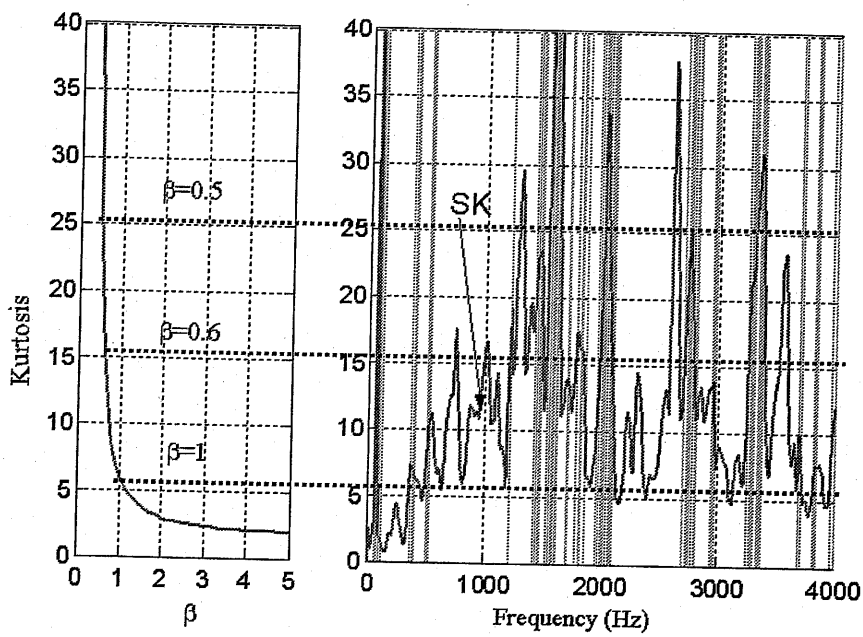


Figure 4.14 Threshold determination for the blind detection of CLT-disobeying TFSS. Left part of the figure shows variation of kurtosis of GGD with shape parameter β while right part shows plot of SK over the CLT disobeying bins (shown as gray vertical lines in the background). The dashed horizontal lines across the two figures are threshold levels used to detect CLT complying and non-complying frequency bins.

that of ICA. Thus combination is effective only under no reverberation or moderate reverberation. The important thing in this context is that FDICA algorithm has to be robust against such phenomena because spectral sparseness is one of the natural characteristics of the speech signal and its happening can't be avoided.

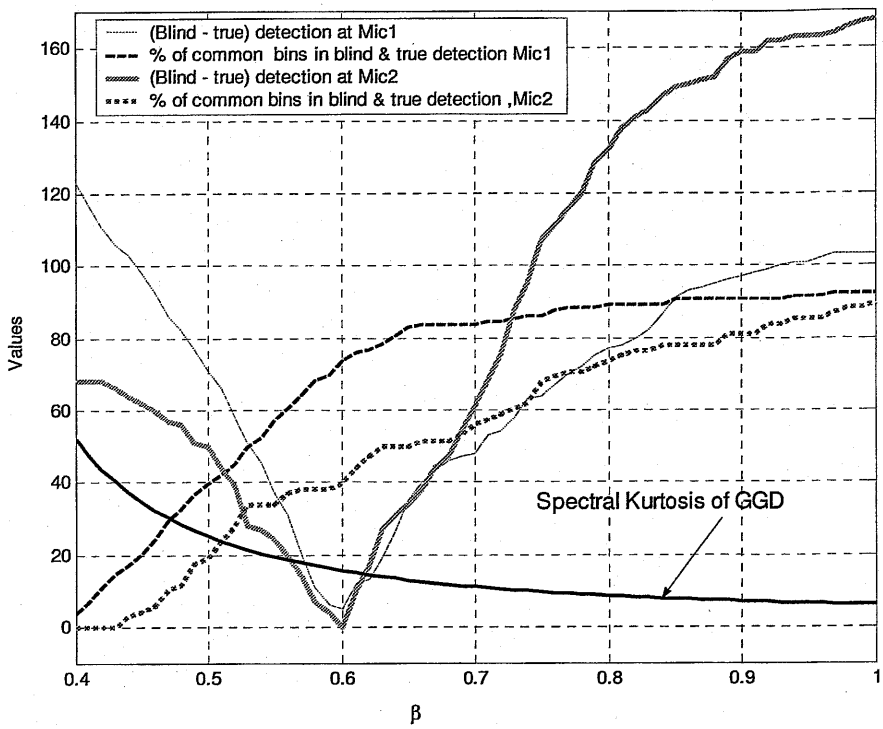


Figure 4.15 Comparison of blind and true detection. Error in detection is shown by the line with legend Blind-true which is minimum for β=0.6, and 70-80% bins are correctly detected.

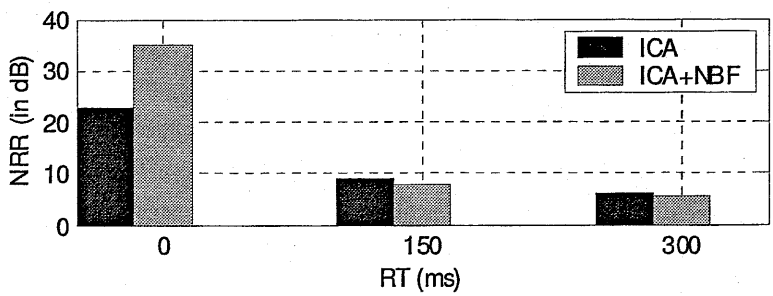


Figure 4.16 Overall NRR averaged for four combinations of the speech data.

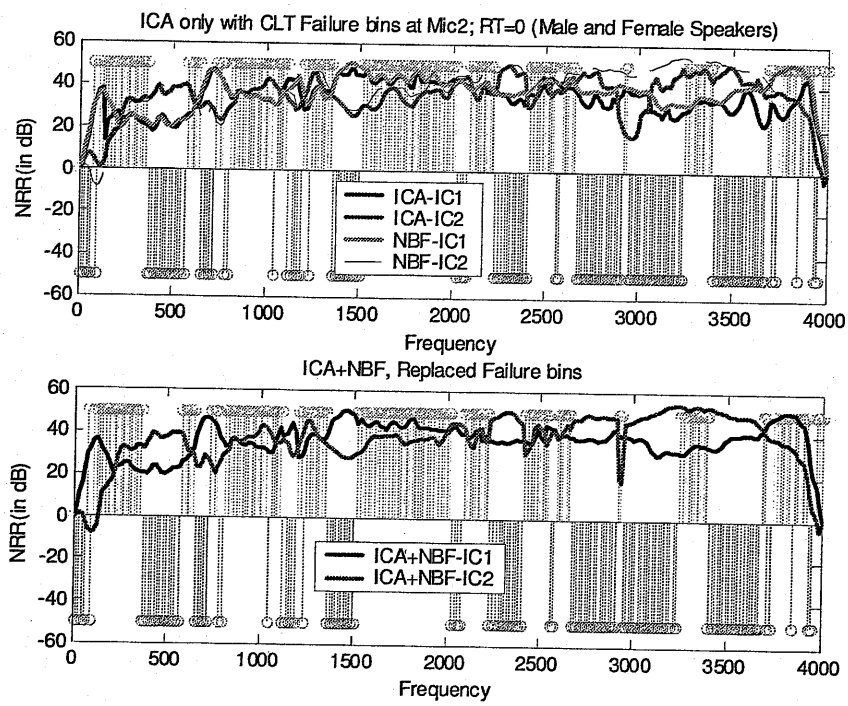


Figure 4.17 Spectral NRR obtained using ICA, NBF and ICA+NBF. The positive stem plots show the CLT-obeying frequency bins while negative stems show CLT-disobeying bins (Voice from male and male speaker).

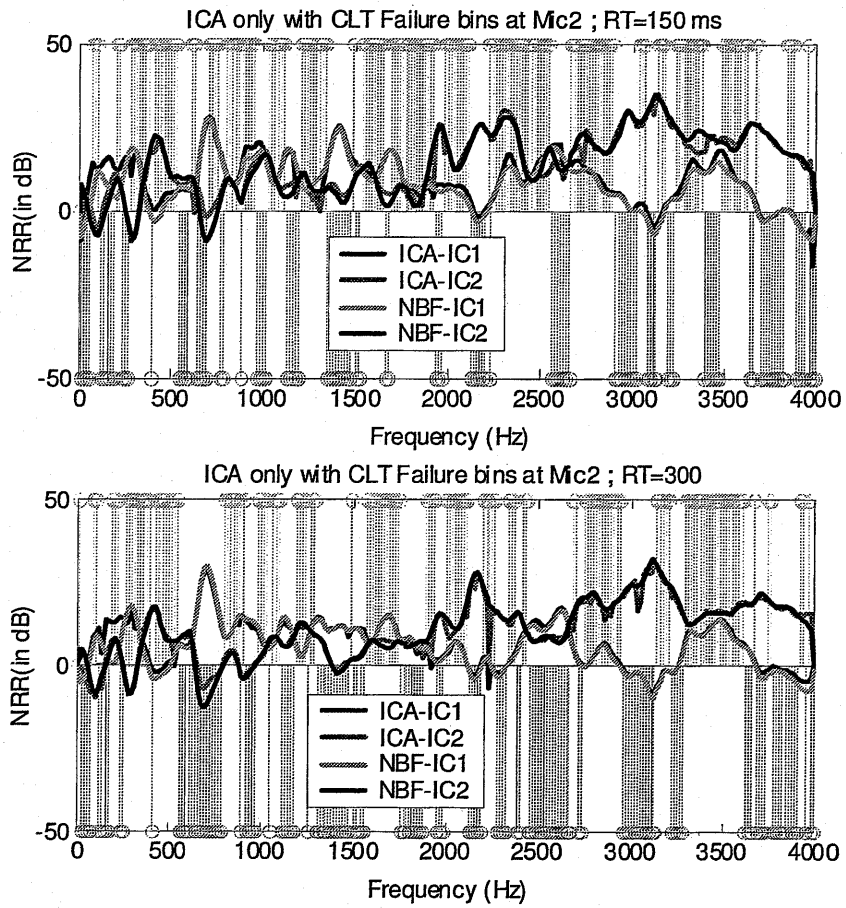


Figure 4.18 Spectral NRR for RT=150 ms and RT=300 ms. CLT-disobeying and CLT-obeying frequency bins are shown with negative and positive stems, respectively (both speakers are male).

Chapter 5

GGD based Negentropy Approximation and Application in Fixed-point FDICA

5.1. Introduction

In the fixed-point ICA by negentropy maximization, negentropy of the data, approximated using generalized Higher Order Statistics (HOS) of the non-quadratic non-linear function, is used as a measure of non-Gaussianity. The choice of non-linear function for negentropy approximation is a crucial task and is highly dependent on the PDF of the TFSS of the data [19][36]. In previous chapter it has been shown that the statistical distribution of TFSS in each frequency bin is not same and can be better approximated by GGD function against the most commonly used PDF of LD and GD functions. Despite, many general purpose non-linear functions have been proposed and have been used in speech signal separation as discussed in Chapter 3 of the thesis. Based on the study in Chapter 4, the issues of this chapter are focused on the novel research questions such as can negentropy approximation by different non-linear functions influence the separation performance of the fixed-point FDICA algorithm and if GGD based function is better approximation of underlying PDF of the TFSS does a non-linear function based on it show superiority in separation too? Accordingly, performance of the FDICA algorithm, based on negentropy maximization, under the negentropy approximation of TFSS by conventional non-quadratic non-linear functions and a new non-linear function based on the PDF modeling of TFSS by the GGD function will be examined.

5.2. Approximation of Negentropy of TFSS:

As a measure of non-Gaussianity, negentropy provides better performance than others such as kurtosis. [19]. As defined in Chapter 3, the term negentropy represents negative of entropy. The negentropy *J(y)* of a random variable y, is given by (reproduced from previous Chapter 3)

$$J(y) = H(y_{gauss}) - H(y), \tag{5.1}$$

where *H(.)* is the differential entropy of (.) and y_{gauss} is the Gaussian random variable with the same covariance as of *y*. As among the distributions of given covariance a Gaussian distribution represents distribution of maximum entropy, the definition of negentropy in Eq.(5.1) ensures that it will be zero (minimum) if *y* is Gaussian and will be increasing if *y* is becoming non-Gaussian. Thus negentropy based contrast function can be maximized to obtain optimally non-Gaussian components. However, estimation of true negentropy, as in Eq.(5.1), is difficult and it requires knowledge of probability density function of the data. However, it is possible to use some approximation of it and several approximations for negentropy estimation have been proposed and used. The moment based approximation of negentropy is given as [17][19]

$$J(y) = \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} kurt(y)^2 + \frac{7}{48} E\{y^3\}^4 - \ldots\ldots, \tag{5.2}$$

where *kurt*(.) represents kurtosis of (.). But this is equivalent to kurtosis which is very raw, loose and rough approximation; however it is also extensively used as a non-Gaussianization measure for ICA algorithm[17][19] [25]. The other more accurate approximations have been based on the use of generalized HOS of some non-linear non-quadratic functions G(y). In terms of such a function the most widely used approximation of negentropy is given in Eq.(3.11) which is reproduced for convenience.

$$J(y) = \sigma[E\{G(y) - E\{G(y_{gauss})\}]^2, \tag{5.3}$$

where σ is a positive constant and y_{gauss} is a Gaussian random variable with same covariance as that of y. In Chapter-3 based on such approximation of negentropy a deflationary learning rule for the separation vector was derived in Eq.(3.22) which involves derivatives of the first and 2nd order denoted, respectively, by $g(y)$ and $g'(y)$, of the used non-linear function $G(y)$.

The separation performance of the fixed-point algorithm depends on the used non-quadratic non-linear function $G(y)$. It is desirable that the function $G(y)$ should provide robustness toward outlier values in the data and should provide better approximation to true negentropy. For the better robustness to outliers $G(y)$ should show slow variation with respect to change in data and at the same time very close approximation of negentropy can be expected if statistical characteristics of $G(y)$ inherit PDF of the data. The statistically efficient and optimal $G(y)$ that can accommodate maximum information about HOS of the data is chosen as the function that can minimize trace of the asymptotic variance of $\mathbf{w}$. The trace of asymptotic variance of $\mathbf{w}$ for the estimation of source s_i is given by

$$V_G = C\frac{E\{g^2(s_i)\} - (E\{s_i g(s_i)\})^2}{(E\{s_i g(s_i) - g'(s_i)\})^2}, \tag{5.4}$$

where constant C depends upon the mixing matrix. As shown in [19] the value of V_G is minimized if the chosen non-linear function G is of the form

$$G(y_i) = c_1 \log p(y_i) + c_2 y_i^2 + c_3, \tag{5.5}$$

where c_1, c_2, c_3 are arbitrary constants. Again, from Eq.(5.5) a simplified form of function G for TFSS can be taken as by truncating higher order term and can be given as follows

$$G(y_i) = c_1 \log p(y_i) \tag{5.6}$$

where c_1 is an arbitrary constant and $p(y_i)$ represents PDF of y_i. The optimal function based on GGD, denoted by $G_3(y)$, can be obtained by using GGD function of Eq.(4.10) for $y = |\mathbf{w}^H \mathbf{X}_w|^2$ and is given by (subscript i is dropped hereafter)

$$G_3(y) = \alpha^{-\beta} |y|^{\beta} + \log A. \tag{5.7}$$

The statistical characteristics of the function depend on the value of shape and scale parameters. The non-linear function in Eq.(5.7) has been plotted in Figure 5.2 for different values of the shape parameter. The value of functions are normalized. Its 3-D shapes are plotted in Figure 5.2. Its smoothness changes with change in the value of shape parameter such that it is less smooth for lower values of shape parameters.

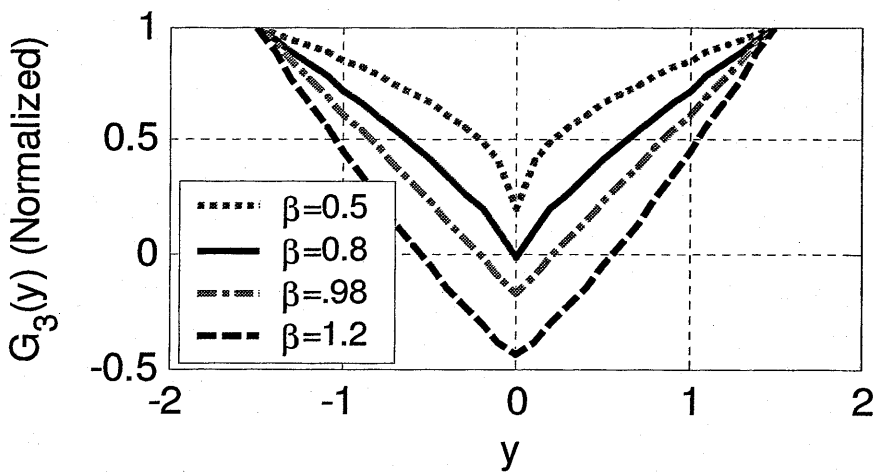


Figure 5.1 GGD based non-linear functions for different values of the shape parameter β. For the lower values of β the non-linear behavior shown by function is less smooth.

The other functions have also been proposed and one of them has been used in Chapter 3. Here, too, they are cited again for convenience and comparative study of performance. For the super-Gaussian signals following functions has been recommended [19] and have been used in the speech signal separation [28][29]

$$G_1(Y) = \log(a_1 + Y); a_1 = 0.01, \quad (5.8)$$

$$G_2(Y) = \sqrt{a_2 + Y}\ ; a_2 = 0.01. \quad (5.9)$$

5.3. Error Estimation in Negentropy Approximation

In order to judge the relative suitability of these non-linear functions we will evaluate their performance for negentropy approximation and robustness to outliers, and capacity of signal separation. The statistical technique of Jackknifing can be used to evaluate relative error in the approximation of negentropy and robustness to outliers [84]. Jackknife is one of the powerful tools for the data partitioning and can be used to estimate bias and standard error occurring in negentropy approximation by the non-linear functions $G_k (for\ k = 1,2,3)$ from Jackknife replicates. The Jackknife replicates for the negentropy are obtained by approximating negentropy of Jackknife samples which are created by omitting, in turn, one data sample from the original TFSS. Let us consider the TFSS in any frequency bin f consisting of U samples. The ith Jackknife replicate for negentropy approximation by function G_k is given by

$$J_k^{(-i)}(f) = G_k([Y^2(f,1), Y^2(f,2) \ldots Y^2(f,i-1), y^2(f,i+1) \ldots Y^2(f,U)]), \quad (5.10)$$

and this is carried out independently in each frequency bin for each sample. The bias $J_k^B(f)$ in the negentropy approximation by function G_k is given by

$$J_k^B(f) = (N-1)\{\bar{J}_k^T(f) - J_k(f)\}, \quad (5.11)$$

where $\bar{J}_k^T(f) = \frac{1}{N}\sum_{i=1}^{R} J_k^{(-i)}(f)$. The standard error in negentropy approximation by G_k is given by

$$\hat{J}_k^{SE}(f)=\left[\frac{(N-1)}{N}\sum_{i=1}^{R}\{J_k^{(-i)}(f)-\bar{J}_k^{T}(f)\}^2\right]^{0.5}. \tag{5.12}$$

This represents standard deviation of the Jackknife replication, however, it is unbiased due to the presence of factor (N-1)/N [85]. Since TFSS in each frequency bins are assumed to be independent the above estimates for bias and standard error can be averaged over the no. of frequency bins and can be given by

$$\bar{J}_k^{B}=\frac{2}{P}\sum_{i=1}^{P/2}J_k^{B}(f) \text{ and } \bar{J}_k^{SE}=\frac{2}{P}\frac{2}{P}\sum_{i=1}^{P/2}\hat{J}_k^{SE}(f). \tag{5.13}$$

5.4. FDICA with Flexible Non-linearity

The most important thing for any contrast function is it separation capacity. However, if the contrast function inherits maximum statistical information of the data, it may provide better separation [19][34]. The separation performance of each non-linear functions will be judged using the deflationary learning rule given in Eq.(3.22). Obviously, that requires first and second order derivatives of the non-quadratic functions $G_k(y)$ which are given by

$$g_1(y)=(a_1+y)^{-1} \text{ and } g_1'(y)=-(a_1+y)^{-2}, \tag{5.14}$$

$$g_2(y)=0.5(a_2+y)^{-0.5} \quad g_2'(y)=-0.25(a_2+y)^{-3/2}, \tag{5.15}$$

$$g_3(y)=-\beta\alpha^{-\beta}\left[|y|^{\beta-1}\,sign(y)\right], \tag{5.16}$$

$$g_3'(y)=-\beta\alpha^{-\beta}\left[|y|^{\beta-2}+y^2(\beta-2)|y|^{\beta-4}\right]. \tag{5.17}$$

As a performance measure NRR, SCRF and number of iteration consumed by algorithm to converge, under the given stopping criterion δ, will be used. The number of iteration taken by the algorithm depends on the nature of convergence. The nature of convergence depends upon chosen non-linear function *G(y)* and on existence of its higher order derivatives. It can be shown that the value of diminishing component of the separation vector, denoted by w^+ after one iteration, is given by [19]

$$w_i^+ = \frac{1}{2}E\{y_i^3\}E(g''(y_1)\}w_i^2 + \frac{1}{6}kurt(y_i)E(g'''(y_1)\}w_i^3 + \ldots\ldots\ldots; \text{for i>1} \quad (5.18)$$

This equation includes higher order derivatives of $G(y)$ as the coefficient of error terms. It can be imbued that if the 3rd order derivative $g''(y)$ of G(y) vanishes i.e. $E(g''(y)\}=0$ the convergence becomes cubic and is governed by the value of 4th order derivative $g'''(y)$ and so on. The 3rd and 4th order derivatives of the used non-linear functions are given by

$$g_1''(y) = 2(a_1 + y)^{-3}; g_1'''(y) = -6(a_1 + y)^{-4}, \quad (5.19)$$

$$g_2''(y) = .38(a_1 + y)^{-2.5}; g_2'''(y) = -.95(a_1 + y)^{-3.5}, \quad (5.20)$$

$$g_3''(y) = K_1\left[|y|^{\beta-3} + 2|y|^{\beta-5} + (\beta-4)y^2|y|^{\beta-5}\right]\text{sgn}(y),$$

$$g_3'''(y) = K_1[|y|^{\beta-4} + (\beta-4)y|y|^{\beta-5}\, sign(y) + 2|y|^{\beta-6} + 2y(\beta-6)|y|^{\beta-7}\, sign(y) + 3(\beta-4)y^2|y|^{\beta-6} + (\beta-6)y^3|y|^{\beta-7}\, sign(y)], \quad (5.21)$$

where $K_1 = -\beta(\beta-2)\alpha^{-\beta}$. (5.22)

In order to avoid singularity of derivatives of $G_3(y)$ at y=0, it is replaced by very small (10^{-4}) number. The parameters of GGD are estimated using maximum likelihood approach as is described in Chapter-3.

