Development, Simulation and Implementation of New Strategies based on Soft Computing for Real Time Speech Processing in Multimedia Applications

A thesis submitted for the award of the

Degree of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

By

Milind Uttam Nemade

THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA

सत्यं शिवं सुन्दरम्

ELECTRICAL ENGINEERING DEPARTMENT
FACULTY OF TECHNOLOGY AND ENGINEERING
THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA
VADODARA – 390 001
GUJARAT, INDIA

December 2013

CERTIFICATE

This is to certify that the thesis entitled***, 'Development, Simulation and Implementation of New Strategies based on Soft Computing for Real Time Speech Processing in Multimedia Applications'*** submitted by ***Milind Uttam Nemade*** in fulfillment of the degree of ***Doctor of Philosophy in Electrical Engineering***, is a bonafide record of investigations carried out by him at the Electrical Engineering Department, Faculty of Technology and Engineering, The M. S. University of Baroda, Vadodara under my guidance and supervision. In my opinion the standards fulfilling the requirements of the Ph.D. Degree as the prescribed regulations of the University has been attained.

December 2013

Prof. Satish K. Shah
Guide,
Department of Electrical Engineering,
Faculty of Technology and
Engineering,
The M. S. University of Baroda,
Vadodara – 390 001

Head,
Department of Electrical Engineering,
Faculty of Technology and Engineering,
The M. S. University of Baroda,
Vadodara – 390 001

Dean,
Faculty of Technology and Engineering,
The M. S. University of Baroda,
Vadodara – 390 001

CERTIFICATE

This is to certify that the thesis entitled**, *'Development, Simulation and Implementation of New Strategies based on Soft Computing for Real Time Speech Processing in Multimedia Applications'*** submitted by ***Milind Uttam Nemade*** in fulfillment of the degree of ***Doctor of Philosophy in Electrical Engineering***, is a bonafide record of investigations carried out by him at the Electrical Engineering Department, Faculty of Technology and Engineering, The M. S. University of Baroda, Vadodara under my guidance and supervision. In my opinion the standards fulfilling the requirements of the Ph.D. Degree as the prescribed regulations of the University has been attained.

December 2013

Prof. Satish K. Shah
Guide,
Department of Electrical Engineering,
Faculty of Technology and
Engineering,
The M. S. University of Baroda,
Vadodara – 390 001

DECLARATION

I, **Milind Uttam Nemade** hereby declare that the work reported in this thesis entitled '**Development, Simulation and Implementation of New Strategies based on Soft Computing for Real Time Speech Processing in Multimedia Applications**' to be submitted by me for the award of the degree of **Doctor of Philosophy in Electrical Engineering** is original and has been carried out at the Department of Electrical Engineering, Faculty of Technology & Engineering, M. S. University of Baroda, Vadodara. I further declare that this thesis is not substantially the same as one, which has already been submitted in part or in full for the award of any degree or academic qualification of this University or any other Institution or examining body in India or abroad.

December 2013 **Milind Uttam Nemade**

(Roll No. 436)

In memory of my mother Smt. Shalini U. Nemade, father Late Shri. Uttam R. Nemade, brothers and my beloved wife whose beautiful smile transform my darkest days into sunshine

ACKNOWLEDGMENTS

It gives me immense pleasure to express my sincere gratitude towards my guide Prof. S. K. Shah for his constant help and support without him this thesis would never have been written. Hence, I would like to share the following short tale.

I started working as a research scholar four years ago, I met Prof. S. K. Shah and he agreed to be my supervisor and letting me start as a doctoral candidate at his department, since he has been truly supportive and encouraged me to develop my research interests. He taught me how to locate the problem of study and how to have wide freedom of thought in selecting the research area. His motivation and moral support helped me to expand my perspective by attending national and international conferences/ short term workshops and had an opportunity to interact with many eminent personalities in the field. I would like to thank him for his valuable suggestion related to writing the good research proposal and research papers, which will enhance quality of paper. It has been an honor to be his research scholar. Apart from research, he has taught me many ethics which I am sure going to be extremely useful in my professional life.

I take this opportunity to thank Dr. Archana Bhise, Principal of K. J. Somaiya Institute of Engineering and Information Technology, Sion (E), Mumbai for giving me the opportunity for joining the doctorate studies and providing me financial support for participation/ publication of my work in international conferences. She has taken a keen and active interest in timely completion of my Ph.D. work and helped me a lot in the preparation of research proposal, synopsis and thesis writing. I am very thankful to all faculty members and supporting staff of Electronics and Telecommunication Department and trekking group of KJSIEIT, Sion, Mumbai for providing stimulating working conditions, their great support and for playing the role of listening group of various speech outputs and their evaluations in my PhD work.

I would also like to give thanks to the Electrical Engineering Department and Faculty of Technology and Engineering, Vadodara, for providing an excellent research and study environment. I thank, Prof. S.K. Joshi (Head of Department), Prof. P. P. Nehte, Prof. Heena

Chandwani and all other faculty members of electrical department, Dr. R. D. Kanphade (Principal, Dhole Patil College of Engineering, Pune), Dr. Pande (IIT Mumbai), Dr. Milind Shah (IIT Mumbai), Prof. S. S. Shitole (Vice Principal, SNDT Santacruz(W), Mumbai) for their valuable suggestions and help which directly and indirectly had a great impact on my work. I also wish to acknowledge, Dr. Kotcha, for her proofreading of the manuscript and good sense of humor.

During my stay at Anand, I would like to thanks two professors. First, I would like to thank Prof. Jitendra Chaudhari for his kind support and help. Secondly, I would like to thank Prof. M.A. Lokhandwala, my roommate for his valuable suggestion and comments in my work. During my stay in V.S. Hall (Post Graduate Hall), I would like to thanks Dr. N. A. Patil, Mr. Sunil Patil and Dr. D. P. Bharambe for his constant encouragement, moral support and taking parental care such as always asking me to keep track of deadlines for submission of semester wise progress report, synopsis, thesis etc.!!. I thank my friends Prof. Rane, Dr. Harish, Mr. Hemant, Mr. Riyaz, Mr. Mohit, Mr. Rahul and all other for providing me nice company and discussion in my work. I am also thankful to Prof. Nimit Shah and Prof. Rakeshbhai Patel, my colleagues, for their timely help in understanding the concept of Genetic Algorithm and its implementation in MATLAB.

Finally, I owe my parents for always being there for me and teaching me the inherent value of education. Also special thanks to my beloved wife and son, for their consistence support with unlimited love, endurance and scarifies. My wife is my most valuable assets and I am eternally grateful for her emotional steadfastness. Last but not the least; I am thankful to all who helped me directly or indirectly in moving ahead towards my objective. The research grants provided by University of Mumbai are gratefully acknowledged.

Milind Uttam Nemade

ABSTRACT

The “speech signal” is an integral part of most of the multimedia applications apart from the personal communication. Desired speech signal is usually impure by background noise. As a result, the speech signal has to be “cleaned” with signal processing utensils before it is played out, transmitted, or stored. The broad objective of this thesis is to devise new strategies based on soft-computing techniques for real time speech processing in multimedia applications.

The speech recognition, being a multimedia application under consideration here, has been a very important system in almost every area of life. The intelligent speech enhancement techniques can raise the outcome of speech recognition and hence it is very important to know the basics involved in it. Here attempt has been made to survey the broad categories of speech enhancement techniques such as speech filtering techniques, beam forming techniques, active noise cancellation methods and to discuss, how these techniques affect the performance of speech recognition applications.

This thesis concerned with designing speech recognition system using beamforming technique which has basically two fold objective. First is to improve the speech recognition performance in multi-microphone environment and second, the attempt has been made to analyse the performance of speech recognition against the filter-bank parameters; filter length and number of subbands.

In experimental setup, dataset is constructed using beamforming parameters and optimization in all the subsets of experiments with different parameters of beamforming using soft computing technique such as Genetic Algorithm (GA) have been explained. Experimental setup also described how to improve performance of beamforming based speech recognition system using GA and how the system is made to be real time by reducing the time required for classifier dramatically.

Finally the speech recognition task is implemented on TMS320C6713 DSP from Texas Instruments using DSP Starter kit- DSK 6713 from Spectrum Digital Incorporation. Code Composer Studio version 3.3 from Texas Instruments is used as development tools. The results of implementations are generated and commented.

CONTENTS

Chapter-4

Chapter-5

Chapter-6

Chapter-7

Chapter-8

Evolutionary Computation based Real Time Speech Beamforming for Multimedia Applications 74-84

Chapter-9

Real Time Implementation of Speech Recognition 85-105

LIST OF ABBREVIATIONS

ANC	Active Noise Cancellation
ANE	Active Noise Equalization
ANN	Artificial Neural Network
AR	Autoregressive Process
ASR	Automatic Speech Recognition
AWF	Adaptive Weiner Filtering
CCS	Code Composer Studio
CNS	Central Nervous System
CSA-BF	Constrained Switched Adaptive Beamforming
DASB	Delay and Sum Beamforming
DAT	Data File
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DIT	Integrated Digital Audio Interface Transmitter
DSK	DSP Starter Kit
ED	Euclidian Distance
EDMA	Enhanced Direct-Memory-Access
EMIF	External Memory Interface
FIR	Finite Impulse Response
FIS	Fuzzy Interface System
FKBS	Fuzzy Knowledge Based Systems
FS	Fuzzy System
FV	Feature Vector
FXLMS	Filtered X Least Means Square Algorithm
GA	Genetic Algorithm

GMM	Gaussian Mixture Model
GPGPU	General Purpose General Processor Unit
GPIO	General-Purpose Input / Output
GPU	General Purpose Unit
GUI	Graphical User Interface
GSC	Generalized Sidelobe Canceller
HPI	Host Port Interface
HMM	Hidden Markov Model
IDE	Integrated Development Environment
I^{2}C	Inter-Integrated Circuit
I^{2}S	Inter-IC Sound
JTAG	Joint Test Action Group
LAR	Log Area Ratio Measure
LLR	Log Likelihood Ratio Distance Measure
LSA	Log Spectral Amplitude
LMS	Least Mean Square
LPC	Linear Predictive Coefficients
McASP	Multichannel Audio Serial Ports
McBSP	Multichannel Buffered Serial Ports
MIPS	Million Instructions Per Second
MFCC	Mel Frequency Cepstral Coefficients
MFC	Mel Frequency Cepstrum
MFLOPS	Million Floating-point Operations Per Second
MMACS	Million Multiply-Accumulate operations per Second
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
Ninput	Number of input samples
NN	Nearest Neighbor
Noutput	Number of output samples