5.5. Experiments and results

Experimental setup was same as described in Chapter 3. The experiments were carried out in two parts separately for the jackknifing and blind separation. The TFSS of the speech data were generated by doing STFT analysis of the mixed signals under the signal analysis conditions shown in the Table 3.1. In order to estimate bias and standard error occurring in negentropy approximation by $G_k(y)$ (k=1,2, 3) six unmixed speech signals from different speakers were used. In this analysis unmixed signal were used because in the separation algorithm $G_k(y)$ has to ultimately approximate negentropy of the separated signal which

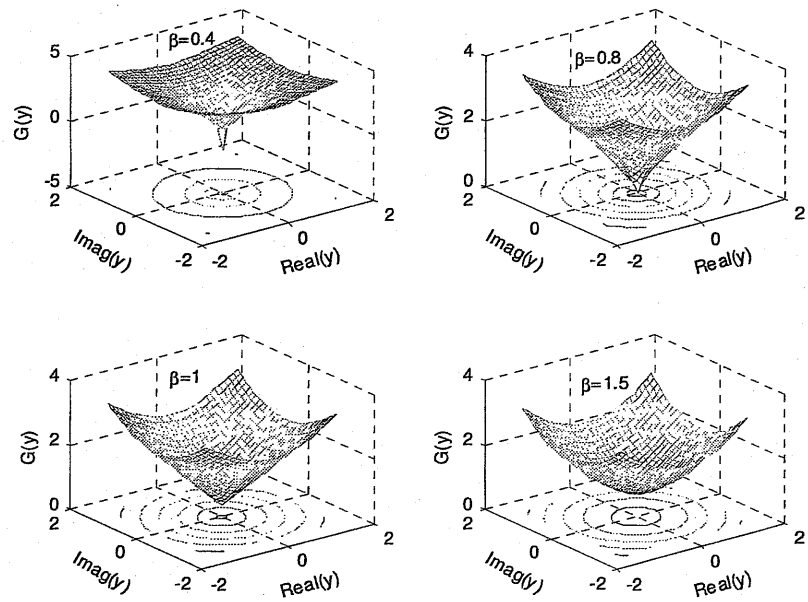


Figure 5.2 3D Shape of the GGD based non linear function for different values of shape parameters.

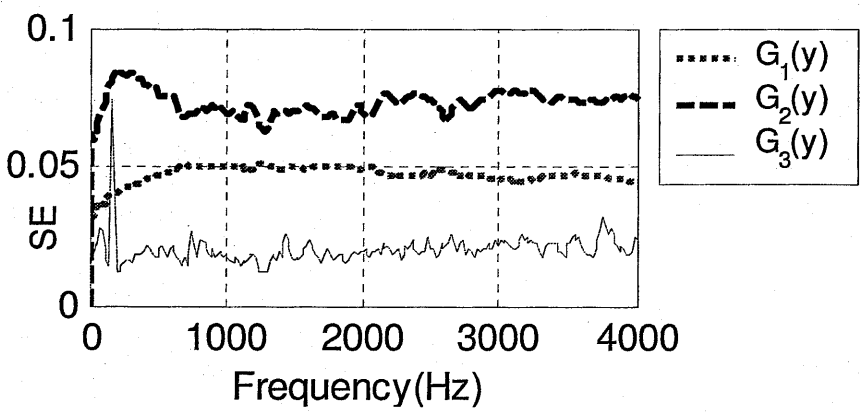


Figure 5.3 Averaged SE ($\hat{J}_k^{SE}(f)$) for different G(y) in different frequency bins.

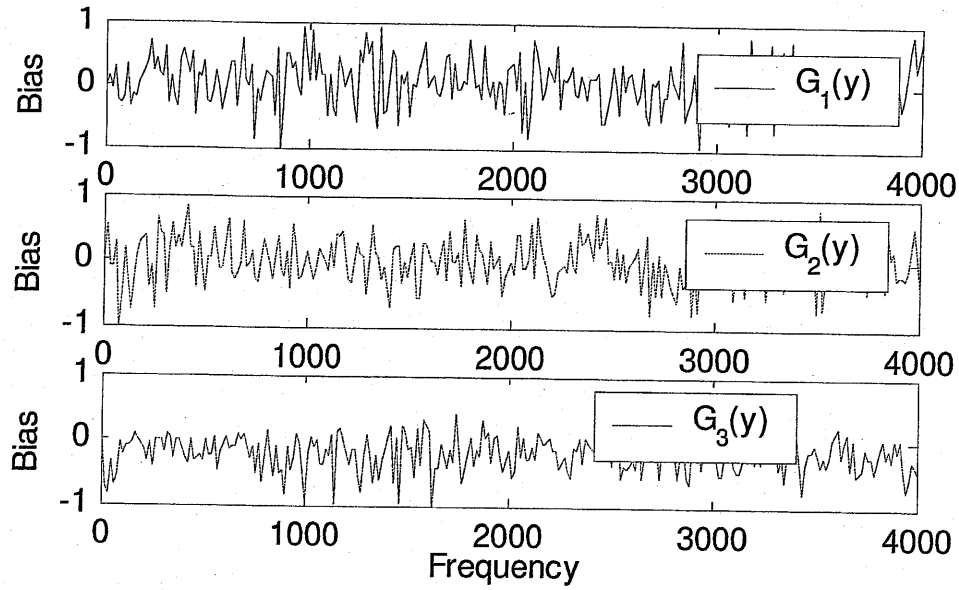


Figure 5.4 Normalized bias $J_k^B(f)$ for different G(y) in different frequency bins.

should be ideally clean and unmixed. The bias and standard error in the negentropy approximation by each of $G_k(y)$ were estimated in each frequency bin using Eq.(5.11) and Eq.(5.12) for sequential delete–one Jackknife method. The estimated standard error, averaged for six combinations of the mixed speech signals including male and female speakers, are compared in Figure 5.3 for each $G_k(y)$ .The averaged bias estimate of $J_k^B(f)$, for different non-linear functions are shown in the Figure 5.4. It is evident from these figures that the standard error and bias is minimum for the GGD based non-linear function which implies that its robustness and closeness to true negentropy of the TFSS signal is better than that of approximated by $G_1(y)$and $G_2(y)$. In the Jackknifing process the value of shape parameter β was fixed at 0.86 under the light of results reported in [82]. The separation performances of the fixed-point FDICA with the use of these three non-linearity functions were also studied under different RTs. The stopping criterion for algorithms was set at $\delta = \| w_{new} - w_{old} \|^2 < .0001$. First the separation performance for different value of β with non-linear

function of Eq.(5.7) were studied. The NRR, which is defined in Eq.(3.46) , SCRF $\gamma(f)$ of Eq.(3.47), and no. of iteration taken to converge up to satisfaction of δ were used as the performance measures. The learning rules of was initialized using null-beam former based value of the separation vector. The results of NRR, $\gamma(f)$, and no. of iterations, averaged for the six combinations of mixed speech data are shown in Figure 5.6, Figure 5.7 and Figure 5.8 respectively. In Figure 5.6 the separation performance is found to be optimum for β values between 0.8 to 0.98 for reverberant and non-reverberant acoustical conditions, however, NRR is very low in the reverberant conditions. Similar trend can be observed in Figure 5.7 for SCRF graphs. It is important to point out that NRR and SCRF performance figures are good for those values β that are close or equal to values of shape parameters corresponding to PDF modeling of TFSS. This is indicative and in supplementation of the fact that for the better separation used non-linear function should be in possession of maximum statistical information about the data [19][34]. It is evident from Figure 5.8 that the no. of iteration too is varying with the shape parameter. For the very lower values of shape parameter the no. of iterations taken increases highly but with the increasing value of β , no. of consumed iterations decreases. The result is interesting in the sense that for the shape parameter value β=2, corresponds to GD, no. of consumed iterations by algorithm is lower than that of for β values representing the PDF of TFSS. However, NRR is getting low. The reason behind this can be understood with help of Eq.(5.18) which shows how the fixed-point algorithm converges to the optimum separation vector. The coefficients of 2nd and 3rd terms of Eq.(5.18) have been plotted in Figure 5.5 for different value of β for non-linear function as well as for data for which shape parameter has been denoted by β_y. The used data were artificially generated by fitting GGD parameters with zero mean and unit variance. It is evident that the annihilations of third order and higher order derivatives are starts earlier and is faster for the higher value of shape parameters which ensures higher convergence speed with increasing values of β. In order to compare the separation performance of all the three non-linear functions another experiment was performed by changing

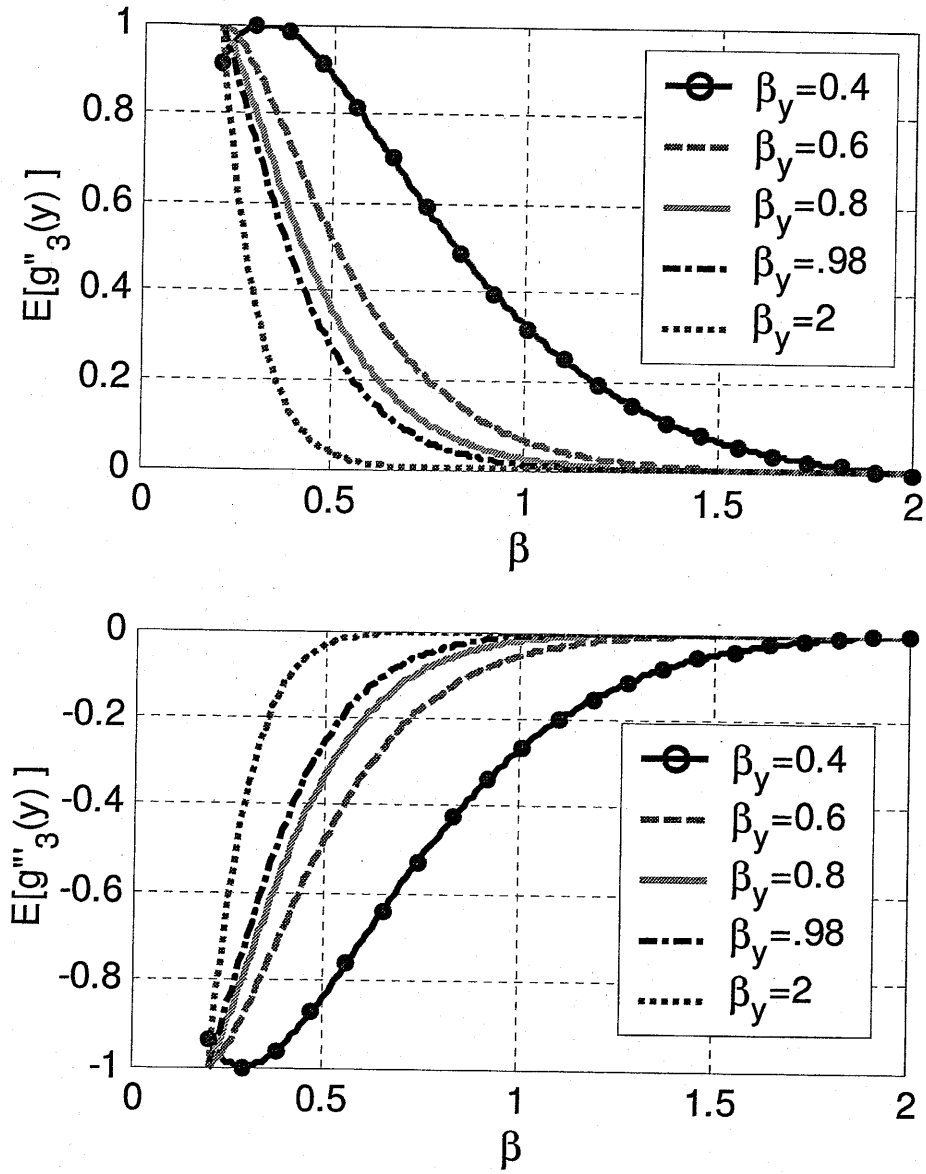


Figure 5.5 Showing normalized mean of 3rd and 4th order derivatives of non-linear function $G_3(y)$. This shows how quickly these terms are vanishing for different value of β which results in different convergence speed. For $\beta=2$ the proposed function $G_3(y)$ acts as a kurtosis and shows cubic convergence.

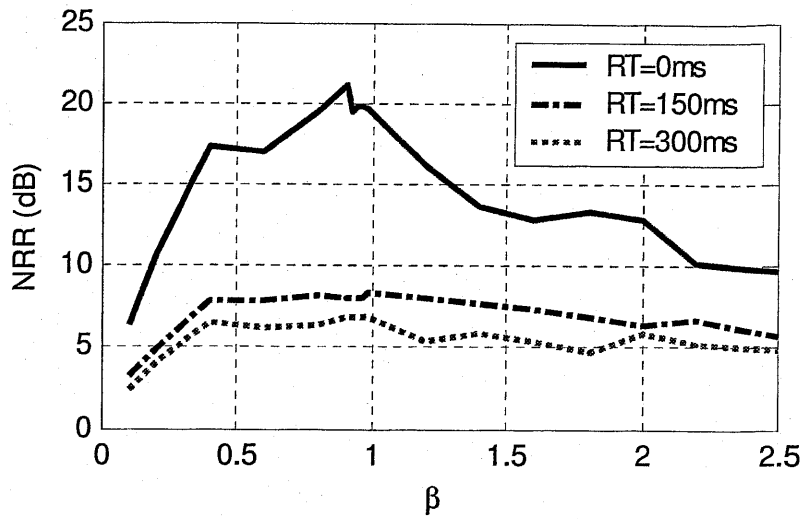


Figure 5.6 Showing averaged NRR for different RT for different value of shape parameter.

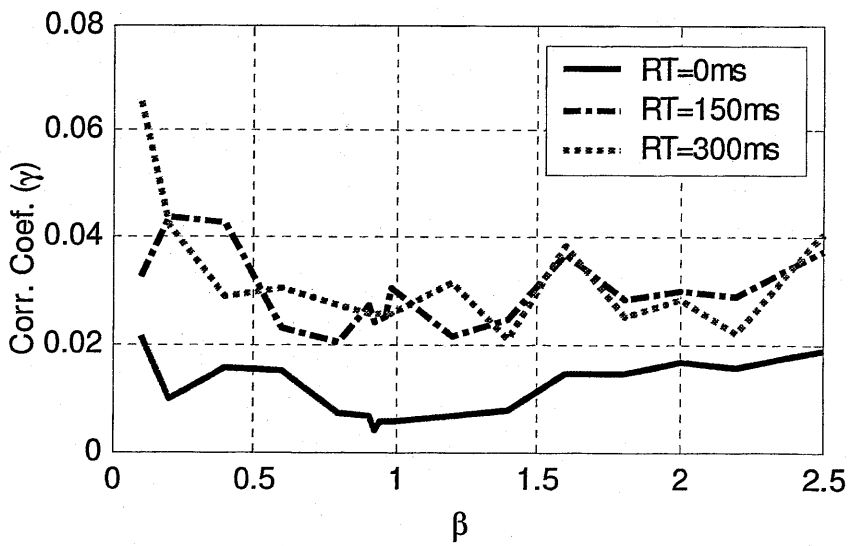


Figure 5.7 for different value of shape parameter β in different frequency bins.

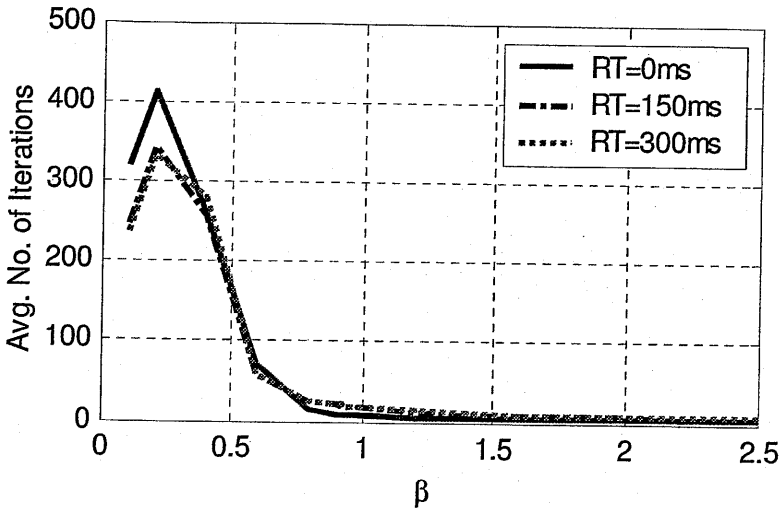


Figure 5.8 Average Number of iteration taken in separation in different frequency bins.

the non-linear functions in the learning rule. The value of parameter of the GGD function is estimated after each iteration, however, the shape parameter was fixed to $\beta = 0.9$ following the above results. The algorithm was initialized by the null-beam former based initial values of the separation vectors. The averaged NRR and no. of iterations consumed, for 6 combinations of mixed signals, are plotted in Figure 5.11and Figure 5.9 respectively. It is evident from these figures that there occurs no significant difference in the achieved NRR, however, significant difference occurs in the number of iterations consumed by different non-linear functions. In this respect, the GGD based non-linear function outperforms the other two with handsome margins for no. of consumed iterations in both the reverberant and non-reverberant conditions. The GGD based functions shows higher convergence speed because the third order derivative and 4th order derivative for it are much less than that of for $G_1(y)$ and $G_2(y)$. These derivatives control quadratic and cubic convergence of the algorithm and are shown in Figure 5.9 for all the three non-linear functions with data with different statistical distribution. It is evident from there that $G_3(y)$, with β=0.90, has very low value of these derivatives in comparison to that of for $G_1(y)$ and $G_2(y)$.