NRMSE	Normalized Root Mean Square Error
PLL	Phase-Locked-Loop
PSD	Power Spectral Density
PSEQ	Perceptual Evaluation of Speech Quality
PSNR	Peak Signal To Noise Ratio
RASTA	Relative Spectral Analysis
RFFT	Real Fast Fourier Transform
RLS	Recursive Least Square
RTDX	Real Time Data Exchange
SPI	Serial Peripheral Interface
SRP-PHAT	Steered Response Power with Phase Transform
SS	Spectral Subtraction
SSNR	Segmental Signal to Noise Ratio Measure
STDFT	Short Time Discrete Fourier Transform
STSA	Short Time Spectral Amplitude
TDE	Time Delay Estimator
TF	Transfer Function
ULA	Uniform Linear Array
VAD	Voice Activity Detector
VLC	Video LAN Client
VLIW	Very Large Instruction Word
VoIP	Voice over Internet Protocol
WAV	Wave file
WER	Word Error Rate
WSS	Weighted Spectral Slope Measure
WT	Wavelet Transform

LIST OF SYMBOLS

$x(n)$	Noise free speech signal
$\widehat{x}$ (n)	Estimation of noise free speech signal
m_x	Mean of noise free speech signal
Vx	Variance of noise free speech signal
Vd	Variance of noisy speech signal
$\vec{a}_{\emptyset}$	Linear Prediction (LP) coefficient vector
$\vec{a}_{d}$	Processed speech coefficient vector
σ_d^2	All pole gain for processed speech signal
$\sigma_{\emptyset}^2$	All pole gain for clean speech signal
r_{Φ}	Reflection coefficients for original signal
r_d	Reflection coefficients for processed signal
K	Sound pressure level of original signal
$\widehat{K}$	Sound pressure level of enhanced signal
σ_x^2	Mean square of speech signal
σ_d^2	Mean square difference between original and reconstructed speech signal
N	Length of the reconstructed signal
$\|x-r\|^2$	Energy of difference between original and reconstructed signal
wi	Weighted link
η	Learning constant
d	Desired output
s(n)	Spoken speech
A(s, k)	Amplitude coefficients
b(s, n)	Filter Coefficients
n	Number of microphones
Y_s	sth microphone received speech

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Speech is the fundamental and common medium, hence important for us, to communicate and most effective and reliable means for expressing oneself for personal communication. With advancement in hardware technologies, there are so many electronic and mobile personal communication based devices available, today in market and that too in cheaper cost and with easy availability. Fig.1.1 shows some typical speech communication applications [1] and most effective and reliable means for expressing oneself for personal communication.

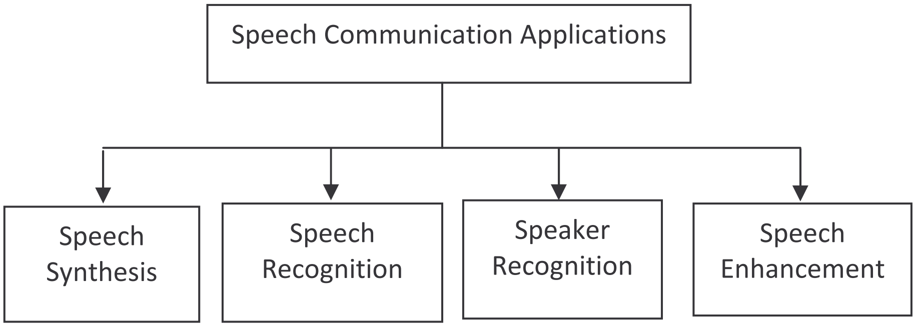


Fig. 1.1 Speech communication applications

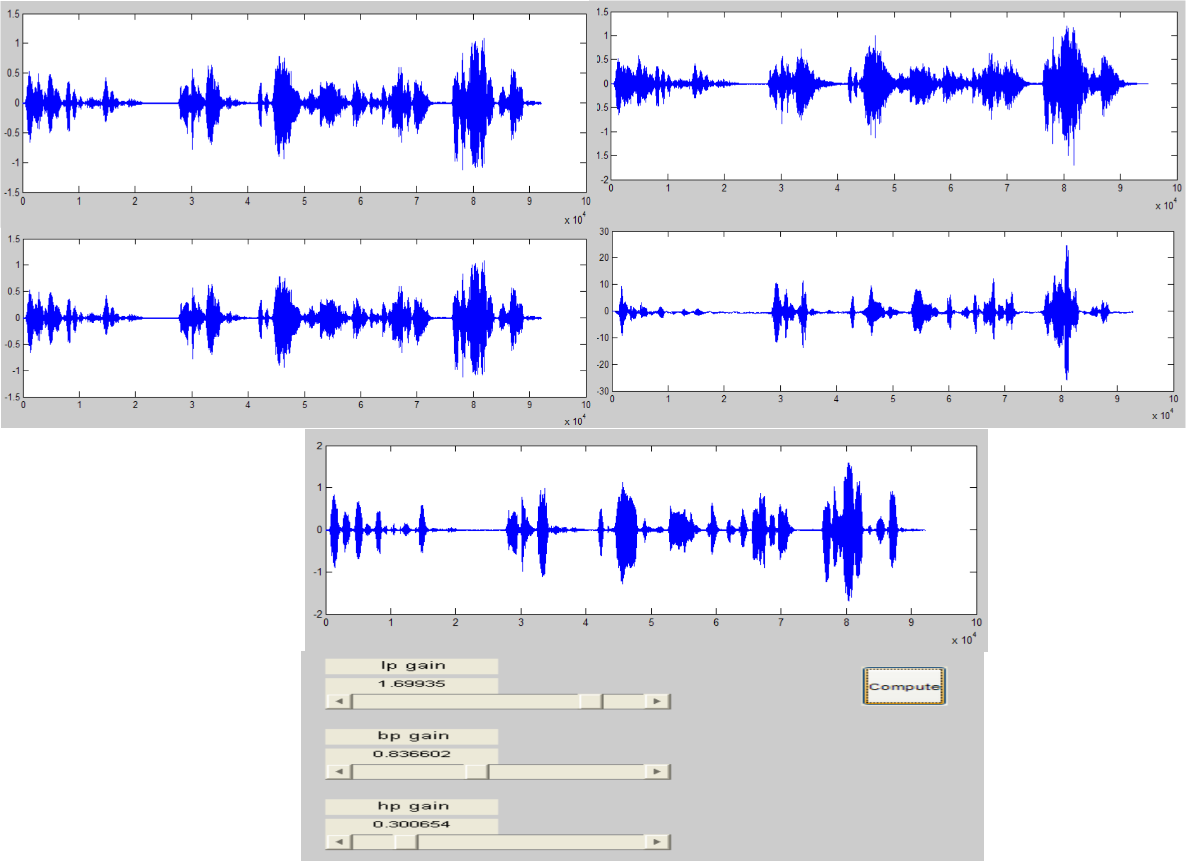
The applications like speech recognition, mobile and personal communication, public address system are few of the applications from long list of speech based systems. However, undesired noises in environment like sound from heavy machines, vehicles are also present in one or other form everywhere. These noises cause undesired effects in speech transmission and acquiring systems. Recently, restricted or usable vicinity of applications is moving from one place and close room to more open and multiple locations, leading to several types of undesired signals of mixing with desired speech signal making speech more corrupt with noise. Not only human communications but intelligent machines which trying to automate the things and sometimes also takes decision based on what it receives as a speech, also suffers from the degraded performance.