This conspicuous feature ensures higher convergence speed for it. The proposed non-linear function based on the statistical modeling of TFSS by GGD function is adaptive in the sense that it depends on the parameters of the data and accordingly provides non-linear behaviors. Favorable, results for the proposed non-linear functions for spectral separation shows effectiveness of statistical

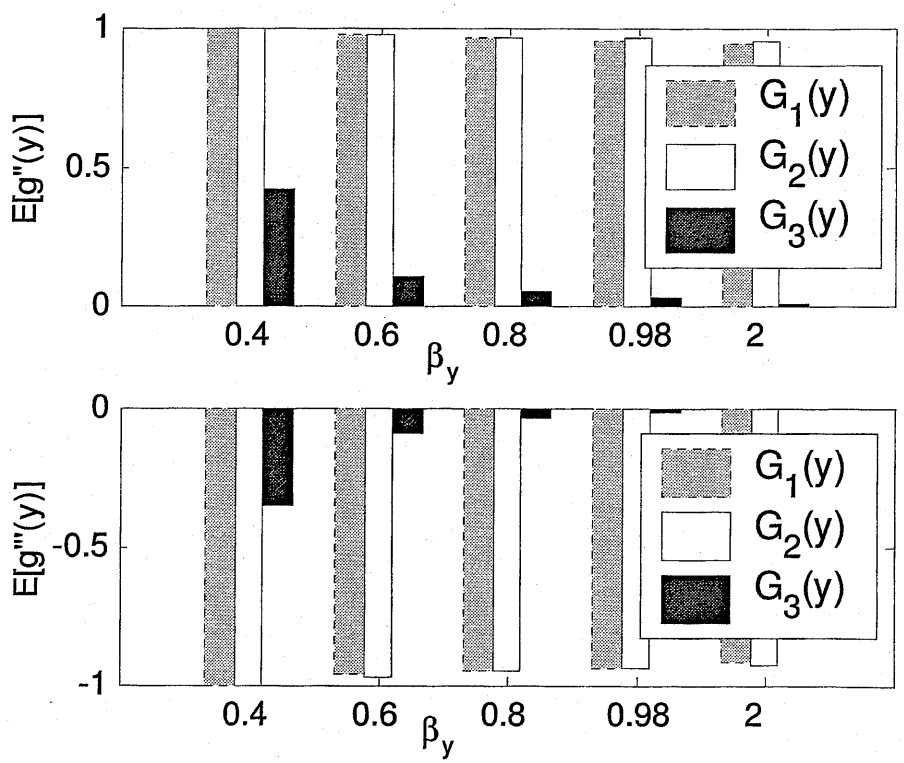


Figure 5.9 These bar plots show normalized mean of 3rd and 4th order derivatives $E\{g''(y)\}$ and $E\{g'''(y)\}$ respectively for different types of synthetic data with different shape. The β for GGD based G(y) is 0.9.

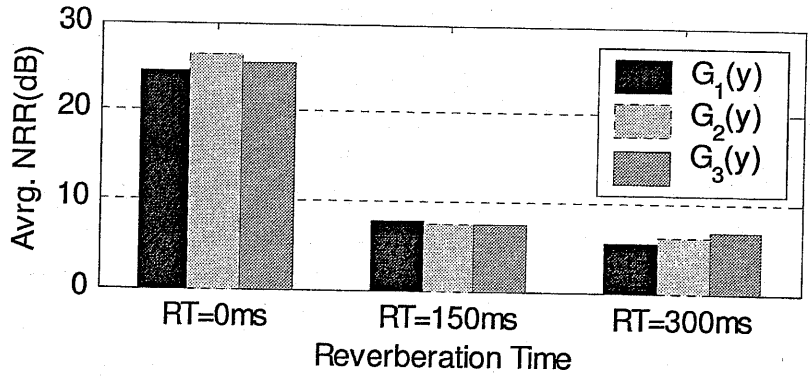


Figure 5.10 Averaged (for 6 pairs) NRR for different G(y) under different RT.

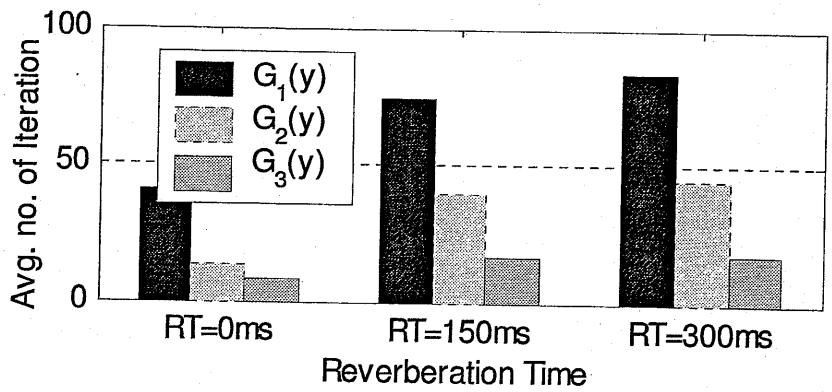


Figure 5.11 Averaged (for 6 pairs) NRR for different G(y) under different RT.

modeling of the TFSS by GGD function. It can be concluded that as the GGD function can better represent statistical model of TFSS, the GGD based

non-linear function can incorporates much information about HOS of the TFSS. Due to this it provides better results than the conventional non-linear functions.

Chapter-6

Enhancement of Separated Independent Components

6.1. Introduction

In this chapter a novel method for denoising speech signal in DFT domain is presented. In general it is a speech enhancement technique that can work for noise with different statistical distributions. The proposed denoising method will be also applied to denoise separated independent components. The idea of denoising ICs is based on the signal mixing model of Eq.(3.7). Under no background noise, it can be assumed that the only source of noise in the separated components is the residual speech signal from other sources which is present even after separation. Under such circumstances one separated independent component can be assumed to be corrupted by the other. Thus for one independent component, other components are assumed to be source of noise and accordingly a novel denoising algorithm will be presented using GGD based statistical modeling of TFSS of both sources.

6.2. Working Signal model

In this chapter, too, the signal model of Eq.(3.7) will used , however, it is essential to give some explanation in the context of enhancement algorithm for speech signal in the DFT domain. The signal model of Eq.(3.7) expresses that the mixed signal in frequency domain is just superposition of spectral contribution of each source in every frequency bin. The output of the FDICA gives independent components that contain interference from others in the residual form which is only source of contamination. This gives additive noise like model

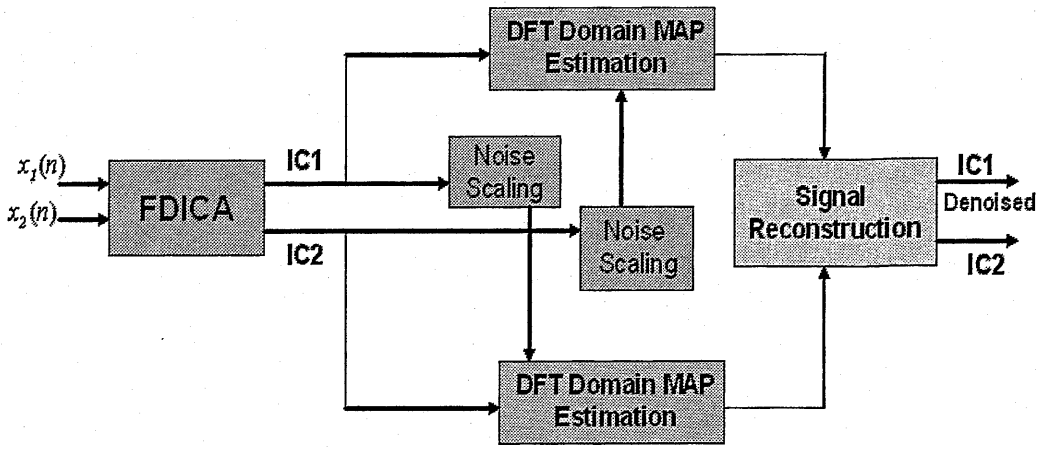


Figure 6.1 Showing denoising scheme for ICs obtained from FDICA.

for each IC and it can be cleaned by single channel enhancement algorithm used for removal of additive noise. Accordingly, we will start from here with speech signal contaminated by additive noise and develop a noise suppression rule for it that will be applied to clean ICs. For the single channel signal capture, the observed speech $y(n)$ in the presence of additive noise $d(n)$ is given by

$$y(n) = x(n) + d(n), \tag{6.1}$$

where $x(n)$ represents clean speech signal, n is the time-index, and random noise $d(n)$ is uncorrelated with the clean speech signal. The aim of the enhancement technique is to estimate clean signal $\hat{x}(n)$ from the observed noisy signal $y(n)$. As we said earlier that our aim is to do estimation in the DFT domain, where DFT coefficients of the clean speech are estimated. The observed speech signal is subjected to STFT analysis, as depicted in Figure 3.2, to produce TFSS. The TFSS $Y_i(f)$ of the ith independent component in any frequency bin f is supposed to be composition of original contribution $X_i(f)$ and cross-channel interference signal $D_j(f), i \neq j$; and it can be represented as follows

$$Y_i(f) = X_i(f) + D_j(f), \text{ for } i \neq j. \tag{6.2}$$

The interference signal component $D_j(f)$ is derived from the independent component $y_j(n)$ by scaling down in accordance with the NRR achieved by the FDICA and by doing STFT analysis. This can be expressed as

$$D_j(f) = STFT\left[\frac{\sigma_r}{\sigma_{y_j}} y_j\right], \tag{6.3}$$

$$\text{where } \sigma_r = \frac{\sigma_{y_i}}{\sqrt{10^{\frac{NRR_i}{10}} + 1}}.$$

In generating contaminating noise level like above it is assumed that the independent component y_j does not contain contribution of y_i but practically it is not so because y_j is also contaminated by y_i depending upon its NRR ensured by FDICA. However, hereafter we will go further with such assumptions. The aim of the enhancement algorithm is to make modification by some function $G(f)$, known as a noise suppression rule, to estimate the spectral component $X_i(f)$ of the clean speech. i.e.

$$\hat{X}(f) = G(f).Y(f). \tag{6.4}$$

The modification function $G(f)$ is also called gain function. Its value lies between 0 and 1 meaning there by it produces more suppression for the lower SNR of the input and less suppression for the inputs with higher SNR. Thus the problem of enhancement of ICs in DFT domain is reduced to find a suitable function $G(f)$ that can reduce the residual interference signal available in any IC. This problem is not new. This is one of the very old problems in the area of speech enhancement but still challenging and chasing the state of art ASRs. This problem has been addressed in many recent research reports and books [49] [86]. In the speech enhancement landscape, the basic assumption under such methods

is that only contaminated signal is available to the enhancement system and thus a classical adaptive noise canceling techniques using reference noise are useless [88][89] in such scenario. There has been development of different algorithms for the enhancement of speech signal, corrupted by broadband noise, based on the short-time analysis of the signal in the frequency domain. Such algorithms are able in accessing and manipulating each spectral component of very short-segments of speech. There have been developments of different algorithms in the DFT domain to enhance the magnitude of spectral components. The report of pioneering effort in this direction appeared in [86]. After then there came many algorithms for enhancement in the DFT domain as the variants of popularly known technique of spectral subtraction [90][91][92]. In the spectral subtraction the Short-Time Spectral Amplitude (STSA) of clean signal is estimated from that of noisy signal and combined with the phase of STSA of noisy signal to get spectral components of enhanced signal [87]. There have been developed speech enhancement algorithms using estimation techniques such as Maximum Likelihood and MAP estimation [92]. The other most important algorithms were developed based on Gaussian statistical models for the magnitude of the DFT coefficients of the speech [93][94][95][96][97]. The assumed PDF for the DFT coefficients of speech and noise plays important role in the enhancement algorithm. The Gaussian PDF for speech spectral components was assumed under the implication of the CLT as the DFT coefficients are weighted sum of the random data samples. Speech signal is naturally non-stationary, however, statistical stationarity in speech signal is created artificially by dividing speech signal into very short time segments, which are supposed to be quasi-stationary, and then DFT of each segment is taken to represent signal in the frequency domain. In Chapter-4 we have described lots on the statistical modeling of TFSS and GGD based model were proposed. However, the statistical modeling of spectral component of speech has been controversial since past and different researchers have used different statistical models for the DFT components of speech which have been discussed before in this thesis. However, in the context of speech enhancement we place here some of such applications e.g. authors in

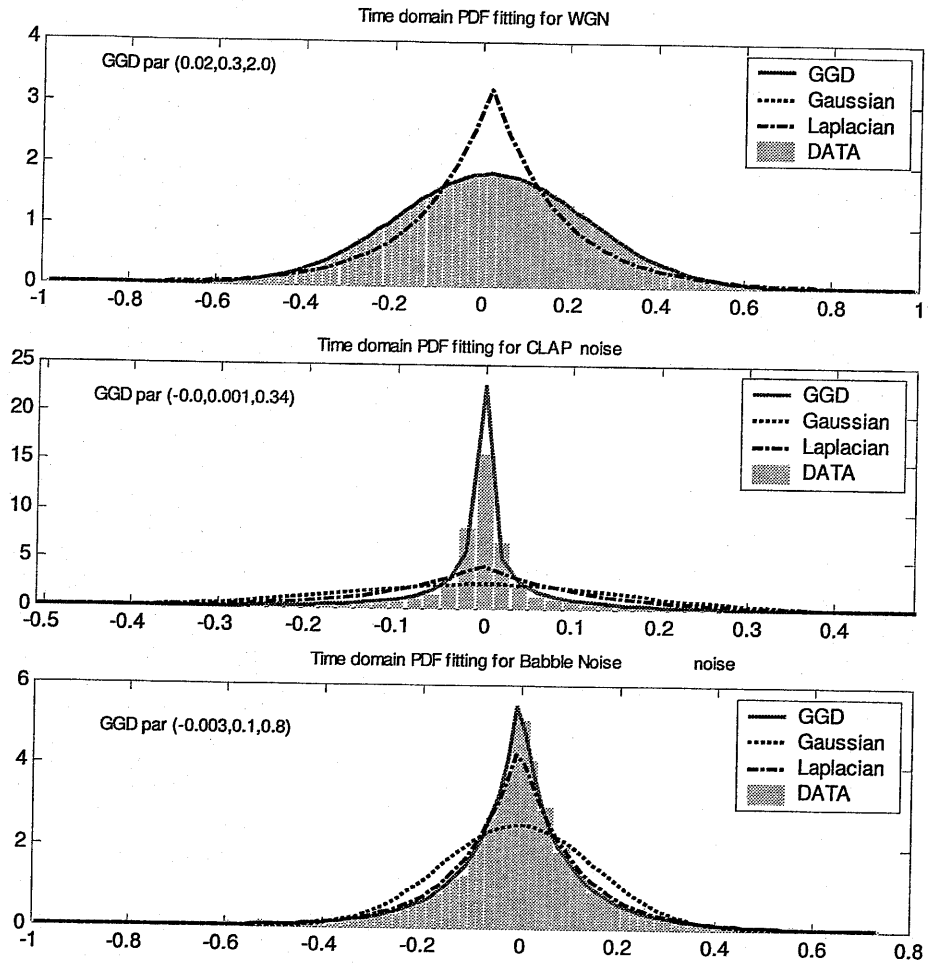


Figure 6.2 Histogram of WGN (Top), Clapping (middle), and babble noise (bottom). The fittings of GD, LD, and GGD function are also shown in the histogram of the noise. GGD parameters (mean, scale, shape) were estimated using ML approach and are also shown.

[61][94] have used Gaussian model. Recently, in [98][99] Laplacian model has been used to derive speech enhancement algorithms. Similar, mismatch between actual and used statistical models for the noise signal also arises. In

many algorithms for speech enhancement e.g. Wiener filtering, in [95], noise and speech both have been assumed to be Gaussian. However, many real world noise signals such as chair crack, clapping, object dropping, babble noise etc. are neither Gaussian nor exactly Laplacian [100]. For example PDFs of three noise signals namely White Gaussian Noise (WGN), babble and clapping noise are shown in Figure 6.2. These figures also contain fittings of the Gaussian, Laplacian and GGD functions. Inspired by these facts on statistics of the spectral components of noise a flexible enhancement algorithm has been proposed here using GGD based statistical modeling. Since GGD based modeling for noise and speech can capture wide range of noise, it can be used to enhance speech signal corrupted by speech like noise.

The denoising situation of the output of FDCA is little bit different with that of enhancement under noise. Under the no external background noise any IC component is considered to be contaminated by scaled version, depending on the achieved NRR by FDICA, of other ICs which are roughly, not exactly, known. Thus in denoising one IC is taken as speech source while other is taken as interference or noise contributing source. It is important to note that noise source is also speech. So here a general method for enhancement in DFT domain is introduced by using GGD models for the TFSS of both ICs. An MAP estimator for the STSA, for speech enhancement in the DFT domain using a flexible GGD function as the prior PDF model for the DFT coefficients of speech will be derived. Also, spectral components of the noise are modeled with GGD.

6.3. Bayesian Estimation

Bayesian estimation is a classical method of statistical estimation that will be used here to develop denoising algorithm for the ICs. In the Bayesian framework the estimate of unknown signal is obtained by minimizing Bayes risk B_r which is given in terms of cost $C(s,\hat{s})$

$$B_r \triangleq E\{C(x,\hat{x})\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(x,\hat{x})p(x,y)dxdy = \int_{-\infty}^{\infty}\left[C(x,\hat{x})\,p(x\mid y)dx\right]p(y)dy, \tag{6.5}$$

where y is the observed data and x is the true value of data hidden in observation y. In the above equation $p(y)$ is non-negative and thus the minimization of B_r puts constraint on selection of $\hat{x}$ which should be chosen such that for every fixed value of y the bracket term in Eq.(6.5) becomes minimum. Thus the minimization of error in estimation in Eq.(6.5) melts down to following

$$\hat{x} \Leftarrow \min_{\hat{x}} E\{C(x,\hat{x}) \mid y\} \tag{6.6}$$

The choice of cost function depends on the problem at hand and leads to different estimation techniques. However, cost functions are chosen to satisfy ones requirements as well as tractable formulation of the problem. In general the cost functions are chosen as the function of error $x_\varepsilon = x - \hat{x}$ in estimation. This makes task of minimizing cost function easier as it becomes on single variable function. Usually, linear, quadratic and uniform cost functions are used. Such cost functions are shown in Figure 6.3. These three cost gives different estimators in terms of median, mean and mode of the posterior PDF as shown in Figure 6.4. Linear cost function varies linearly with the absolute value of error i.e. $C(x_\varepsilon) = |x_\varepsilon|$. In the quadratic cost function the cost is taken as the function of square or error i.e. $C(x_\varepsilon) = x_\varepsilon^2$ and estimate is known as Minmum Mean Squared Error (MMSE) estimate $\hat{x}_{mmse}$ and is given by

$$\hat{x}_{mmse} = \int_{-\infty}^{\infty} xp(x \mid y)dx. \tag{6.7}$$

It is important to note that it is not possible to get tractable solution of the above integral for all types of PDF.

The other very important cost function is uniform cost function. Such cost function assigns zero cost for all error less than some certain value and uniform value for errors outside that limit. Thus under such cost function the estimation is carried out for error lying between very small values $\pm\delta/2$. The cost function is given as

$$C(x_\varepsilon)=\begin{cases} 0 & \text{for } x_\varepsilon \le \pm\delta/2 \\ \dfrac{1}{\delta} & \text{for } x_\varepsilon > \pm\delta/2 \end{cases} \tag{6.8}$$

Under such a situation Eq.(6.5) is given by

$$E\{C(x,\hat{x})\,|\,y\}=\int_{-\infty}^{\infty} C(x,\hat{x})\,p(x|y)dx=\frac{1}{\delta}\left[1-\int_{\hat{x}-\delta/2}^{\hat{x}+\delta/2} p(x|y)dx\right]=1/\delta-p(\hat{x}|y). \tag{6.9}$$

This equation shows that for any fixed value of δ the minimum of Eq.(6.6) can be obtained by maximizing $p(x\,|\,y)$. Such estimator is well-known as MAP estimator and is given by

$$\hat{x} \Leftarrow \max_{\hat{x}} p(x\,|\,y)=\max_{\hat{x}} p(y/x)p(x)/p(y). \tag{6.10}$$

Thus the MAP estimator is given as the mode of the posterior density which is modified prior PDF in accordance with the observed data.

6.3.1. MAP Estimation Under GGD Prior

As said before, MAP estimation uses some prior knowledge about the quantity to be estimated and updates that prior knowledge with the likelihood function that contains information available in the new data. As a prior knowledge, a prior PDF, based on previous knowledge about the event, is taken which is

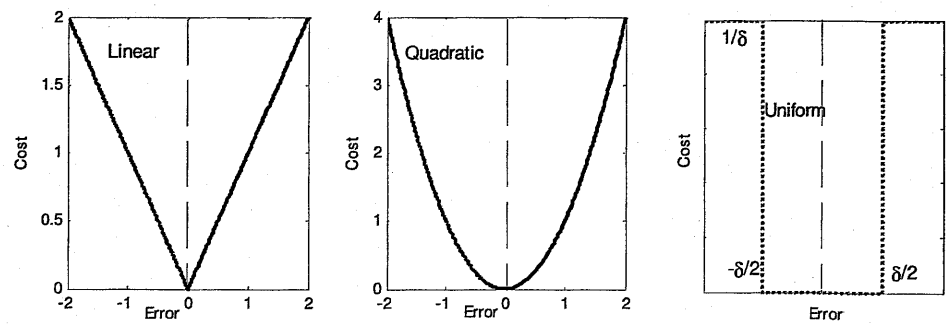


Figure 6.3 Different types of cost function used in Bayesian estimation.

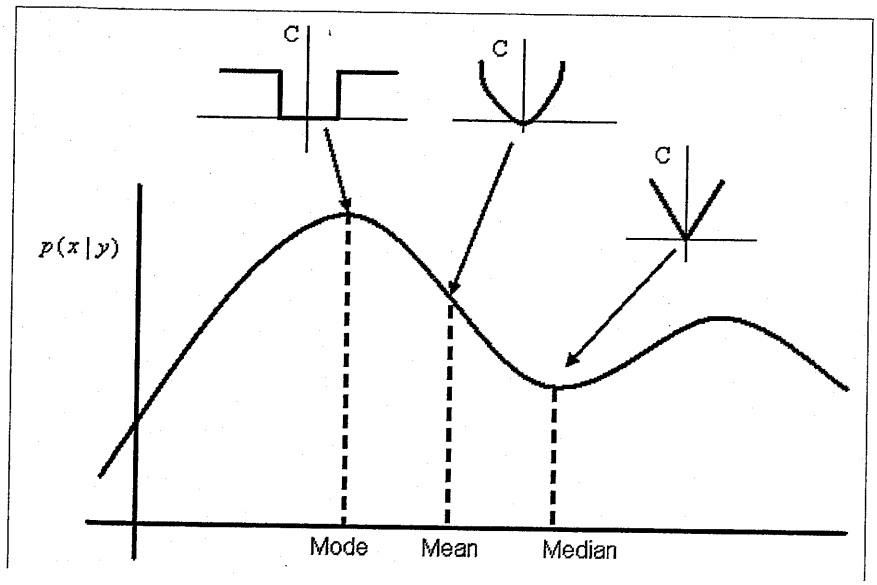


Figure 6.4 Generalized posterior PDF showing different Bayesian estimators under linear, quadratic and uniform cost functions.

further modified, according to Bayes theorem, by the likelihood for the new samples to form a new PDF called as the Bayesian posterior. The improved posterior PDF contains all known information, both old and new about the event. The maximum of posterior PDF under the uniform cost function gives MAP estimator that is also an optimal estimator [101]. MAP estimation for denoising ICs is similar to MAP estimator of spectral components of the clean speech from that of the observed noisy speech. The problem of estimation of spectral components in any frequency bin can either be formulated as the task of estimating real part and imaginary part or estimation of spectral magnitude and related phase. Here we will obtain joint estimator for magnitude and phase of the time-frequency series of speech.

Let in the *kth* frequency bin $Y_k = R_k e^{i\nu_k}$ represents noisy signal and $X_k = a_k e^{i\alpha_k}$ represents spectral components of clean signal in the polar form. Thus the problem of estimating clean signal can be formulated in terms of estimation of magnitude a_k and phase α_k. Accordingly, MAP estimator of a_k and phase α_k are given as the mode of the posterior PDF $p(a_k, \alpha_k \mid Y_k)$, which can be

obtained by maximizing the posterior PDF $p(a_k,\alpha_k \mid Y_k)$. The desired posterior PDF is given by Bayes' theorem in terms of likelihood function $p(Y_k \mid a_k,\alpha_k)$ and a prior PDF $p(a_k,\alpha_k)$ as follows

$$p(a_k,\alpha_k \mid Y_k) = p(X_k \mid Y_k) = \frac{p(Y_k \mid a_k,\alpha_k)p(a_k,\alpha_k)}{p(Y_k)}. \tag{6.11}$$

Since $p(Y_k)$ is constant with respect to (w.r.t) spectral magnitude a_k and phase α_k, only numerator of Eq.(6.11) is significant in the optimization landscape and denominator will be dropped hereafter. The natural logarithmic function of only numerator is optimized, which is given by

$$J = \ln[p(Y_k \mid a_k,\alpha_k)p(a_k,\alpha_k)]. \tag{6.12}$$

The MAP estimators of magnitude a_k and phase α_k are given by

$$(\hat{a}_k,\hat{\alpha}_k) = \underset{(a_k,\alpha_k)}{\arg.\max}\{J\} = \underset{(a_k,\alpha_k)}{\arg.\max}[\ln\{p(Y_k \mid a_k,\alpha_k)p(a_k,\alpha_k)\}]. \tag{6.13}$$

Obviously, MAP estimation needs knowledge of conditional probability $p(Y_k \mid a_k,\alpha_k)$ and prior probability $p(a_k,\alpha_k)$ of the spectral components of clean speech for which GGD will be used here. The GGD model for the magnitude of the DFT coefficients of the clean speech signal in the kth frequency bin, is given by

$$p(a_k) = \begin{cases} A_x e^{-\left[\frac{|a_k|}{b_x}\right]^{\beta_x}}, & for\ 0 \le a_k \le \infty \\ 0 \qquad\qquad else \qquad\qquad , \end{cases} \tag{6.14}$$

where b_x is the scale parameter, β_x is shape parameter,

$$A_x = \frac{\beta_x}{2b_x\Gamma(1/\beta_x)} = \frac{\beta_x}{2\Gamma(1/\beta_x)}\frac{1}{\sigma_x}\sqrt{\frac{\Gamma(3/\beta_x)}{\Gamma(1/\beta_x)}}; \tag{6.15}$$

$$\sigma_x = \text{Stdv.of clean speech} = \sqrt{E\{a_k^2\}}\ . \tag{6.16}$$

Since the positions of analysis window in STFT analysis are arbitrary, the PDF of phase α_k, follows uniform distribution and is expressed as

$$p(\alpha_k) = \textit{Uniform}\ \text{PDF} = \begin{cases} \dfrac{1}{2\pi} & \text{for } -\pi \le \alpha_k \le \pi \\ 0 & \text{else} \end{cases}. \tag{6.17}$$

The joint PDF of the magnitude a_k and phase α_k is given as

$$p(a_k, \alpha_k) = \frac{A_x}{2\pi} e^{-\left[\frac{|a_k|}{b_x}\right]^{\beta_x}}, \quad \textit{for } 0 \le a_k \le \infty \text{ and } -\pi \le \alpha_k \le \pi. \tag{6.18}$$

The conditional probability $p(Y_k \mid X_k)$ of the observed data, given clean signal inherits randomness of noise and can be given as the PDF of noise, which is also modeled by GGD as follows

$$\begin{aligned} p(Y_k \mid X_k) &= p(Y_k \mid a_k, \alpha_k) = \textit{Noise}\ \text{PDF} \\ &= (A_N/2\pi) e^{-\left[\frac{|Y_k - X_k|}{b_n}\right]^{\beta_n}}, \\ \textit{for } 0 \le |Y_k - X_k| &\le \infty, -\pi \le \gamma_k \le \pi \end{aligned} \tag{6.19}$$

where b_n and β_n are scale and shape parameters, respectively, for the GGD distribution for noise spectral component in the frequency bin k, and γ_k is the

corresponding phase for noise.

$$A_N = \frac{\beta_n}{2b_n\Gamma(1/\beta_n)} = \frac{\beta_n}{2\Gamma(1/\beta_n)}\frac{1}{\sigma_n}\sqrt{\frac{\Gamma(3/\beta_n)}{\Gamma(1/\beta_n)}}, \tag{6.20}$$

and

$$\sigma_n = \text{Stdv.of noise} = \sqrt{E\{|D_k^2|\}}\ . \tag{6.21}$$

Now the desired posterior density in Eq.(6.11) can be given by using Eq.(6.18) and Eq.(6.19) and dropping denominator as follows

$$p(a_k,\alpha_k \mid Y_k) \propto p(Y_k \mid a_k,\alpha_k)p(a_k,\alpha_k) = \left(A_N A_x/4\pi^2\right)e^{-\left[\frac{|Y_k - X_k|^{\beta_n}}{b_n^{\beta_n}}\right]-\left[\frac{|a_k|^{\beta_x}}{b_x^{\beta_x}}\right]}. \tag{6.22}$$

Using Eq.(6.22) in Eq.(6.12) gives

$$J = -\frac{|R_k e^{iv_k} - a_k e^{i\alpha_k}|^{\beta_n}}{b_n^{\beta_n}} - \frac{|a_k|^{\beta_x}}{b_x^{\beta_x}} + \ln\frac{A_N A_X}{4\pi^2}. \tag{6.23}$$

Now, in order to locate, say at $\hat{\alpha}_k$ and $\hat{a}_k$, the highest of the posterior PDF, differentiating Eq.(6.23) w.r.t. phase α_k and spectral amplitude a_k, and equating derivatives to zero gives

$$\partial J/\partial\alpha_k\big|_{\alpha_k=\hat{\alpha}_k} = B[R_k^2 + a_k^2 - 2R_k a_k\cos(v_k - \hat{\alpha}_k)]^{\beta}$$
$$[2R_k a_k \sin(v_k - \hat{\alpha}_k)],$$

Equating it with zero gives

$$\sin(v_k - \hat{\alpha}_k) = 0 \Rightarrow \hat{\alpha}_k = v_k, \tag{6.24}$$

where $B=0.5\beta_n/b_n^{\beta_n}$; $\beta=0.5\beta_n-1$. Eq.(6.24) gives MAP estimated phase of the spectral components of clean signal which is same as that of that of the spectral components of the noisy speech. Similar treatment of Eq.(6.23) w.r.t spectral amplitude a_k along with use of Eq.(6.24) gives

$$\partial J/\partial a_k \mid_{a_k=\hat{a}_k}=0$$

which further gives

$$2B[R_k^2+\hat{a}_k^2-2R_k a_k]^{\beta}[-R_k+\hat{a}_k]-\beta_x \hat{a}_k^{\beta_x-1} b_x^{-\beta_x}[sign(\hat{a}_k)]^{\beta_x}=0, \qquad (6.25)$$

In order to avoid singularity when $0<\beta_x<1$, and $a_k=0$, $\hat{a}_k^{\beta_x-1}$ in Eq.(6.25) is replaced by $\hat{a}_k^{\beta_x-1}+\delta$, where δ is very small ($<10^{-4}$) number. Further simplification of Eq.(6.25), results in the following radical (power) equation

$$\beta_x \hat{a}_k^{\beta_x-1} b_x^{-\beta_x} sign(\hat{a}_k)^{\beta_x}=2B(R_k-\hat{a}_k)^{2\beta+1}$$
$$\Rightarrow \hat{a}_k^{\beta_x-1}=P(R_k-\hat{a}_k)^{\beta_n-1}, \qquad (6.26)$$

where $P=b_x^{\beta_x}\beta_n/b_n^{\beta_n}\beta_x$. It may be very difficult to find an analytical solution of the Eq.(6.26), however, its numerical solution can be easily obtained by Newton-Rapshon's method under which numerical solution after the ith iteration is given as

$${}^{i+1}\hat{a}_k={}^{i}\hat{a}_k-\frac{{}^{i}\hat{a}_k^{\beta_x-1}-P(R_k-{}^{i}\hat{a}_k)^{\beta_n-1}}{(\beta_x-1){}^{i}\hat{a}_k^{\beta_x-2}+P(\beta_n-1)(R_k-{}^{i}\hat{a}_k)^{\beta_n-2}}. \qquad (6.27)$$

This solution gives MAP estimator of the spectral magnitude which is further combined with the phase of a related noisy spectral component to get a spectral component of the clean signal. The solution in Eq.(6.27) is sensitive to the used initial value. The good initial values can be obtained as the special case solutions of the Eq.(6.26) as described below

6.3.2. Special Cases of GGD based MAP Estimator

Case-1 For $\beta_x = \beta_n = 2$, the spectral components of both the noise and speech signal have Gaussian (assumption working under the conventional Wiener filtering) PDF and solution of the Eq.(6.26) using Eq.(4.11) is given by

$$\hat{a}_k = \frac{b_x^2}{b_x^2 + b_n^2} R_k = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_n^2} R_k, \tag{6.28}$$

which is Wiener filter and can be used as the initial value for the iterative solution in Eq.(6.27). Thus the MAP estimate under the Gaussian model for both the noise and speech is equivalent to Wiener filtering. This is due to symmetry of the posterior PDF, which too is Gaussian, for which mean (Wiener solution) and mode (MAP estimator) are equal.

Case 2. When $\beta_x = 1$ i.e. the clean speech spectral component has Laplacian PDF and that of noise is GGD, the solution to Eq.(6.26) is given by

$$\hat{a}_k = R_k - \left(\frac{1}{P}\right)^{\frac{1}{\beta_n - 1}} = R_k - \left(\frac{1.4142\sigma^{\beta_n}{}_n}{\sigma_x \beta_n}\right)^{\frac{1}{\beta_n - 1}} \left[\frac{\Gamma\left(1/\beta_n\right)}{\Gamma\left(3/\beta_n\right)}\right]^{\frac{0.5\beta_n}{\beta_n - 1}} \tag{6.29}$$

in which further if the PDF of noise spectral components is assumed to be Gaussian, we have $\beta_n = 2$ and Eq.(6.29) can be simplified into

$$\hat{a}_k = R_k - 1.4142\frac{\sigma_n^2}{\sigma_x} = R_k - 1.4142\frac{\sigma_x}{\xi} \tag{6.30}$$

where $\xi = \sigma_x^2/\sigma_n^2$ is the spectral SNR of the noisy speech signal.

For the other two special cases i.e. when $\beta_x = \beta_n = 1$, Eq.(6.26)) fails to give solution for a_k and it can be shown that under such condition it leads to $\sigma_x = \sigma_y$, and for $\beta_n = 1$ estimate for a_k is given by

$$a_k = (P)^{\frac{1}{\beta_x - 1}} = \left[\left(\frac{b_x^{\beta_x}}{b_n} \right) \left(\frac{1}{\beta_x} \right) \right]^{\frac{1}{\beta_x - 1}}, \qquad (6.31)$$

which is independent of R_k. However, such all-special cases are not happening with the speech signal. As shown in [82] the shape parameter of the spectral magnitude is nearly equal to $1 (0.8 < \beta_x < 1)$ and in the majority of the frequency bins spectral amplitudes have strongly Laplacian distribution (GGD with $\beta_x < 1$). The simultaneous happenings of $\beta_x = \beta_n = 1$ can be avoided by making them slightly more or less than 1.

6.4. Voice activity detection

The solutions of Eq.(6.26), require scale and shape parameters of clean speech and noise signals. The estimation of these parameters for general problem of denoising speech signal under additive background noise and ICs enhancement will be a little bit different. In the speech enhancement problem clean signal and noise signals are not known. However, they can be estimated from the noisy data only. The GGD parameters of noise can be estimated, using ML approach as described before, from the noise only portion e.g. a few samples from the beginning or other silent parts of the noisy data can be taken using voice activity detector. Voice activity detection in low SNR condition is problematic. A VAD for this purpose based on negentropy measure of speech signal is described below.

Detection of noise only frames and noisy speech frames is difficult, especially, in a very low SNR condition. In the very low SNR condition conventional energy based VAD detector fails [102]. It is also conceivable from Figure 6.5, which shows how the energies of speech segments under different SNR conditions change. We propose here a statistical VAD detector based on the chaos measure of the spectral magnitude of quasi-stationary segments. For a speech signal, spectral components are well organized, however, for the noise signal it is not well organized e.g. spectrogram of White Gaussian Noise (WGN) and clean

speech signal from the male speaker can be observed in Figure 6.6, to be imbued with such differences in the spectral organization. Accordingly, the observed noisy speech data during the speech period in the signal is less chaotic, as shown in Figure 6.7, than during the noise-only frames and thus chaos-based measure can discriminate noise-only and noisy speech frames. For doing voice activity detection based on the measure of such chaotic characteristic, we have used negentropy as a measure [19]. The benefit of using negentropy over others such as entropy [102][103] is that it is always positive and can be computed in terms of only shape parameters of the used GGD model. The negentropy of each frame in DFT domain is obtained in terms of Differential Entropy (DE) ΔH of the magnitude of spectral components. The DE of the any frame data $U=[Y_1,Y_2.....Y_N]$ is given by

$$\Delta H(U)=-\int_{-\infty}^{\infty} p(U)\log p(U)dU, \tag{6.32}$$

where $p(U)$ represents PDF of the frame data U. The PDF of magnitude of spectral components of each frame is represented by GGD with mean $\mu_u=0$

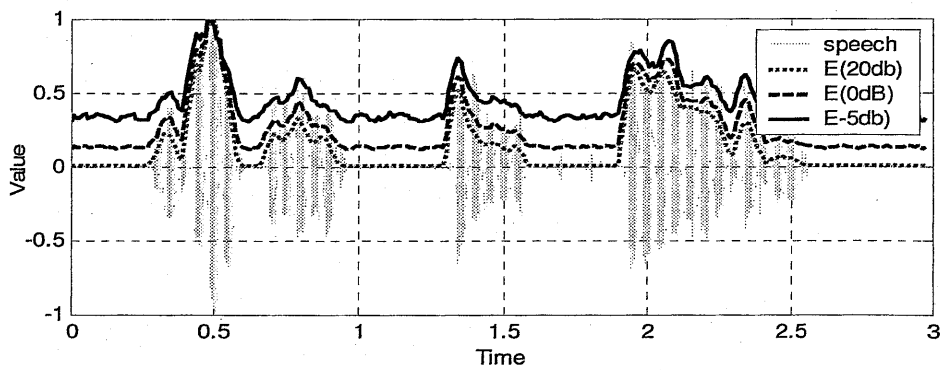


Figure 6.5 Energy of speech segments corrupted by WGN under different SNR conditions.

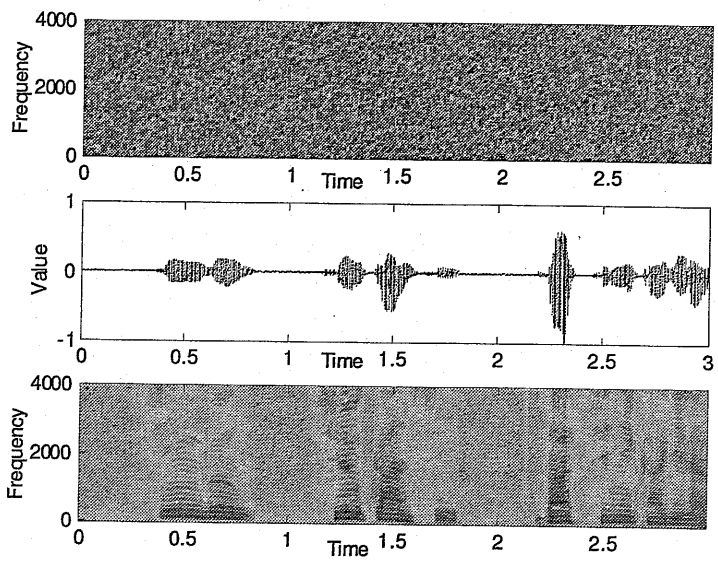


Figure 6.6 Spectrograms of WGN (upper) and clean speech from male speaker (lower and middle figures).