Since last five decades, various approaches for noise reduction and speech enhancements have been investigated and developed. Among, very early and fundamental approach of noise reduction was introduced to use the theory of the optimum Wiener filter. Given a desired signal and an input signal, the Wiener filter produces an estimate of the desired signal that is optimal, i.e. the squared mean error or difference between the signals is minimized. The Wiener filter can also be adaptively estimated used in an environment where the surrounding noise has time-varying characteristics. Adaptive algorithms such as Least Mean Square (LMS) and Recursive Least Squares (RLS) are well known examples and also widely used. There are so many applications of speech still to be far from reality just because of lack of efficient and reliable noise removal mechanism and preserving or improving the intelligibility for the speech signals [1-2]. In audio signal processing applications auditory system parameters[2] like echo, multiple echo, reverberation, flanging and equalizer reduces the overall computational complexity and memory requirement of the system. Before doing speech enhancement we need to generate and analyze these parameters [3]. An acoustic *echo* is one of the simplest acoustic modeling problems. Echoes occur when a sound arrives via more than one acoustic propagation path. Reverberation is the persistence of sound in a particular space after the original sound is removed. A reverberation, or reverb, is created when a sound is produced in an enclosed space causing a large number of echoes to build up and then slowly decay as the sound is absorbed by the walls and air [4]. The "flange" effect originated when an engineer would literally put a finger on the flange, or rim of one of the tape reels so that the machine was slowed down, slipping out of sync by tiny degrees. Equalization is the process of adjusting the strength of certain frequencies within a signal. The most well known use of equalization is in sound recording and reproduction [5]. Various audio effects were simulated in MATLAB. The implementation of effects performed using digital signal processing components. To get the simulated results, we have taken a sample wav file. Using this input sample file various audio effects were simulated. Figure 1.2 shows the sample input wav file, response of echo, reverberation, flanger and equalizer effects. Figure1.3 shows responses of LP, HP and BP filters for equalizer.

Recent advances in CPU and multi-core hardware has provided ample amount of computational power and thus, need for today is to design the complex but yet efficient and realistic approach for noise reduction to achieve speech enhancement. The speech enhancement is not only useful for storage and transmission of speech data but it can play vital role in improving much need system based speech recognition where accurate identification of words and sentences can provide automation in most of the human-machine based interface and also be useful in machine-machine interaction based automation. Robotics is a familiar example where speech recognition systems can become boon for today's advanced society at social level in addition to during natural calamities and on war fields.

It is obvious that speech enhancement can boost up the performance of speech recognition systems by keeping low word error rate (WER). Types and sources of noise that can be considered in speech enhancements are also discussed in further section.

[1] Paper presented on "**Digital Signal Processing based Implementation of Auditory System Parameters**", under International Society of Science and Technology, Mumbai in *National Conference on Emerging Technologies and Applications in Engineering and Science (NCETAES)* organized by Saraswati College of Engineering, Kharghar, Navi Mumbai, Pages 44-49, ISSN: 0974-0678, February, 3-4, 2011.



lp gain
1.69935

bp gain
0.836602

hp gain
0.300654

Fig. 1.2 The sample input wav file, response of echo, reverberation, flanger and equalizer effects

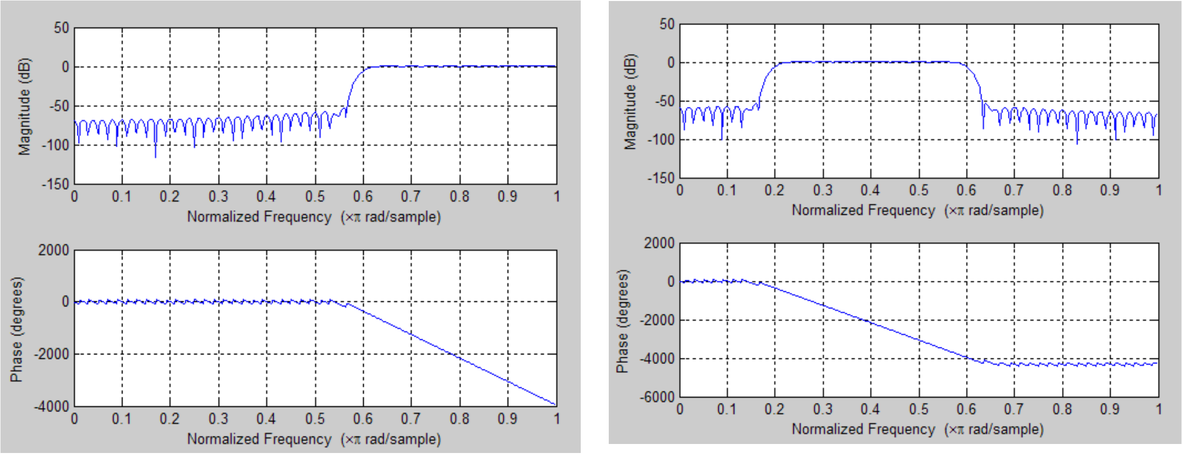


Fig. 1.3 Responses of LP, HP and BP filters

There are various types of advanced speech enhancement algorithms in literature and they can be classified in main three categories, namely; filtering/estimation based noise reduction, beam forming and active noise cancellation (ANC) techniques. Detailed knowhow of these techniques can aid the research in speech enhancements. In this thesis, we have attempted towards surveying the methodologies for speech improvement. It is also investigates, how these techniques affect the performance of various application systems like speech recognition and speech communication.

Speech is the fundamental way for “we humans” to communicate. This way of expressing oneself is probably one of the most effective and reliable means for personal communication. For centuries, efforts have been made to enable individuals to communicate over great distances, distances that render normal face-to-face speech communication impossible. The invention of the radio telegraph and the telephone in the nineteenth century was a great leap forward in the direction of seamless personal communication between persons on the geological locations.

At the same time the industrial revolution introduced new difficulties for personal communication in the form of high sound pressure levels from vehicles and other kinds of machine. Today, we live in a world where silence is a rarity and noise is almost constantly present. This noise sometimes impairs our ability to communicate reliably regardless of what communication means we choose. Not only human communications but intelligence machines which trying to automate the things and sometimes also takes decision based on what it receives as a speech, also suffers from the degraded performance. During the last decades, since the 1940’s, different approaches to noise reduction and speech enhancements have been developed. One early and fundamental method of noise reduction was to use the theory of the optimum Wiener filter. Given a desired signal and an input signal, the Wiener filter produces an estimate of the desired signal that is optimal, i.e. the squared mean error or difference between the signals is minimized. The Wiener filter can also be adaptively estimated used in an environment where the surrounding noise has time-varying characteristics. Adaptive algorithms such as Least Mean Square (LMS) and Recursive Least Squares (RLS) are well known examples and also widely used.

We can today retrospect notice a development from the earlier analogue techniques into digital techniques used since 1980’s. Recent advances in computers have rendered it possible to implement rather complex signal processing algorithm in real-time on digital processors. These processors are capable of performing rapid additions and multiplication which are two fundamental arithmetic operations used when filtering a digital signal. There are some well researched techniques and well studied situations. Like, frequency domain techniques, voice activity detection (VAD), multi-microphone techniques (Beam-forming), noise level estimation (SNR estimation) and active noise cancellations (ANC) techniques are few of many popular terms in speech enhancements.

Detailed knowhow of these techniques can aid the research in speech enhancements. Our research work aims at developing effective speech enhancement techniques particularly to improve the performance of applications like speech recognition or personal communication via mobile or blue-tooth.

1.1 Scope of Research

Speech is medium of communication to express message of the speaker. Along with the message of the speaker other information like language, dialect, gender and age of the speaker are embedded in the speech signal. Listener can perceive this information along with the message in the speech. In fact, human ears are capable of decoding this supplementary information in order to be more informed about speaker. Due to automation and law enforcement needs, there are so many applications derived from the field of speech processing like speech recognition system, gender classification based on the speech of speaker etc. However, most of the times, the speech signal is affected by interference from various sources of noise and consequently, speech processing in various application takes place with degraded speech, which may result in poor performance of the developed system.

In general, there exists a need for digital voice communications, human-machine interfaces, and automatic speech recognition systems to perform reliably in noisy environments. For example, in hands-free operation of cellular phones in vehicles, the speech signal to be transmitted may be contaminated by reverberation and background noise. In many cases, these systems work well in nearly noise-free conditions, but their performance deteriorates rapidly in noisy conditions. Therefore, development of pre-processing algorithms for speech enhancement is always of interest. The goal of speech enhancement varies according to specific applications, such as to boost the overall speech quality, to increase intelligibility, and to improve the performance of voice communication devices.

Enhancement of speech is useful in many applications like aircraft, mobile, military and commercial communications. An enhancement of speech is also useful to improve perceived speech for hearing impaired persons or to improve the speech of speaker with defective speech production process. In all these applications the end users are human beings. Apart from these, there are other applications which involves enhancement as a pre-processing step for other speech processing tasks such as speaker recognition.

1.2 Problem Formulation

Speech enhancements involve processing of speech signals in temporal and/or spectral domains. Any such processing introduces distortion into speech signal. Trade-off between the reduction of noise and introduction of new distortion depends on the perception by the human auditory system. Enhancement in the processes signal is measured in terms of quality and intelligibility. Quality refers to naturalness and case of listening to speech, whereas intelligibility refers to ease of understanding speech. In general, high quality speech signal can be considered as highly intelligible, but highly intelligible speech signal need not be of high quality.