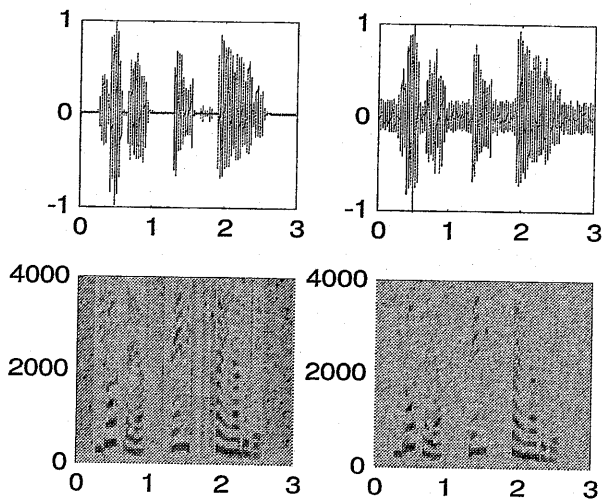


Figure 6.7 Spectrograms clean speech and noisy speech degraded by WGN (First row clean and noisy waveforms, Second row corresponding spectrograms).

scale parameter α_u and shape parameter β_u, estimated from the data. Using GGD model for $p(U)$ in Eq.(6.32) it can be integrated to give

$$\Delta H(U) = f(\alpha_u, \beta_u) = \log\left[\frac{2\alpha\,\Gamma(1/\beta_u)}{\beta_u}\right] + \frac{1}{\beta_u}, \tag{6.33}$$

which depends on the scale and shape parameters. The negentropy $H(\beta_u)$ is computed as the difference of DE of Gaussian RV, with same variance as of that of spectral components of speech, and DE of speech spectral components modeled by the GGD $f_{GG}(0, \alpha_u, \beta_u)$. Accordingly, negentropy $H(\beta_u)$ is given by

$$\begin{aligned} H(\beta_u) &= \Delta H(\alpha_g, \beta_g = 2) - \Delta H(\alpha_u, \beta_u) \\ &\approx \log\left[\frac{\beta_u}{2}\sqrt{\frac{6.29\Gamma\left(3/\beta_u\right)}{\Gamma\left(1/\beta_u\right)^3}}\right] + \left(0.5 - \frac{1}{\beta_u}\right). \end{aligned} \tag{6.34}$$

The theoretical variation of negentropy of GGD with shape parameter is shown in the Figure 6.8. It is obvious from there that the negentropy is zero for the Gaussian distribution and goes up in the positive direction for the spiky distribution. Since the speech frames are more parsimonious than noise frames, the noise-only frames will have lower negentropy while for the noisy speech frames negentropy will be relatively high and thus a threshold value of the negentropy can be chosen to demark noisy speech frames and noise-only frames. The threshold value of negentropy can be decided on the basis of the global statistics of the negentropy. The negentropy of the frames itself is a random variable and its PDF represents joint probability of occurrence of noisy speech frame and noise-only frames. The PDF of negentropy of each frame can also be modeled by GGD with mean μ_h, scale parameter α_h and shape parameter β_h as follows

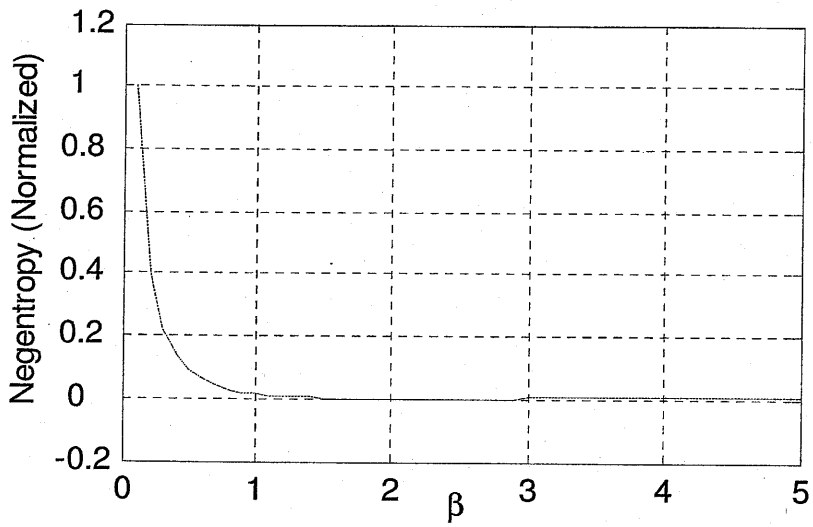


Figure 6.8 Shape parameter versus negentropy of the GGD. It is zero for Gaussian distribution and positive for the spiky distribution ($0<\beta<1$).

$$p_v(H) = \frac{\beta_h}{2\alpha_h \Gamma\left(1/\beta_h\right)} e^{-\left[|H-\mu_h|/\alpha_h\right]^{\beta_h}} \tag{6.35}$$

where $v = [speech,\ noise]$ to represent noisy speech frame and noise-only frames respectively. Since the occurrence of noise and speech frame is independent

$$p_v(H) = p(noise \mid H)\, p(speech \mid H) \tag{6.36}$$

The threshold value H_{TH} of the negentropy is estimated under assumption that the conditional probabilities of noise-only frame and noisy speech frames are same

(say p_1) at the threshold [104]. Accordingly, the threshold H_{TH} is given by

$$H_{TH} = \mu_h \pm \left[\frac{-\alpha_h \log\left(2p_1(1-p_1)\Gamma\left(1/\beta_h\right)\right)}{\beta_h} \right]^{1/\beta_h}. \tag{6.37}$$

The similarity between the spectral bands of the estimated noise and original noise can be measured by measuring the Kullback Leibler Divergence between the PDF of their spectral bands.

6.5. GGD parameters for noise and speech

Using the threshold value in Eq.(6.37), total noise-only frames (say L) are stacked together in the time succession and GGD parameters for noise spectral components in each frequency bins are estimated from these data using ML technique as described in Chapter 3. Due to unavailability of clean signal, the GGD parameters for the clean speech cannot be obtained directly as has been done for the noise spectral components; however, they can be estimated using GGD parameters of noise spectral components and higher order statistics of the spectral components of the observed noisy data. The shape parameter of the spectral magnitude of clean speech can be estimated from their kurtosis K_x using the following relation valid for the GGD function

$$K_X = \frac{\Gamma(1/\beta_x)\Gamma(5/\beta_x)}{\Gamma(3/\beta_x)^2} = V(\beta_x). \tag{6.38}$$

where V is some function. The shape parameter can be estimated by inverting the relation in Eq.(6.38) such that

$$\hat{\beta}_x = V^{-1}(K_x). \tag{6.39}$$

It is difficult to find an analytical inverse function for V, however, it can be easily done with the look-up table by storing values of shape parameter and corresponding value of kurtosis. The theoretical variation in the kurtosis with β is shown in Figure 6.9. The estimation of shape parameter using Eq.(6.39) needs kurtosis of the clean signal which is not available, however, it can be estimated from the higher order statistics of spectral components of the noisy speech and estimated GGD parameters for noise. Starting from Eq.(6.2) it can be shown that the kurtosis of clean signal is related to kurtosis, skewness, variances and means of noise and noisy data as follows

$$K_x = \frac{\left[K_y\sigma_y^4 - 4S_x\sigma_x^3\mu_n - 4S_n\sigma_n^3\mu_x - 6\sigma_y^2\sigma_n^2 - (K_n - 6)\sigma_n^4\right]}{(\sigma_y^2 - \sigma_n^2)^2}. \tag{6.40}$$

where K_z =Kurtosis of spectral components of signal z, S_z =coefficient of skewness of spectral components of signal z, , σ_z and μ_z denotes standard deviation and mean of the signal indicated by subscript z, and $z=(x,y,n)=$(noisy speech, clean speech, noise) signal. The coefficient of skewness of the clean speech signal is estimated from the skewness and lower order statistics of the noisy data and noise signal as follows

$$S_x = Skewness = \frac{\left[S_y\sigma_y^3 - 3(\sigma_y^2 - \sigma_n^2)\mu_n - 3(\mu_y - \mu_n)\sigma_n^2 - S_n\sigma_n^3\right]}{\left(\sigma_y^2 - \sigma_n^2\right)^{3/2}} \tag{6.41}$$

The variance and means of the clean speech signal are estimated as follows

$$\hat{\sigma}_x^2 = \sigma_y^2 - \sigma_n^2; \tag{6.42}$$

$$\mu_y = \mu_x + \mu_n \Rightarrow \mu_x = \mu_y - \mu_n. \tag{6.43}$$

However, to prohibit σ becoming negative, it is approximated as

$$\hat{\sigma}_x^2 = \max\left(\sigma_y^2 - \sigma_n^2, 0\right), \tag{6.44}$$

in which the subscripts n denotes noise and y denotes noisy speech signal. The scale parameter of the GGD for clean speech is then obtained using value of σ_x and β_x in Eq.(4.11). The whole process of the speech enhancement, as described and derived above, in the DFT domain under the proposed framework, is shown in Figure 6.10.

Performance Evaluation Score

In the above described MAP estimation of the clean signal, the estimated parameters for the noise and clean signal plays important role and the accuracy of the estimation can be checked by measuring the distance between the PDFs of the original spectrum and estimated

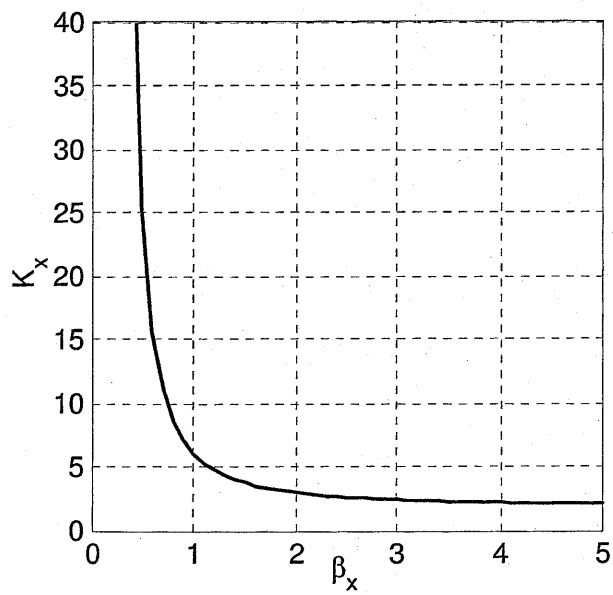


Figure 6.9 Theoretical variation of kurtosis of GGD with shape parameter.

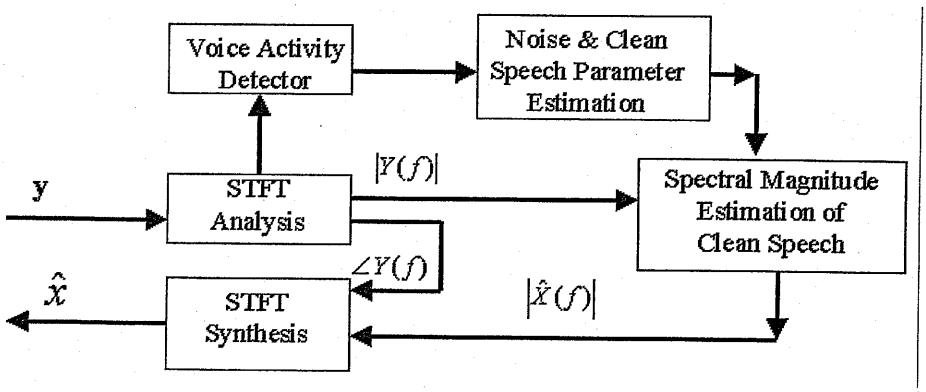


Figure 6.10 Speech enhancement scheme used to estimate spectral components of clean speech in the DFT domain. The phase of noisy data is used to reconstruct the original signal.

spectrums of the related signals. The similarity between the spectral bands of the estimated noise and original noise can be measured by measuring the Kullback Leibler Divergence (KLD) between the PDF of their spectral bands. Since the PDF of the spectral bands are modeled by the GGD, KLD between them can be measured in terms of GGD parameters. The KLD between two GGD functions defined by scale parameters α_i, α_j and shape parameters β_i, β_j is given by

$$D_{ij} = \log\left(\frac{\beta_i \alpha_j \Gamma(1/\beta_j)}{\beta_j \alpha_i \Gamma(1/\beta_i)}\right) + \left(\frac{\alpha_i}{\alpha_j}\right)^{\beta_j} \frac{\Gamma((\beta_j+1)/\beta_i)}{\Gamma(1/\beta_i)} - \frac{1}{\beta_j}. \qquad (6.45)$$

Further DFT coefficients in each frequency bin are assumed to be independent, so the overall distance between two noise spectrums can be given by averaging the distance calculated in Eq.(6.45) for each pair of the spectral bands of the original(*Org.*) noise and estimated (*Est.*) noise as follows

$$D_{avg.} = \frac{1}{N}\sum_{k}^{N} D_{ij}(f_k),\ (\ i = Org.,\ j = Est.). \tag{6.46}$$

The GGD parameter estimation for ICs enhancement is same as above except the noise parameters were estimated from the TFSS of noise signal scaled from a IC using Eq.(6.3). Then the clean signal parameters were estimated as mentioned above.

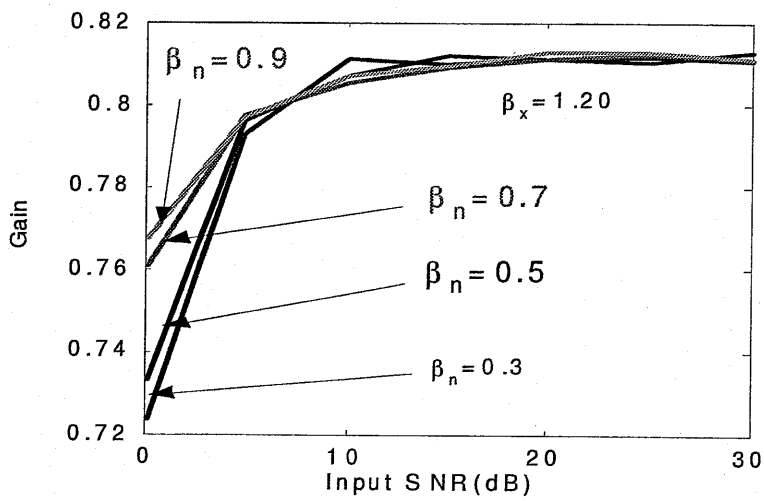
In order to evaluate performance of the proposed denoising algorithm global SNR and segmental SNR of the estimated signal will be measured [88]. The global SNR provides error measurement over time and frequency and is defined as

$$SNR_{dB} = 10\log_{10}\frac{\sum x^2(t)}{\sum[x(t)-\hat{x}(t)]^2}, \tag{6.47}$$

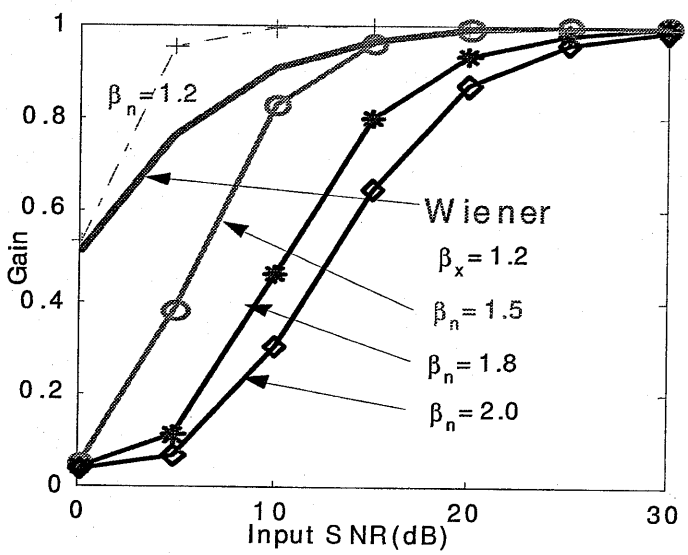
In order to evaluate performance of the FDICA with MAP enhancement the NRR defined in Eq.(3.46) with and without MAP enhancement will be used. As a subjective test preference test for enhanced speech signal has been done.

6.6. Experimental Evidences

The experiments in this chapter are placed in three separate parts. First we place characteristics of noise suppression rule derived in Eq.(6.27) for the MAP estimator. Then speech enhancement experiment under different noise conditions will be presented. Finally, enhancement experiments for separated ICs will be placed. The characteristics of noise suppression rule are shown in the Figure 6.11. These gain curves were obtained for the 5000 samples of random variables (RV) generated for given GGD parameters. The β parameters of GGD for RV corresponding to clean speech were held constant at 1.2. It was done so, as the average value β for $|x(f)|$, speech amplitude was found around 1, but at exactly 1 Eq.(6.26) vanishes. Obviously, the shape of the gain function depends on the GGD parameters of the clean speech and noise signal. The proposed noise suppression rule offers more noise suppression at the lower



(a)



(b)

Figure 6.11 a) Characterestic of noise suppression rule for different values of GGD parameters. The shown curves were obtained from artificially generated random variables with GGD parameters. (a) shows plots for highly spiky noise (b) shows plots for less spiky noise. Shape of the curve changes with the characteristic of noise.

SNR and the amount of noise suppression decreases as the signal's SNR level goes up which means the clean signals are not cleaned further. In the experiments for the speech enhancement we have used four sentences, of time length 3 sec., sampling frequency 8 kHz and spoken by two male and two female speakers, from the ASJ continuous speech corpus for the research[80] and noise data from the NOISEX-92 database freely available at http://mi.eng.cam.ac.uk/comp.speech/Section1/Data/noisex.html. The clapping noise was self-recorded. The enhancement experiments has been done with the speech signals degraded to different SNR levels e.g. –5 db, 0 db, 5 db, 10 db, 15 db, and 20 db. In the first part of the experiment, statistical characteristics of the spectral components of different noise were investigated. The signal analysis conditions are kept same as mentioned in Table.3.2. The GGD parameters for the spectral components of WGN, babble (BAB) noise, and clapping noise were estimated using ML approach. It is the shape parameter that decides shapes of the PDF, the value of shape parameters for them are shown in Figure 6.12. Upper subplot in that figure shows shape parameter β for the magnitude of spectral components of the WGN, babble noise, and clapping noise. The bar plot in lower subplot of the same figure shows values of shape parameters, averaged over the total number of frequency bins, for the real part, imaginary part and the magnitude of the spectral components of the same noise. It is evident from these figures that the spectral components for WGN have Gaussian distribution but the babble noise and clapping noise have relatively spiky distribution. The PDF of the time domain samples of all these three noise signals are already shown in the Figure 6.2 with fittings of Laplacian, Gaussian and GGD functions in which too, similar differences in PDF can be observed. Thus it is very loose assumption to use Gaussian model for all noise signals, as is done in the Wiener filtering and can affect performance of the related speech enhancement algorithms. In Figure 6.13, PDFs of the spectral components of the WGN, clapping and babble noise in frequency bin f=688 Hz are shown. It is evident from there that GGD with measured parameters provides better fitting in the PDF of the spectral components. Similar results were observed for other frequency bins too. In

Figure 6.14 the performance of energy based VAD and negentropy based VAD has been shown under very low SNR condition. It can be seen in that figure how the energy based method fails to demark speech segments and noise-only segments. The result of negentropy based VAD for the speech signal, from male speaker, degraded to 0db SNR level by WGN and clapping noise are shown in Figure 6.15. In that figures, patterns in spectrograms of the noisy speech data revel chaos of noisy speech and noise-only frames can be observed. The upscaled negentropy curve is also plotted to show how it tracks signal frames with different chaotic conditions. The negentropy of each frame is also plotted over noisy and clean speech waveforms. In the case of lower SNR or higher SNR almost similar results were found. For the discrimination of the noisy speech signal into noisy speech frames and noise-only frames, a threshold is required which can be estimated using Eq.(6.37) and is shown in Figure 6.16. That figure shows theoretical value of threshold as a function of the probability of occurrence of a speech segment. The shown threshold curves were estimated for the speech signal degraded to 0 dB and 15 dB SNR levels by WGN and clapping noise. It is evident that there is a little variation in the negentropy value where the probability of occurrence of each is assumed to be equal (=0.5). Also, if the probability of occurrence of speech frame is increased, threshold goes down and chances of taking larger number of frames as noisy speech and less number of frames as a noise-only frame increase. In our experiments, the used value of thresholds for WGN, babble and clapping noise were 0.1, 0.2 and 0.19 respectively. After discriminating the noisy speech signal into noisy speech frames and noise-only frames, the GGD parameters for noise were estimated. The estimated noise parameters and statistics of the noisy speech signal were used to estimate GGD parameters for the clean signal using Eq.(6.40) -Eq.(6.43). The estimated parameters for the babble noise and clean speech signal along with the corresponding parameters for their original versions are shown in Figure 6.17. As it is evident from that figure that the parameter estimation for the clean signal is not so much accurate, however, the estimated

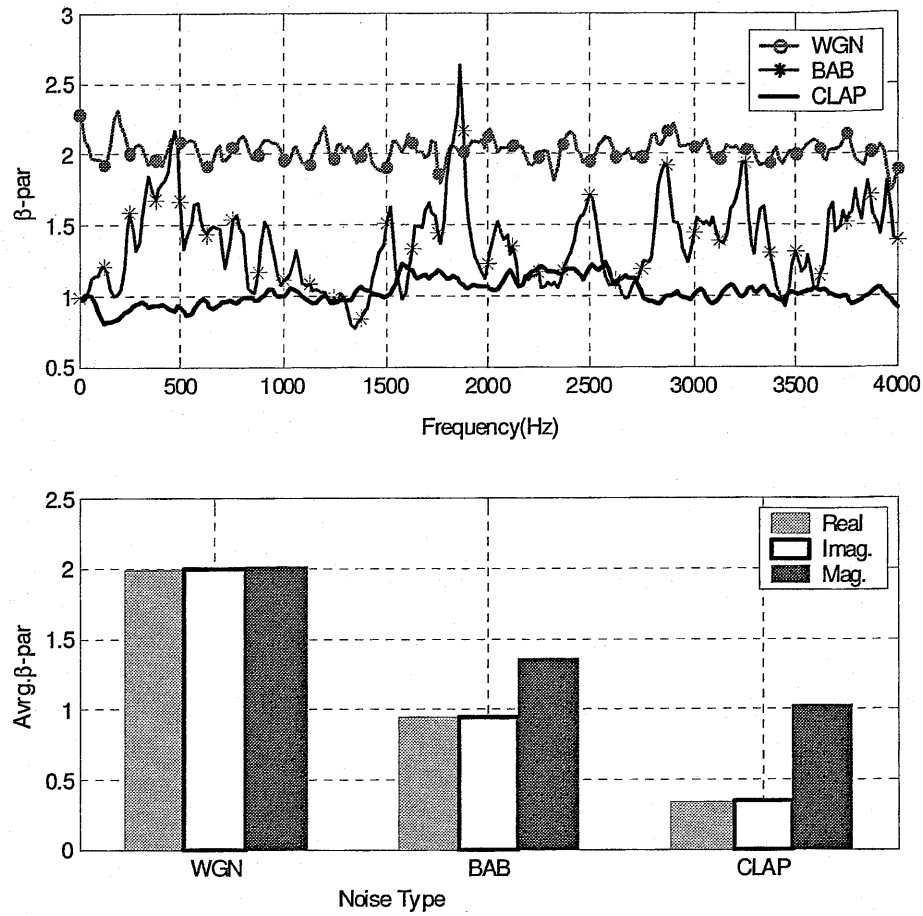


Figure 6.12 Shape parameters of the WGN, babble and clapping noise. Upper figure shows shape parameters of magnitude of noise spectral components and lower bar plot shows shape parameters, averaged over frequency bin, for real part, imaginary part and polar magnitude of the WGN, babble and clapping noise.