In order to have effective speech enhancement techniques applicable for wide range of applications, following are the research objectives of our work:

- Literature survey for existing techniques and modifications suggested by various researchers in present application scenario.
- To study the effect of various auditory parameters, this reduces computational complexity and memory requirement of digital processor system.
- To understand the effect of noise on various applications based on speech processing like speech recognition, speaker identification, gender classification, personal and mobile communication etc.
- To identify the specific noise based issues and their severity for different speech processing applications.
- To explore and implement the various noise removal and speech enhancement techniques.
- To analyze the effectiveness and usefulness of speech enhancement techniques in one or more than one speech processing applications.
- To study the effect of various noise levels on the speech processing applications and formulating the new SNR estimation strategy in order to improve the performance of particular speech enhancement technique.
- To conduct the experiment based on dataset available publicly or self created dataset to evaluate the performance of speech enhancement techniques.
- To embed the speech enhancement technique in one or more applications of speech processing and observe the change in performance of the system using dataset available publicly or self created dataset.

- Real time and hardware implementation of speech recognition technique using MATLAB and CCS V3.1 on DSK 6713 from Spectrum Digital Corporation.
- Hardware profiling of technique considering it as real time speech processing for multimedia applications.

1.3 Thesis Contribution

In this research work, the problem of speech enhancement to be used in multimedia application like speech recognition has been considered. In order to improve the performance of speech recognition using specialized speech enhancement technique. In the part of our work, we proceed with two-fold objectives. First is to improve the speech recognition performance in multi-microphone environment. Second, we attempted to analyze the performance of speech recognition against the filter-bank parameters; filter length and number of sub bands. In the remaining part of the research work, we have improved the performance of beamforming based speech recognition system using evolutionary computational algorithms (Genetic algorithm, GA). Additionally, the system is made to be working in real-time as time required for classifier has been reduced dramatically. This is particularly achieved by including the zeros at random places and in random amount in initial population chromosomes, which were generated randomly in the range of 0 to 1. This results in the reduction of feature elements in feature descriptor and have feature vector length.

1.4 Limitations and remedial action during research work

This research work involves development of new strategies based on soft computing for real time speech processing in multimedia application. We will be using MATLAB software for speech processing. The speech signals used in our work will be either dataset available at reputed research groups working in same field or new dataset designed by us. Validation of results will be with either annotated dataset or manual annotation using ground truth. Despite of the pre-structured work planned, there could be limitations as described below. In case of some limitation occurred, the corrective action is taken as indicated in table 1.1.

Table 1.1 Limitations and remedial action

Limitation	Remedial Action
The speech signal dataset available is not capable of serving the purpose of our research work.	Construction of new data set or changing the objective.
Non-availability of methods to annotate the data.	Finding method to know approximated ground truth subjectively.
The new methodology to be explored is not giving satisfactory performance.	Change the methodology or changing the objective.
Software constraints or unavailability.	Choosing the new software or developing the methodology with basic software such as C/C++ or JAVA or using speech processing libraries.
Lack of validation measures.	Coming up with new measures or taking experts views.
Non-availability of real speech data from real environment.	Simulation of speech signals or changing the objective.
Non-availability of annotated data.	Manual annotation or changing the objective.
Impossible to implement mathematical process.	Approximating the mathematical process.

1.5 Outline of Thesis

The thesis is organized in the form of eleven chapters as follows:

Chapter-1 Introduction

The “speech signal” is an integral part of most of the multimedia applications apart from the personal communication. Here we wish to produce and analyze a variety of effects, viz. echo, multiple echo, reverberation, flanging and equalizer on pre-recorded speech, which will be used in sound processor to generate various musical effects. The broad objective of this thesis is to devise new strategies using soft-computing techniques so that performance of speech based application can be improved. In this chapter, introduction of speech recognition and enhancement is presented. How speech is influencing lifestyle of citizens and what kind of research is required for speech related applications are briefly discusses here. The brief note on scope of this topic is placed in this section. Apart from scope and applications, the specific problems dealt in this thesis work are mentioned. More precise contributions of this thesis are also highlighted here. It is interesting to foresee the effect of this research in future speech related applications especially in multimedia area.

Chapter-2 Basic Concepts of Speech Enhancement and Recognition

The effect of speech purity is visible at the performance of any speech based applications. The speech recognition, being a multimedia application under consideration here, has been a very important system in almost every area of life. Here an attempt has been made towards studying and implementation of speech enhancement techniques like Spectral Subtraction, Minimum Mean Square Error (MMSE), Kalman, Wavelet Transform, Wiener and Adaptive Wiener filter. The intelligent speech enhancement techniques can raise the outcome of speech recognition and hence it is very important to know the basics involved in it. This chapter is devoted to know the fundamental concepts behind the various classes of speech image enhancements techniques. The short description of speech recognition is also included in this chapter. Finally, brief note on various performance measures in speech enhancements and recognition has been incorporated.

Chapter-3 Literature Review

The speech, being a fundamental way of communication for the humans, has been embedded in various essential applications like speech recognition, voice-distance-talk and other forms of personal communications. There are so many applications of speech still to be far from reality just because of lack of efficient and reliable noise removal mechanism and preserving or improving the intelligibility for the speech signals. The broad categories of speech enhancement techniques can be listed as speech filtering techniques, beam forming techniques and active noise cancellation methods. In this chapter, an attempt has been stepped towards surveying the methodologies for speech improvement. It was also interesting to discuss, how these techniques affect the performance of various application systems like speech recognition and speech communication. Essentially, we also discussed here about the types and sources of noise that can be considered in speech enhancements.

Chapter-4 Soft-computing: Evolutionary Computations

The continuing advances of computational technology such as availability of large memories in small space, parallel GPUs, have changed the paradigm of the way the problem used to be solved. The soft-computing is the new paradigm for the computationally solvable complex problems and has been heavily relied on the computational power of devices. It includes the probabilistic theory in addition to the elements covered by computational intelligence. In this chapter, we have given brief description of main elements of the soft-computing, such as neural network, fuzzy logic and genetic algorithm.

Chapter-5 Speech Enhancement and Beamforming

A new generation of speech acquisition applications is emerging as a result of advances in technology and the prevalence of mobile and broadband communication. Thus, it becomes essential to have reliable speech processing based applications. The speech is corrupted with so many different types of noises and by cross voices. This presents the need of cleaning out the speech so that applications can perform without any flaws. In this chapter, we have described in detail about the speech enhancement theory and especially with beamforming techniques.

Chapter-6 Experimental Setup and Dataset

While experimenting with speech enhancement using beamforming technique and later for the speech recognition experiments, there is a need of dataset with ground truth. There are not many datasets available in public across the world. It was necessary to construct the dataset using beamforming parameters. In doing so, we have attempted to do the simulation of speech database to be used for speech recognition experiments with beamforming parameters. In this chapter, the detail of this simulation has been provided about this dataset.

Chapter: 7 Speech Recognition using Beamforming technique

In this chapter, our work has two-fold objective. First is to improve the speech recognition performance in multi-microphone environment. Second, we attempted to analyse the performance of speech recognition against the filter-bank parameters; filter length and number of subbands. The experiments were performed for 20 words including numbers and commands, 10 words of numbers only and 10 words of commands only for different values of filter bank parameters. The results obtained have proved the speech enhancing capability of the beamforming technique in multi-microphone network where noise and echo-interference can degrade the original speech signal.

Chapter-8 Evolutionary Computation based Real Time Speech Beamforming for Multimedia Applications

Here we have presented the approach of evolutionary computation in form of genetic algorithm to select the features that are responsible for discriminating the different words. The system is made to be working in real-time as time required for classifier has been reduced dramatically. This is especially an important requirement in the mobile devices where power, memory and processing power are available with large constraints. The experiments were performed for 20 words including numbers and commands, 10 words of numbers only and 10 words of commands only for different values of filter bank parameters. The results show the effectiveness of the GA optimization in all the subsets of experiments with different parameters of beamforming.

Chapter-9 Real-time implementation of speech recognition

In this chapter we depicted the real time hardware implementation of speech recognition using DSP processor software development kit, DSK-TMS320C6713 with Code Composer Studio (CCS). MFCC algorithm calculates cepstral coefficients of Mel frequency scale. After feature extraction from recorded speech, each Euclidian Distance (ED) from all training vectors is calculated using Gaussian Mixture Model (GMM). The command/voice having minimum ED is applied as similarity criteria. The timing analysis is done for various individual blocks of algorithm. The time required for processing in DSP and PC processors are compared.

Chapter-10 Conclusions and future scopes

In this chapter final conclusions, future extension of the work and future scope in this field are elaborated.

Chapter-11 References

Thesis ends with Bibliography which includes the list of references used in each chapter, research project details, list of short term program attended, list of publications and presentations, additional resources used and list of MATLAB programs simulated for research work.

1.6 Conclusion

The “speech signal” is an integral part of most for the multimedia applications apart from the personal communication. The broad objective of this thesis is to devise new strategies using soft-computing techniques so that performance of speech based application can be improved. In this chapter, introduction of speech recognition and enhancement is presented. How speech is influencing lifestyle of citizens and what kind of research is required for speech related applications are briefly discusses here. The brief note on scope of this topic is placed in this section. It is interesting to foresee the effect of this research in future speech related applications especially in multimedia area. Apart from scope and applications, the specific problems dealt in this thesis work are mentioned. More precise contributions and outline of this thesis are also highlighted here.

CHAPTER 2

BASIC CONCEPTS: SPEECH ENHANCEMENT AND RECOGNITION

2.1 Introduction

With the rapid advancements in industrial and technology applications, the demand from consumer for transmission and manipulation of data, primarily speech, audio and images, are at its peak. The speech, being a fundamental way of communication for the humans, has been embedded in various essential applications like speech recognition, voice-distance-talk and other forms of personal communications. There are so many applications of speech still to be far from reality just because of lack of efficient and reliable noise removal mechanism and preserving or improving the intelligibility for the speech signals. The broad categories of speech enhancement techniques can be listed as speech filtering techniques, beam forming techniques and active noise cancellation methods. In this thesis, an attempt has been stepped towards surveying the methodologies for speech improvement. It is also interesting to investigate, how these techniques affect the performance of various application systems like speech recognition and speech communication. Essentially, we also discuss about the types and sources of noise that can be considered in speech enhancements.

2.2 Speech

Traditionally, the speech signals (sound) of spoken words have been studied with two different perspectives: (1) Phonetic components of spoken words, e.g., vowel and consonant sounds, and (2) Acoustic wave component. A language can be broken down into a very small number of basic sounds, called phonemes. An acoustic wave is a sequence of changing vibration patterns (generally in air), however we humans are more accustom to "seeing" acoustic waves as their electrical analog on an oscilloscope (time presentation) or spectrum analyzer (frequency presentation). During speech analysis, signals are decomposed and represented on time-frequency axis, also called as spectrograms. Spectrogram has two axes, one displays frequency contents in speech at particular instance along the vertical axis and another shows the time variation of each frequency components on horizontal axis. The intensity (amplitudes) of particular frequency at particular instant is represented by figure intensity or color. The human speech is generated as air from the lungs passes through the larynx producing perturbations (vibrations) in the vocal cords and/or noise in any regions of oral or nasal cavity as shown in figure 2.1. The consonant part of speech is generated because of restriction brought to the flow of air through vocal cord. The shape of passages across the vocal cord is modified to travel the airflow to form vowels [CH-1(1-3)].