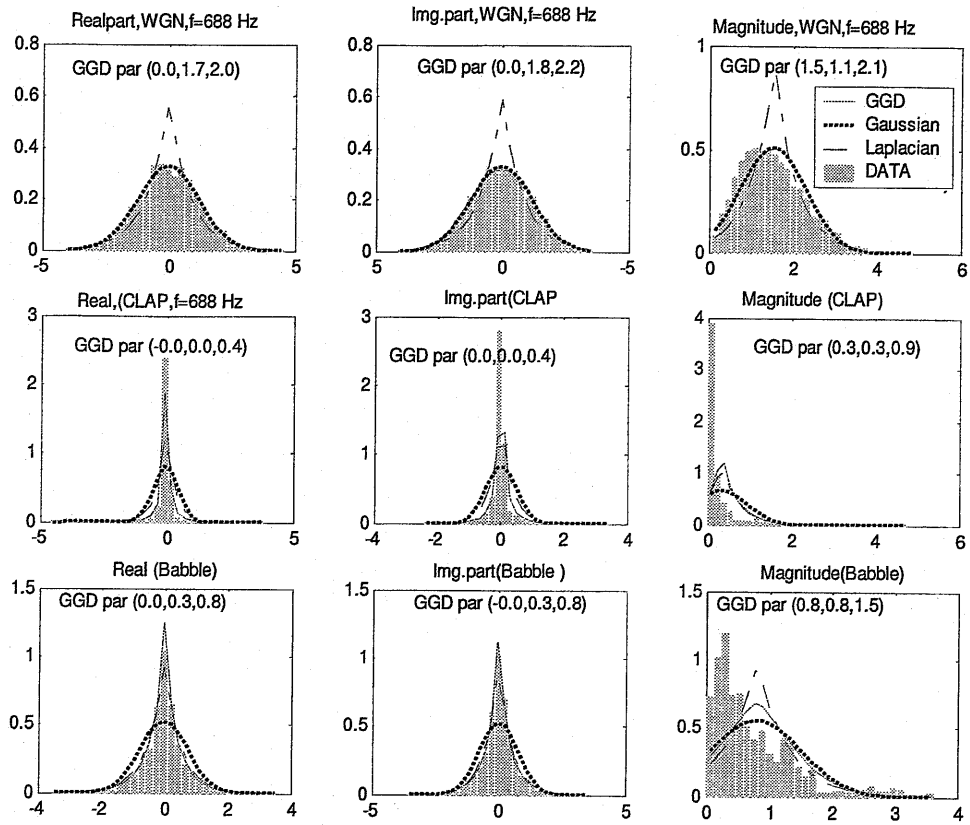


Figure 6.13 Fitting of GGD, Gaussian and Laplacian PDF in the histograms of magnitude of noise spectral components in frequency band f=688 Hz. Figures in the first column (left) of every row is for imaginary part, middle column is for the real part and rightmost column is for the polar magnitude. Each successive row from top to bottom is for WGN, clapping and babble noise respectively. The GGD parameters shown in each figure represent (mean, scale, shape). (Legend indication is same for each plot which has been shown in one plot for clarity)

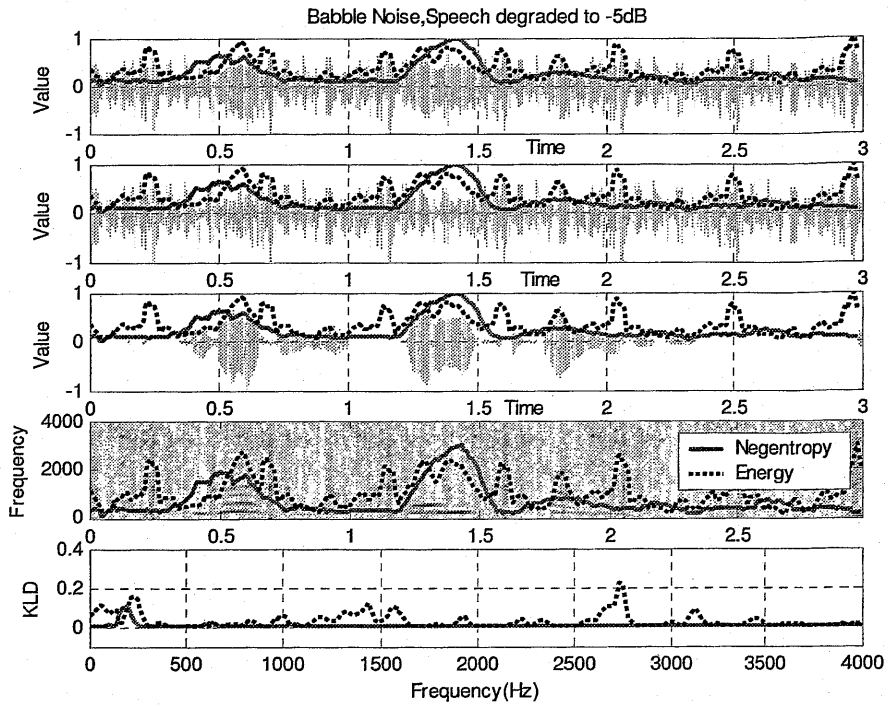


Figure 6.14 Performance of energy based VAD and negentropy based VAD. The clean speech signal is corrupted by speech like babble noise to SNR level of -5dB. Subplots from top to bottom are noisy speech signal, noise signal, clean speech signal, Spectrogram of noised speech and KLD between estimated and original noise spectrum.

parameters for the noise are very near to that of the original noise signals. Next we performed denoising experiments to estimate clean speech spectral components using Eq.(6.27). The four clean speech signals, two from male and two from female speakers, were noised to the SNR levels of –5 db, 0 db, 5 db, 10 db, 15 db, and 20 db by WGN, babble noise and clapping noise. we used Eq.(6.29) to initialize the iterative process of the Eq.(6.27). Use of this initial value looks logically better than that of in Eq.(6.28) in the light of PDF of the speech spectral components. However, comparative study on the appropriateness of these two initial values and their influence on the overall performance is still

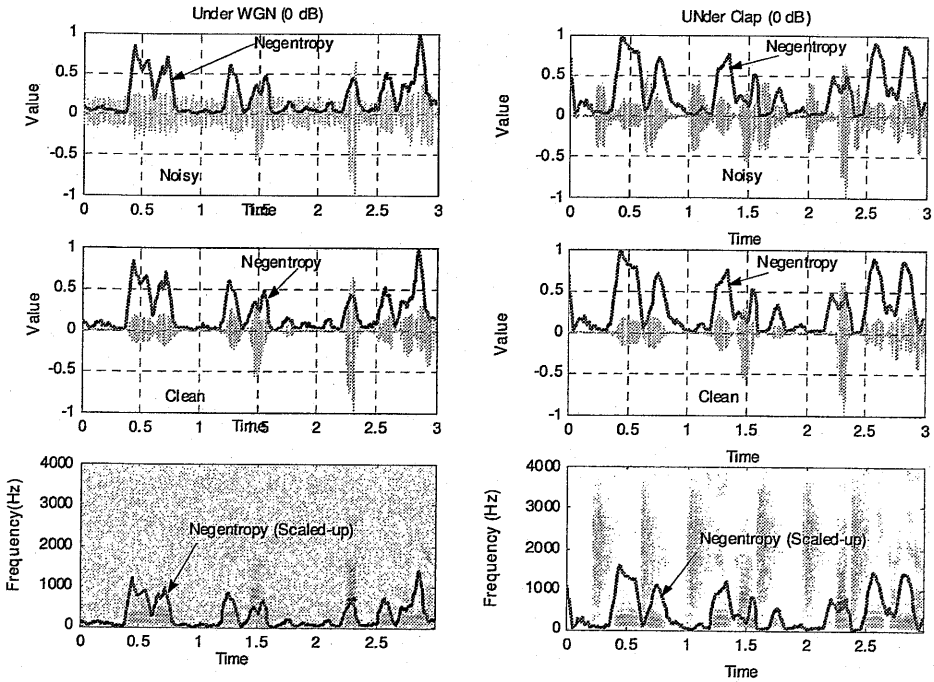


Figure 6.15 Performance of negentropy based VAD under WGN (Left column) and Clap noise (right column). SNR=0 dB. Speech signal is degraded to 0db, very low SNR condition. Negentropy of each frame is plotted over the degraded speech as well as clean speech waveform and spectrogram of the degraded speech signal. Negentropy values plotted over the spectrogram are upscaled to fit into the plot

unexplored. As a performance measure SNR level and segmental SNR were measured in accordance with Eq(6.47) for the enhanced speech signal. The SNR levels of the degraded speech and denoised speech, averaged for all the four speech signals, are depicted in the figures of Figure 6.18. The same figures also contain SNR improvement result obtained using Eq.(6.28) which is a Wiener filter. The performance of the proposed MAP estimator in the lower SNR conditions is better than that of in the higher SNR conditions which is indicative of the fact that algorithm provides stronger noise suppression in the low SNR conditions. The estimated segmental SNR for the

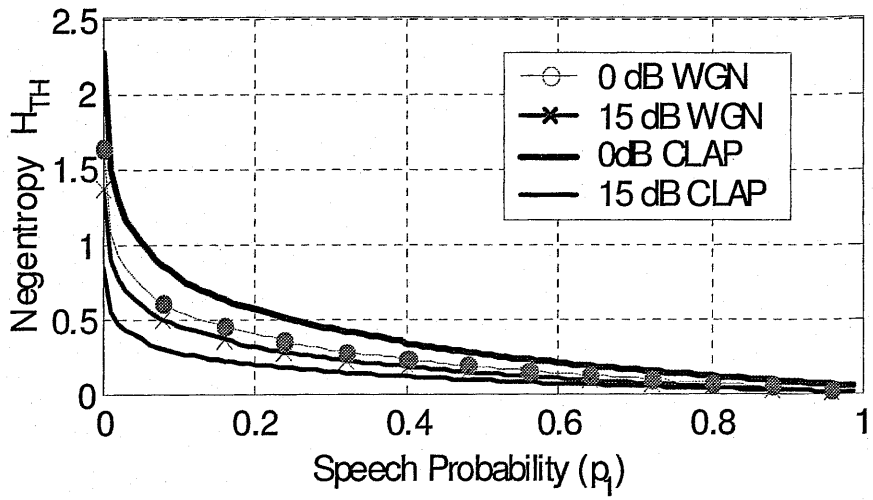


Figure 6.16 Value of threshold for different types of noise under different SNR conditions.

speech signals enhanced by Wiener filtering and MAP estimator are shown in Figure 6.19 for the speech signal degraded by babble noise. For different SNR levels of the input signal, the difference in segmental SNR of both gets lessened with increasing input SNR. The spectrograms of the noisy signals, under WGN, estimated clean speech signals by Wiener filtering and proposed method are also shown in .The averaged segmental SNR and Euclidian distances of MFCC parameters of clean speech from that of noisy and enhanced signals under the WGN and babble noise are shown in Figure 6.21and Figure 6.22. In the denoising experiments for the ICs from FDICA, the fixed-point obtain separated sources. Then one separated source was used as source of noise or cross-channel interference for other and vice-versa. The residual noise going with any separated source was estimated using Eq.(6.3) which requires SNR of the signal as a priori. The achieved NRR by FDICA algorithm for each separated

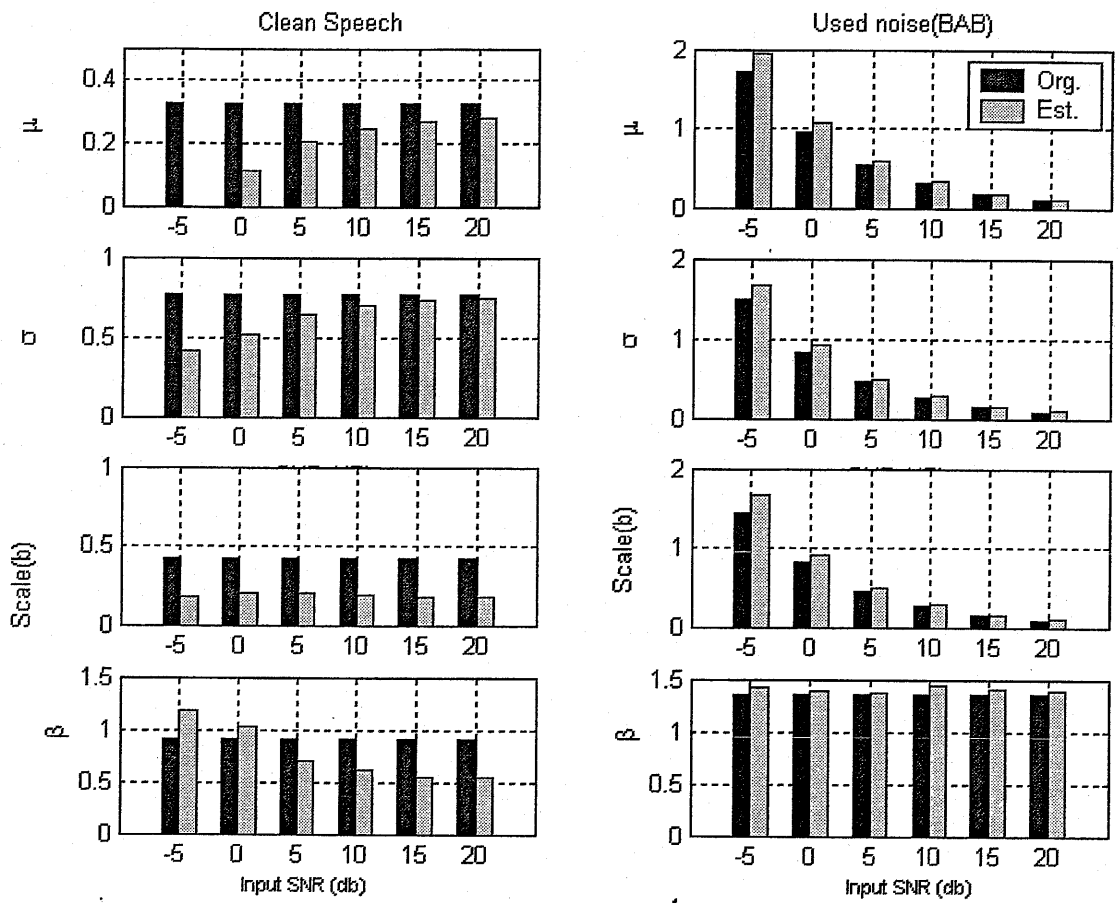
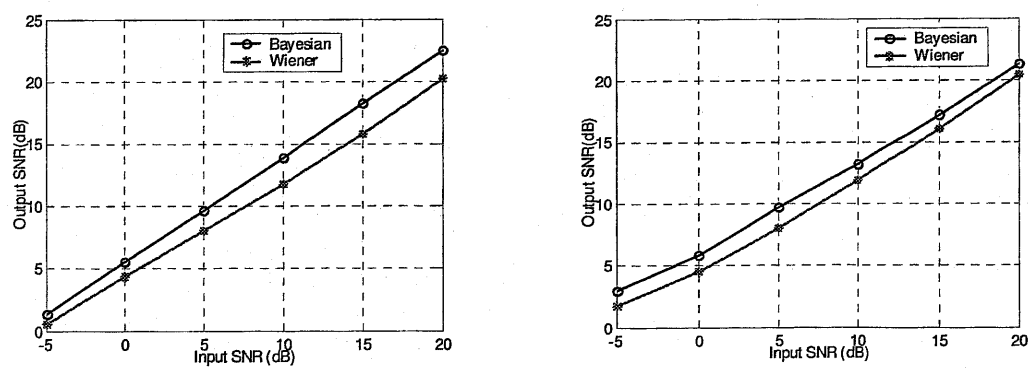


Figure 6.17 Estimated and original parameters for the clean speech and noise from the noisy speech data degraded to 0 dB SNR by babble noise. Subplots in the left column show mean, standard deviation, scale and shape parameters for the speech signal (from top to bottom) while subplots in right column show the same for the noise signal.

source has been used as a priori SNR. However, under the blind setup it is not permissible thus it is an ad hoc method for the experimental purpose. Then the noise parameters obtained from estimated residual noise were used to estimate

GGD parameters for the clean version of the separated source. Then FDICA algorithm described in Chapter-3 and Chapter 4 were used to learn ICs and MAP estimator was applied to estimate spectral components of clean signal. As a performance measure the NRR before and after denoising were estimated and are shown in Figure 6.23. It is evident from that figure that the post processing of outputs from FDICA results in cleanliness of the signal and suppression of residual interfering components from other speakers. Since under non-reverberant conditions the output of the FDICA have SNR more than 20 dB thus post processing gives no improvement, however, for the reverberant conditions the NRR achieved by FDICA is low and post-processing gives good improvement. Subjective test were also done to compare performance of the proposed method and conventional Wiener filtering. The subjective test was planned to collect preference of the 10 subjects for the enhanced signals. In the subjective test seventy-two utterances, from male and female speakers, degraded to -5, 0, 5, 10, 15 and 20 db of SNR



(a) (b)

Figure 6.18 SNR of the denoised signal MAP estimation and Wiener filtering of degraded signal by WGN (Fig. a), babble noise (Fig.b). The SNR result is averaged for four speakers (Two male and two female). The clean speech signal was degraded to –5, 0, 5, 10, 15 and 20 dB.

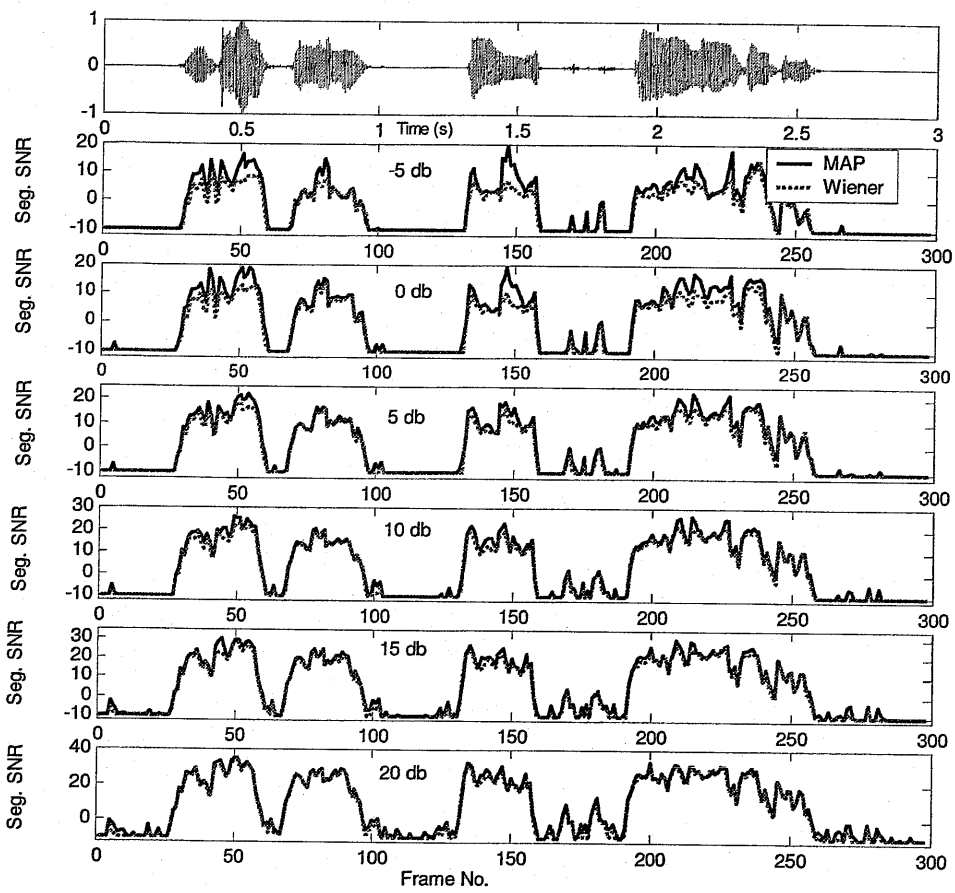


Figure 6.19 Segmental SNR for the speech signal degraded by babble noise to different SNR levels.

by WGN, BAB, AIBO's motor noise (motor noise reaching to the ear microphone of the AIBO robot) were used. The statistical characteristics of TFSS of AIBO's noise can be imbued from Figure 6.24. The shape parameter for polar magnitude of this noise is between that of Gaussian and Laplacian noise signals. The degraded speech signals were played in random manner before subject and their preference were collected. While playing speech signals the signal enhanced by MAP and Wiener filtering were played in succession for each SNR level. The averaged score collected from 10 subjects

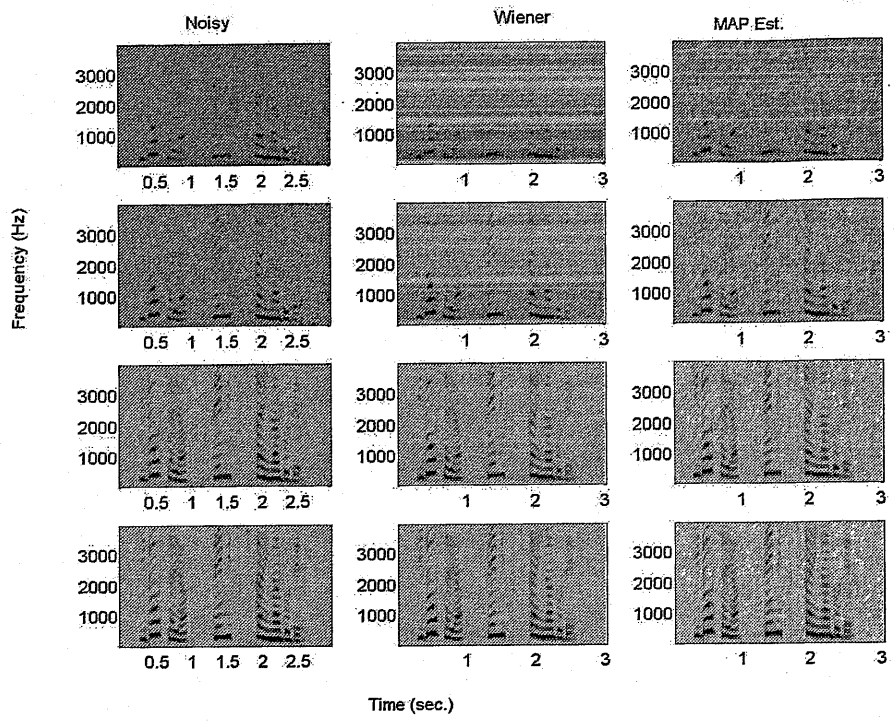


Figure 6.20 Spectrograms of the noisy and enhanced speech signals. The subplots from top to bottom in any column correspond to SNR conditions –5db, 5 db, 15 db and 20 db. Subplots in first column are for noised signals, subplots in second column are for enhanced signals by Wiener filtering, and that of in the third column are for the proposed MAP estimator.

are shown in the Figure 6.25. The averaged score is called here Mean Preference Score (MPS). The individual scoring by different subjects was found to be depending upon their aural taste for residual noise in the enhanced signal. Similar test were also carried out for the separated signal by the FDICA algorithm and for post-processed separated signal by the MAP estimator as described before. For this experiment five mixed signals recorded for RT=0ms, RT=150 ms and RT=300ms were first separated by FDICA algorithm and separated signals were post processed by MAP estimator. In this way two set of separated signals, consisting of 30 pairs, for ICA only and ICA with MAP

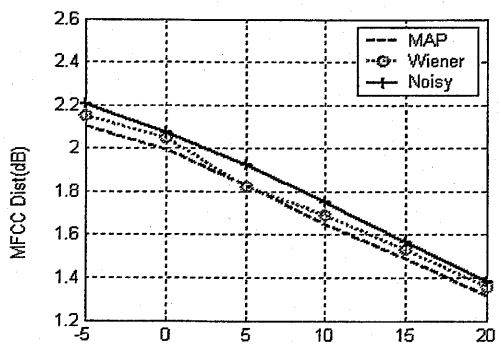


Figure 6.21 MFCC distance between noisy, clean and enhanced speech.

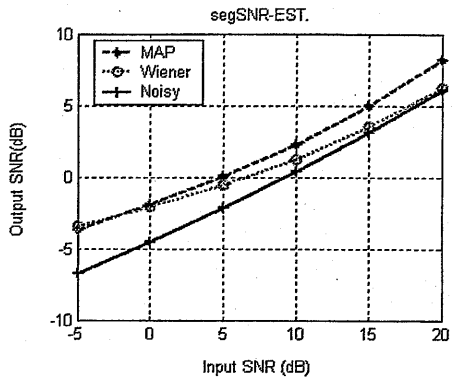


Figure 6.22 Segmental SNR under of noisy and enhanced signal.

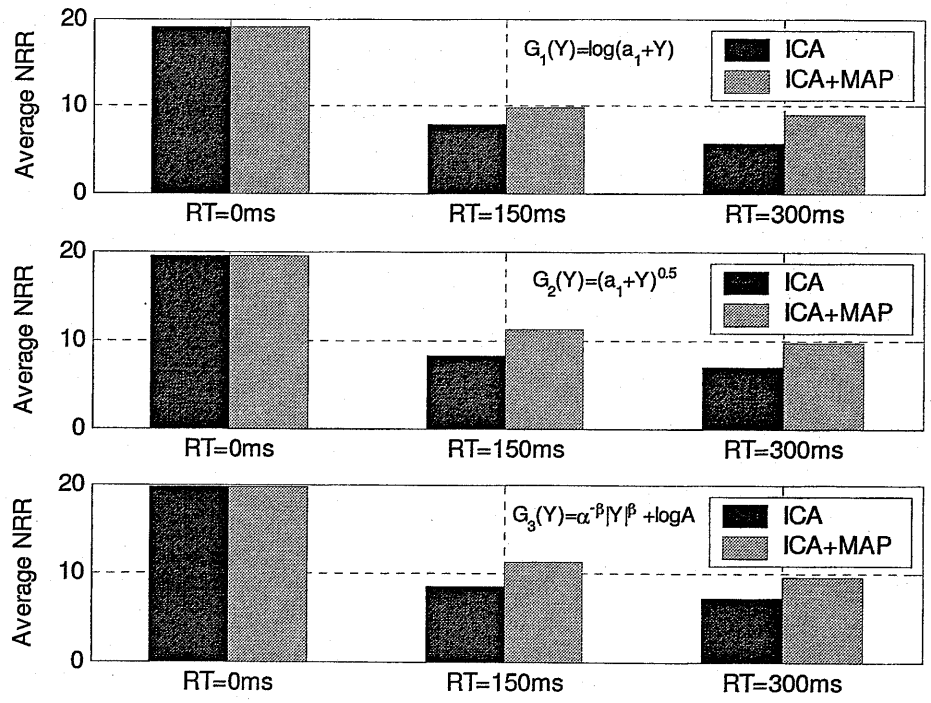


Figure 6.23 NRR performance of FDICA with MAP estimator as the post-processing enhancement scheme. Since under no reverberation ICs are at higher SNR so it remains as it is. With increasing reverberation enhancement is effective.(Shown results are averaged for six combination of speakers).

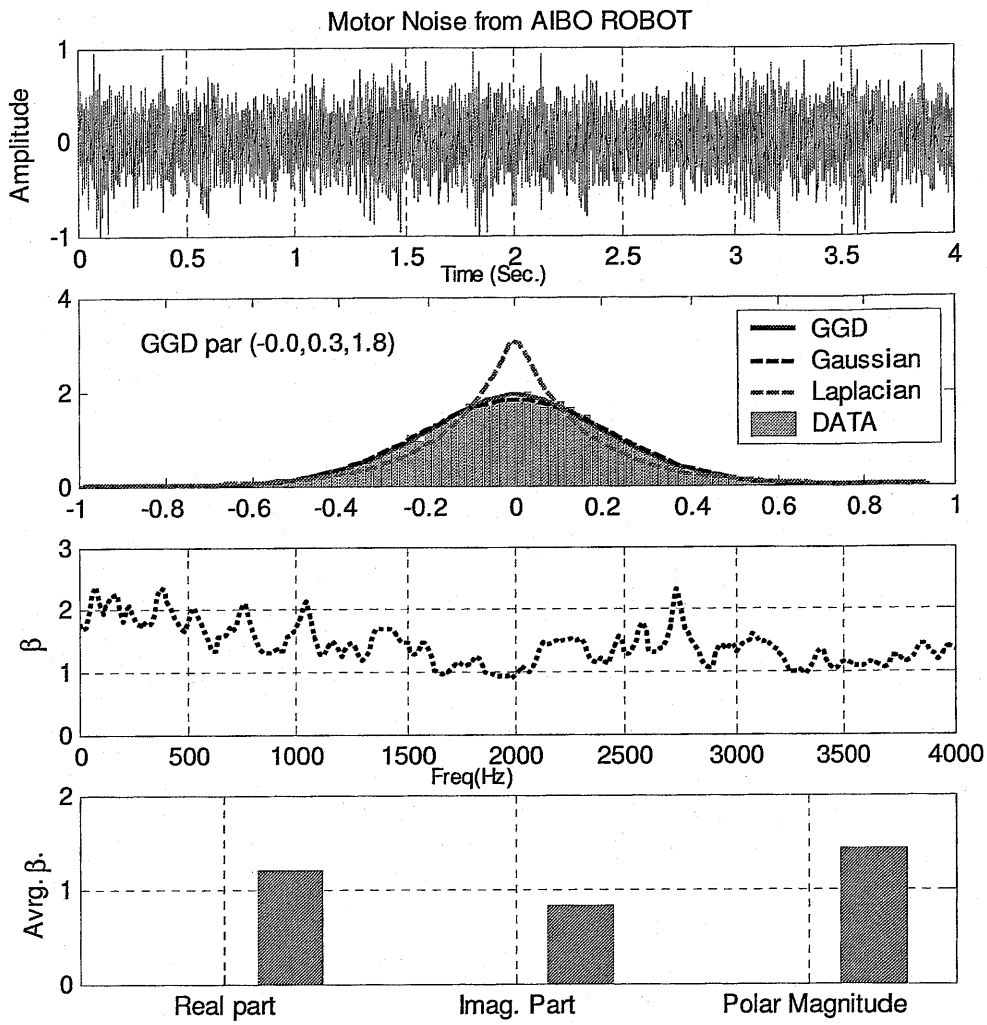


Figure 6.24 Motor noise from AIBO robot. Subplots from top to bottom shows wave form, histogram of time domain samples with GGD, GD and LD fittings, shape parameter for the TFSS and shape parameters averaged over frequency bins respectively. The shape parameter for statistical distribution of TFSS lies between that of for GD and LD.

were created. These signals (in pair) for the given RT were played in random manner before the subjects. The MPS collected from 10 subjects are shown in

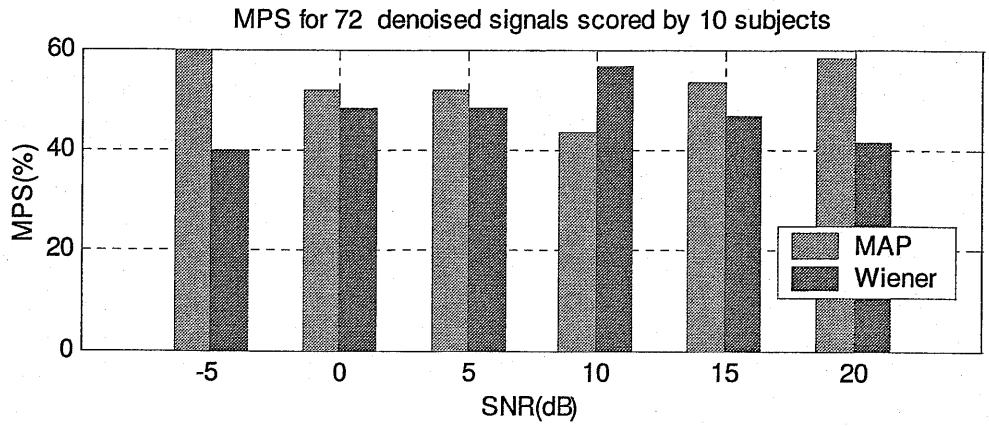


Figure 6.25 Mean Preference Score (MPS) for enhanced signal by proposed MAP estimator and Wiener filter. The used noise signals were WGN, speech like babble noise and motor noise of AIBO robot.

Figure 6.26. Since for RT=0 ms FDICA algorithm produces very clean signal post processing shows no improvement and thus the MPS score under RT=0 was hard to differentiate for subjects. The MPS for RT=150ms and for RT=300ms was found to be more for post-processed separated signal which is supplementing the results shown in Figure 6.23. Thus post processing is beneficial in case of reverberant conditions.

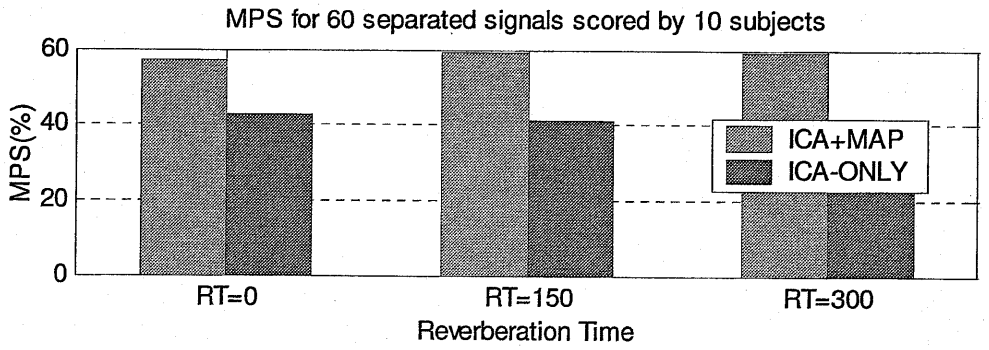


Figure 6.26 MPS (Mean Preference Score) for separated signal by ICA only and ICA with MAP estimator as preprocessing stage.

Chapter -7

Conclusions

"In research, the horizon recedes as we advance, and is no nearer at 60 than it was at 20. As the power of endurance weakens with age, the urgency of pursuit grows more intense.....and research is always incomplete." **Mark Pattison (1875).**

"An intelligent person can choose one of the best objectives, but it can not be achieved without perspiration".....**Indian saying**

In this thesis we have addressed the problem of blind separation of convoluted mixture of speech using microphone array processing technique with ICA. The study in the thesis moves around the application of non-Gaussianization based ICA as a tool for the speech signal separation in the frequency domain. The choice of non-Gaussianization based ICA was emphasized on the fact that in the mixing of speech signals the statistical property of mixed signal is also governed by central limit theorem and it has been shown in the Chapter-4 how the dogma of CLT is instrumental in increasing Gaussianity in each frequency bin in the convolute mixing process. The measure for non-Gaussianization was taken as negentropy and a fixed–point learning rule for the separation vector was derived. Also, the separation performance of the algorithm under different acoustic conditions, characterized by different reverberation time, was investigated. The performance of the algorithm drops with increase in reverberation. The effect of initialization of algorithm by random value based initial separation vector and null-beam former based separation vector was also studied to conclude the suitability of null beamformer based initial value. One of the interesting contributions of the thesis originated

from the study of spectral separation performance by observing NRR. In every frequency bin algorithm achieves significantly varying level of NRR. It should not happen in the light of assumption that in each frequency bin TFSS are assumed to be independent and thus algorithm should show matched level of performance for each data set (of each frequency bin). The cause of significant difference was investigated and it has been found that the mixing process of speech signal doesn't follow CLT in each frequency bin. The cause of failure to comply with the CLT has also been investigated to conclude that the spectral sparseness i.e. each speaker does not contribute signal in every frequency bin, of the speech sources leads to disobedience of CLT. Such frequency bins presents ill conditions to the FDICA algorithms based on non-Gaussianization because it does not full-fill the basic principle of working of the algorithm. A method based on spectral kurtosis and PDF of the TFSS has been proposed to detect such frequency bins in advance. The proposed method relies on information only from the mixed signal and thus it is also blind. We also studied the combination of null beamformer and FDICA to overcome the effect of CLT non-compliance by some frequency bins. In the proposed combination the separation process in CLT disobeying bins was switched over to null beamformer and it was found that separation performance of combination increased in the non-reverberant condition. In the reverberant conditions, the combination did not give any improvement, since the separation performance of null beam former is too poor.
This thesis has also contributed on the statistical modeling of TFSS. However, the statistical model for TFSS is a contradictory topic since long. It was an important study because in the proposed ICA algorithm the objective function is based on the PDF of the TFSS. We have compared the suitability of most widely used Gaussian and Laplacian distribution functions against a flexible parametric GGD function. Statistically, the suitability of GGD function, with shape parameter less than one, as a statistical model for TFSS was found more appropriate. Further a GGD based approximation of negentropy and its effect on the separation performance were investigated to found superiority of the GGD based non-linear function over the conventional functions. It was also

investigated how the change in the convergence nature of algorithm from quadratic to cubic is faster in the case of GGD based non-linear function that led to overall faster convergence of the algorithm under the negentropy approximation by the GGD based function.

The other contribution of the thesis was placed in Chapter-6 where an MAP estimator for the speech spectral component in DFT domain has been presented. The proposed estimator can be applied in different noise signals. It uses GGD function as a flexible model for the spectral components of both the noise and speech signals. Thus the proposed estimator is useful under different types of noise signals such as spiky noise and Gaussian noise signals. It has also been shown how the Wiener filter can be derived as the special case of the proposed MAP estimator by imposing statistical assumptions of Wiener filtering on it. The proposed MAP estimator was derived with aim to use it as a post-processing stage for the FDICA algorithms. It has been applied with the FDICA algorithm with assumption that the one IC produces interference or noise signal for other and can be cleaned by MAP estimator. The method was found effective in improving the separation performance of the FDICA algorithm. The proposed MAP estimator can also be effectively used in speech enhancement contaminated by non-speech and speech like noise. However, this requires blind estimation of involved parameters of noise and clean speech signal. In order to demark noise only parts from the noisy-speech signal a VAD algorithm based on chaos measure of signal by negentropy has been proposed. The proposed algorithm is useful and work effectively in very low SNR conditions. Some experiments on denoising speech signal under spiky and Gaussian noise were also presented. The proposed MAP estimator is strict MAP estimator because even the used prior models for the clean signal are also estimated from the ob served data.

Future Work

I started to work with a few aims and as an outcome I got a fewer results with new problems and ideas in plural number. However, some interesting and

important issues clouding my space of thought are pointed here for future research activities. The one of the important aims of the researches in blind methods for speech signal separation is its implementation for artificial audition systems such as in robot audition and in conversational interface for other machines. The main hurdles standing in the way are poor separation quality as well as computational load. However, fixed-point FDICA algorithms are computationally less expensive than the other algorithms such as gradient based algorithms but the separation quality is a bit inferior. The faster computability of the fixed-point FDICA algorithm is very strong plus point for its implementation in the real time application system if its separation capacity in reverberant conditions is enhanced too acceptable level. It can be further explored in this direction. The separation capacity of algorithm may be improved by changing non-linear function or by initialization with good values. It has been found that null beam former based initialization of the algorithm gave good quality of separated signal but for the reverberant conditions null beamformer does not give good initialization. The natural gradient based FDICA algorithm shows better separation performance, however, it takes huge amount of iterations and near convergence the rate of convergence becomes sluggish. The combination of fixed-point FDICA with natural gradient based FDICA can be combined mutualistically. The very slow convergence of natural gradient algorithm near optimal solution can be enhanced by switching over to fixed-point FDICA with initial separation vector learned by natural gradient based algorithm.

The other avenue of for future work is in the area of ICA algorithm. As we have shown how the ICA by negentropy maximization is trapped into problem due to non-compliance of CLT which occurs due to spectral sparseness of the sources. It can not be stopped to occur and algorithm has to be robust against this. Thus some other method of non-Gaussianization can be used to find independent components. As we have shown that the PDF of TFSS can be better approximated by the GGD and that the mixing process increases the Gaussianity of the mixed signal. Accordingly, the shape parameters for the underlying PDF also changes. In the context of GGD it shifts towards 2 (for

Gaussian distribution). Thus an ICA algorithm can be developed by creating some contrast function as a function of the shape parameter of the underlying distribution. Such a cost function can be used to reshape the PDF of mixed signal from more-Gaussian to less Gaussian distribution. Such a cost function can be developed using different distances e. g. KLD between the two GGD can be used, however, the relation between shape parameter and data points for KLD is enough complex. The other possibility is to use statistical rank of the data for the given GGD function. But the relative suitability of these two options, of course others also, need to be investigated before advancing in this direction. Such a method of non-Gaussianization can be effective in the reverberant conditions because its functioning is against the Gaussianization which is also enhanced in the reverberant conditions due to presence and repeated addition of reflected and delays components of the signal sources. Since optimization landscape of such cost function will be based on change in shape parameter, the problem of CLT failure may not affect the performance of algorithm.

Also, it will be not out of place to mention that the post-processing technique introduced in previous chapter for speech signal enhancement in general and for ICs in particular uses fixed GGD parameters for the TFSS. However, these parameters are not really fixed over time. Shape parameter may be fixed to some global value but scale parameter depends on local variance. The fixed parameters may be highly effective in the case of stationary noise but is rough assumption for non-stationary or speech like noise. Here is the scope for further exploration in this direction too. The algorithm with adaptive parameter estimation over time frames may improve the enhancement results.

References

[1] V. Zue,W. Glass and R. James, Conversational interfaces: advances and challenges, Proc. IEEE 88 (8), 1166–1180 (2000).

[2] R.K. Prasad, Saruwatari, H., Shikano K., "Robots That Can Hear, Understand and Talk," Advanced Robotics, Vol.18, No.5, pp-533-564, 2004.

[3] S. Furui, "Steps toward flexible speech recognition — recent progress at Tokyo Institute of Technology," in: Proc. 8th Australian Int. Conf. on Speech Science and Technology, pp.19-29, Canberra (2000).

[4] E. C. Cherry, "Some experiments on recognition of speech, with one and with two ears," J. Acoust. Soc. Amer. 25, 975–979 (1953).

[5] R. Carter, "*Mapping the Mind,*" Phoenix, London, 2000

[6] Y.Nota, K. Honda, "Brain regions involved in motor control of speech," Jr. Acoust. Sci. &Tech., ASJ,25,4,2004.

[7] S.R.Arnot,M.A. Bins, C.L Grady, C. Alain, " Assessing the dual-pathway model in humans," Jr. NeroImage, 22,pp-401-408, 2004.

[8] T.W. Parson, "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am., 60, 911-918, 1976.

[9] K.Kashino, K. Nakadai, T.Kinoshita, and H.Tanaka, "Organization of hierarchical perceptual sounds," Proc. 14th Int. conf. On Artificial Intelligence, vol.1, 158-164, 1995.

[10] M. Unoki, M.Akagai, "A method of signal extraction from noisy signal based on auditory scene analysis," Speech Communication, 27, 261-279, 1999.

[11] Frost, "An algorithm for linearly constrained adaptive array processing," Proc. IEEE, 60, 926-935, 1972.

[12] L.J. Griffiths, C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas and Propag., 30, 27-34, 1982.

[13] Y. Kaneda, J.Ohga, "Adaptive microphone array system for noise reduction," IEEE Trans. Acoust., Speech and Signal procs., ASSP-34, 27-34, 1986.

[14] H. Saruwatari, T. Kawamura, K. Shikano, "Blind source separation for speech based on fast convergence algorithm with ICA and beamforming" Proc. EUROSPEECH-2001, 2603-2606, 2001.

[15] S. Haykin, "Unsupervised adaptive filtering," John Wiley and Sons Inc.,NewYork, 2000.

[16] T.W.Lee, "*Independent component Analysis*," Norway, Kulewar Academic Press, 1998.

[17] P.Comon, "Independent component analysis, a new concept?,"Signal Processing, Vol.36, pp.287-314,1994.

[18] P. Samaragadis, "Blind separation of convolved mixture in frequency domain," Neural Computing Surveys, vol.2, pp-94-128, 1999.

[19] A. Hyvarinen, et al, "*Independent component analysis*," John Wiley & Sons, Inc., 2001.

[20] J.F Cardoso., et al., "Independent component analysis of the cosmic microwave background," Proc. ICA2003, 1111-1116, Nara, Japan, 2003.

[21] J.F. Cardoso, "On the performance of orthogonal source separation algorithms," Proc. of EUSIPCO-94, Edinburgh, 1994.

[22] A. Hyvärinen, "Survey on independent component analysis", Neural Computing Surveys, 94-128, 1999.

[23] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," IEEE Transactions on Neural Networks 10(3):626-634, 1999.

[24] T.W.Lee, "Independent component analysis," Kulewar Academic Press, Norway, 1998.

[25] J.P.LeBlanc, P. L. DeLeon, "Speech Separation by Kurtosis Maximization," IEEE Int.Conf. on Acoustics, Speech and Signal Processing, Seattle, WA, May, 1998.

[26] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," Neural Computation, 9(7):1483-1492, 1997.

[27] E. Bingham et al., "A fast fixed point algorithm for independent component analysis of complex valued signal," Int. J. of Neural System, 10(1)1: 8, 2000.

[28] N. Mitianoudis, N. Davies, "New fixed-point solution for convolved audio source separation," Proc. IEEE Workshop on Application of Signal Processing on Audio and Acoustics, New York, 2001.

[29] R.K. Prasad, H.Saruwatari, A. Lee, K. Shikano, "A fixed point ICA algorithm for convoluted speech separation," Proc. International Symposium on ICA &BSS, pp-579-584, Nara, Japan, 2003.

[30] S.Kurita, H.Saruwatari, S.Kajita, K.Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant condition," Proc. ICASSP2000, vol.5, pp.3140-3143, 2000.

[31] R. K. Prasad, H. Saruwatari, K. Shikano, "Problems in blind separation of convolutive speech mixture by negenotropy maximization", Proc. IWAENC 2003, Kyoto, Japan, 287-290, 2003.

[32] J.F. Cardoso, "Eigenstructure of 4th order cumulants tensor with application to the blind source separation problem," Proc. ICASSP'89, 2109-2112, 1989.

[33] C. Jutten, J. Herault,, "Blind separation of sources part 1: An adaptive algorithm based on neuromimetic architecture," J. Signal Processing, 24, 1-10,1991.

[34] J.F. Cardoso, C.N.R.S, E.N.S.T, "Blind signal separation: statistical principles," Proc. of IEEE, vol.9, no.10, pp-2009-2025, Oct.1981.

[35] S. Araki et al., "The fundamental relation of frequency domain blind source separation for convolute mixtures of speech," IEEE Trans. Speech and Audio Processing, Vol. 11, no.2, 109-116, 2003.

[36] S. Ikeda., S. Murata, "A method of ICA in time-frequency domain", Proc. Workshop Indep. Compon. Anal. Signal Sep., 365-367, 1999.

[37] T. Nishikawa, et al., "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," IEICE Trans. Fundamentals, Vol.E86-A, pp.846-58, no.4, April, 2003.

[38] K. Torkkola, "Blind separation for audio signals-are we there yet?" Proc. Workshop on ICA & BSS, France, 1999.

[39] D.Gabor,"Theory of communication', J. IEE (London), vol.93, pp.429-457, 1946.

[40] L.Cohen, "Time-frequency distributions-a review," In proc. of IEEE, vol.77, No.7, pp-941-979, July 1989.

[41] P.J. Loughlin, J.W. Pitton, L.E. Atlas, "Bilinear time frequency representations: new insights and properties," IEEE Trans. on Signal Processing, vol.41, no. 2, pp-750-767, Feb.1993.

[42] D.R.Brillinger, "Fourier analysis of stationary process," Proc. IEEE, vol.62, pp.1628-1643, Dec.1974.

[43] W. Wokurek, H. Franz,and G.Kubin, " Wienger distribution analysis of speech signals," Proc. of Int. conference on DSP, pp-294-298, Florence, 1987.

[44] F. Halawatsch, "Inference terms in the Weigner distribution," Proc. Int. Conf. Digital Signal Processing, pp-363-367, Flourance, 1984.

[45] E. Bingham, "Advances in independent components analysis with applications to data mining," Ph.D. Thesis, Helsinky University of Technology, Finland, 2003.

[46] A. Cichocki, S. Amari, "Adaptive blind signal and image processing, learning algorithms and application," John Wiley & Sons Ltd., 130-131, 2002.

[47] H.Johnson, et. al., "Array signal processing concepts and techniques," Prentice Hall, 1993.

[48] H. Sawada, R. Mukai, S. Araki, S.Makino, "A robust approach to the permutation problem of frequency-domain blind source separation," IEEE International Conference on Acoustics, Speech, and Signal (ICASSP2003), 381-384, 2003.

[49] S.J. Godsill, et al., "*Digital audio restoration*," Springer Verlag London, 1998.

[50] J.M. Mendel, "Tutorial on higher order statistics (spectra) in signal processing and system theory: theoretical results and some applications, "Proc. IEEE, vol.79, No.3, 277-305, 1991.

[51] R.F. Dwyer, "Use of kurtosis statistics as aid in detecting random signals," IEEE Journal of Ocean Eng., vol.OE-9, no.2.85-92, 1984.

[52] C. Nikias,, A. Petrupulu,, "Higher-order spectral analysis-a nonlinear signal processing framework," 1993. Prentice Hall.

[53] L. Chrysostomos, et al., "Signal processing with higher order spectra," IEEE Signal processing magazine, 10-37, 1993.

[54] P.O. Amblard et al, "Statistics for complex variable and signals-part I," Signal Processing, vol-53, 1-25, 1996.

[55] R. Scott, et al., "Real-time time-frequency based blind source separation," Proc. of ICA2001, Dec. 9-13, San Diego, CA, 2001.

[56] T. Nakamura et. al., "Acoustic modeling based on generalized Laplacian distribution," Proc. EUROSPEECH, Budapest, 1999.

[57] W.B. Davenport, "An experimental study of speech wave probability distributions," J. of Acous. Soc. America, vol.24, no.4, pp.390-399, July 1952.

[58] D.L. Richards, "Statistical properties of speech signal," Proc. IEE, vol.111, no. 5, pp. 941-49, 1964.

[59] H.Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signal," Signal Processing, vol. 12, no. 2. pp. 119-141, Mar. 1987.

[60] S.Gazor and W. Zhang, "Speech probability distribution," IEEE Signal Processing Letters, vol.10, no.7, July 2003.

[61] J.M. Tribolet and R.E. Crochiere, "Frequency domain coding of speech," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27, pp. 522, Oct.1979.

[62] R. Zelsinki and P. Noll, "Adaptive transform coding of speech signal," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, p. 306, Aug. 1977.

[63] J.E. Porter and S.F. Boll, "Optimal estimators for spectral restoration of noisy speech,", in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, March 1984.

[64] A. Ephraim and D.Malah, "Speech enhancement using a spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, no. 6, pp. 1109-21, Dec.1984.

[65] E.A Wan and A.T. Nelson, "Removal of noise from speech signal using dual EKF algorithm," IEEE Conf. Acoust., Speech and Signal Processing, 1998.

[66] S. Jongseo, N.S. Kim, and W. Sung, "A statistical model based voice activity detection," IEEE Signal Processing Letters, vol. 6, no.1, January 1999.

[67] P. P.O. Amblard, M. Gaeta, J.L. Lacoume, "Statistics for complex variable and signals-part II", Signal Processing, vol-53,pp-1-25,1996.

[68] H. Sawada, R. Mukai, S. Araki, S. Makino, "Polar Coordinate based Nonlinear Function for Frequency Domain Blind Source Separation," IEICE Trans. Fundamentals, vol. E86-A, no.3, pp. 590-596, March 2003.

[69] G.E.P. Box, and G.C.Tiao, "Bayesian inference in statistical analysis," Adison Wesley, Readinbg, Massachussete, 1973.

[70] R.V. Hogg, "Adaptive robust procedure: A partial review and some suggestions for the future," Journal of American Statistical Association, vol.69, pp.909-923, 1974.

[71] D.S. Boraca, "The distribution of Indian stock return: A tale of two tails," Decision, vol.29, No.1, Januaary-June, 2002.

[72] L.Balogh, I.Kollar and G.Gueret, "Variance of Fourier coefficients calculated from overlapped signal segments for system identification," IEEE Instrumentation and Measurement technology Conference, Alaska, USA, 2002.

[73] C. A. Greenhall, D. A. Howe and D. B. Percival, "Total variance: An estimator of long term frequency stability," IEEE Transaction on Ultrasonics, Ferroelectrics and Frequency Control, Vol.46, No.5, 1999.

[74] J.L Devore, "Probability and statistics for engineering and science," Duxbury Press, Belmont, 1995.

[75] M.K. Varanasi and B. Aazhang, "Parametric generalized Gaussian density estimation," Journal of Acost. Society of America, 86(4), pp. 1404-1414, October 1989.

[76] M. N. Do, M. Vetterli, "Wavelet based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," IEEE Transaction on IP, vol.11, No.2, Feb.2002.

[77] B. Sanker, A. M. Charles, and A.O. Peder, "Maximum Likelihood estimates for exponential type density function," Proc. of ICASSP, 1998.

[78] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/.

[79] Papoulis A. & Pillai S., "Probability, random variables and stochastic process," McGraw Hill, 2002.

[80] T. Kobayashi, S. Itabashi, S. Hayashi, and T. Takewaza, "ASJ continuous corpus for research," J. Acoustic Soc. Jpn., vol.48, no.12, pp-888-893, 1992.

[81] P. Kismode, "Alpha-stable distributions in signal processing of audio signals," Proc.SIM2000, Technical University Denmark, 2000.

[82] R. K. Prasad, H. Saruwatari, K. Shikano, "Probability distribution of the time-series of speech spectral component," Journal of IEIEC Trans. Fundamental, vol.E87-A, no.3, March 2004.

[83] R.K Prasad, H.Saruwatari, K.Shikano, "Blind Detection of CLT Disobeying Frequency Bind for Audio Source Separation by Fixed-Point ICA," 18th International Congress on Acoustics (ICA2004), vol. IV, pp.3151--3154, April 2004.

[84] J. Shao, D.Tu, "The Jackknife and Bootstrap," Springer, New York, 1995.

[85] B. Efron , R.J. Tibshirani, " An Introduction to Bootstrap," Chapman and Hall, London, 1993.

[86] S. Furui,, M.M. Sondhi,, (Eds.), "Advances in speech signal processing,", Narcel Dekker Inc., USA, 1991.

[87] F. Q. Thomas, "*Discrete Time Speech Signal Processing:Principle and Practice*," Prentice Hall, 2002.

[88] J.H.L. Hansen, B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," International Conference on Spoken Language Processing (ICSLP), vol. 7, pp. 2819-2822, Sydney, Australia, 1998.

[89] H.L. John, and A. C. Mark, "Constrained speech enhancement with application to automatic speech recognition," IEEE Proc. International Conference on Acoustics, Speech, and Signal Processing, pp. 561-564, New York, NY, April, 1988.

[90] M.R. Weiss, E. Aschkenasy,, and T.W. Parsens,, "Study and development of the INTEL technique for improving speech intelligibility," Report NSC-FR/4023,Nicolet Scientific Corp., December, 1974.

[91] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. , Speech, Signal Process. ASSP-27, pp. 113-120, 1979.

[92] Schwartz et al., "Enhancement of speech corrupted by acoustic noise," ICASSP' 79, pp. 208-211, Washington, 1979.

[93] J. Lim, and A.Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. IEEE 67, pp. 1586-1604, 1978.

[94] A. Ephraim, D. Malah, "Speech enhancement using optimal non-linear spectral amplitude estimation," ICASSP'83, Boston, 1983.

[95] A. Ephraim, D. Malah, "Speech enhancement using a minimum mean square error short-term spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-32, no. 6 (Dec.1984) 1109-21, 1984.

[96] R.J. McAulay, and M.L. Malpass, "Speech enhancement using soft decision noise suppression filter," IEEE Trans. Acoust., Speech, Signal Process. ASSP-28, pp.137-145, 1980.

[97] J.E. Porter, and S.F. Boll, "Optimal estimators for spectral restoration of noisy speech," In Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, March, 1984.

[98] R. Martin, B. Colin, "Speech enhancement in the DFT domain using Laplacian priors," Proc. IWAENC, Kyoto, Japan (2003), pp. 87-90, 2003.

[99] L. Thomas, and V. Peter, "Noise reduction by maximum a posteriori spectral amplitude estimation with super-Gaussian speech modeling. Proc. IWAENC, Kyoto, Japan, pp. 83-86, 2003.

[100] G.G. Panayotis, T. Panagiotis, K. Chris, "Alpha-stable modeling of noise and robust time delay estimation in the presence of impulsive noise," IEEE Trans. on Multimedia, Vol.1, No.3, pp. 291-301, 1999.

[101] R.N. McDoough, and A.D. Whalen, "Detection of signals in noise," Academic Press Inc., USA, 1995

[102] J. Stadermann, V. Stahl, G. Rose, 2001. Voice activity detection in noisy environments Proc. of EUROSPEECH 2001, pp. 1851-1854.

[103] J.L. Shen, J.Hung, L.S.Lee, "Robust entropy based end point detection for speech recognition in noise," Proc. 5th International conference ICSLP'98, Sydney, Australia. 1998.

[104] D. V. Compernolle, "Noise adaptation in hidden Markov model speech recognition system," Computer Speech and Language, 3:151-67, 1989.

My Publications related to Thesis

Journal Papers

1. Rajkishore Prasad, Hiroshi Saruwatari, and Kiyohiro Shikano, ``Probability distribution of time-series of speech spectral components," IEICE Trans. Fundamentals, Vol.E87-A, No.3, pp.584--597. 2004.
2. Rajkishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano, "Robots that can hear, understand and talk," Advanced Robotics, Vol.18, No.5, 2004.
3. Rajkishore Prasad, Hiroshi Saruwatari, and Kiyohiro Shikano, "Negentropy based voice-activity detection for noise estimation in very low SNR condition", IEICE Electronics Express, vol.1, No.16, pp.495-500, Nov. 2004.
4. Rajkishore Prasad, Hiroshi Saruwatari, and Kiyohiro Shikano, " An ICA algorithm for separation of convolutive mixture of speech signal," Intl. Journal of Signal Processing (IJSP), vol. 1, no. 3, 2004.
5. Rajkishore Prasad, Hiroshi Saruwatari, and Kiyohiro Shikano, " Blind separation of speech by fixed-point ICA with source adaptive negentropy approximation, IEIEC Journal (Accepted).

Conference Papers

1. Rajkishore Prasad, Hiroshi Saruwatari, Akinobu Lee, Kiyohiro Shikano, "A fixed-point ica algorithm for convoluted speech signal separation," Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Signal Separation, pp.579--584, April 2003
2. Raj Kishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano, "Problems in blind separation of convolutive speech mixtures by negentropy maximization", Proc. International workshop on Acoustic Echo and Noise Control (IWAENC-2003), Kyoto, Japan.
3. Rajkishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano, "Combining null beamformer and fixed-point ICA for the blind separation of convolutive mixture of speech", International conference on Intelligent Signal Processing and Robotics, February 2004.

4. Raj Kishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind detection of CLT disobeying frequency bins for audio source separation by fixed-point ICA," Proc. 18th International Congress on Acoustics (ICA2004), April 2004
5. Raj Kishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano, "Single channel speech enhancement: MAP estimation using GGD prior under blind setup," Proc.5th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2004), Sept. 2004, pp. 873-880.
6. Raj Kishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano, "Noise estimation using negentropy based voice activity detector, Proc. of 47th IEEE Midwest Symposium on Circuits and Systems, Hiroshima, pp. 149-152, Japan.
7. Raj Kishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano, "MAP estimation of speech spectral component under GGD a priori,"Proc. of SAPA 2004, Korea, (CD-ROM Proc.) .