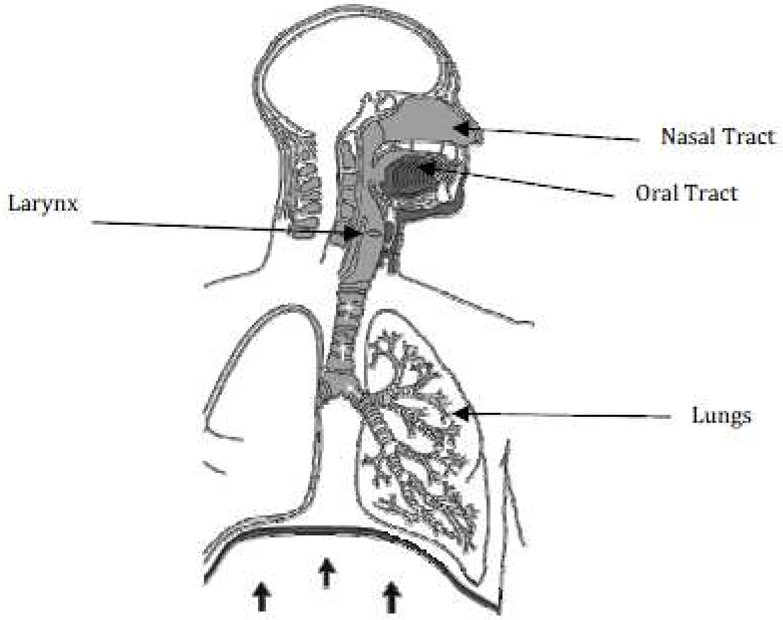


Fig. 2.1 Human Articulatory System

2.3 Speech Enhancement

The aim of speech enhancement is to improve the quality and intelligibility of degraded speech signal. Improving quality and intelligibility reduces exhaustion of listener. It is difficult to measure intelligibility by mean of mathematical algorithm, while we can measure quality of speech signal by the term signal distortion [1]. Intelligibility of speech signal greatly affected by background noise, it will warp clean speech. So to improve intelligibility attenuation of noise is required to enhance speech signal. Figure 2.2 shows basic block diagram of speech enhancement.

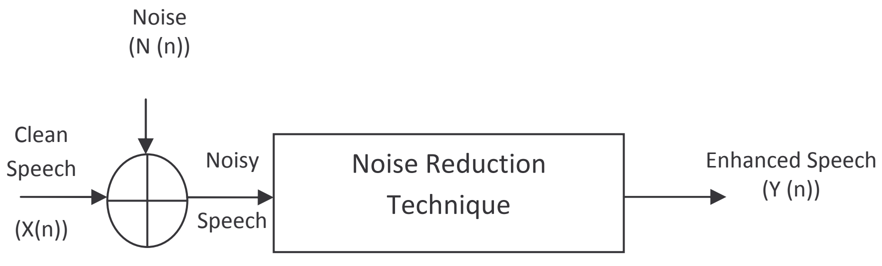


Fig. 2.2 Basic block diagram of Speech Enhancement

Here an attempt has been made towards surveying the methodologies for speech enhancement. We will be analyzing various methods of speech enhancement such as Kalman filter, Wiener filter, Spectral Subtraction method and Minimum Mean Square Error (MMSE). The Spectral Subtraction method is the most widely used due to the simplicity of implementation As given in [2], Spectral subtraction is a method for restoration of the power spectrum or the magnitude spectrum of a signal observed in additive noise, through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. It reduces the effect of background noise based on the STSA estimation technique. Ephraim, Y.[3] formulated MMSE estimation approach for enhancing speech signals degraded by statistically independent additive noise is developed, based upon Gaussian autoregressive (AR) hidden Markov modeling of the clean signal and Gaussian AR modeling of the noise process. The parameters of the models for the two processes are estimated from training sequences of clean speech and noise samples. Paper [4] explains low complexity wiener filtering statistical approach to filter out noise that has corrupted a signal. In this typical filters are designed for a desired frequency response based on knowledge of the spectral properties of the original signal and the noise. Paper [5] describes Kalman filtering, in which speech signal is usually modelled as autoregressive (AR) process and represented in the state-space domain. All the Kalman filter-based approaches proposed in the past operate in two steps. They first estimate the noise and the driving variances and parameters of the signal model, then estimate the speech signal. Based on our observations and analysis of performance parameters such as SNR ratio, Mean square error, etc. we conclude which method is the most suitable for speech enhancement. We have implemented the code using Graphic User Interface (GUI) on MATLAB[2] as shown in figure 2.3. Figure 2.4 shows reconstructed speech signal for Wiener, SS, MMSE and Kalman Speech Enhancement methods and figure 2.5 shows comparison of SNR and PSNR for Wiener, SS, MMSE and Kalman methods.

According to specific application, the requirement of speech enhancement technique varies to increase speech quality, intelligibility and performance of speech communication devices. VoIP (Voice over Internet Protocol) played a vital role in communication system. Echo is the problem occurs in VoIP which reduces the quality of speech signal. It is difficult to remove echo completely but it can remove to tolerable range. If we try to remove it completely then it degrades the quality of speech signal on VoIP system [6]. Speech enhancer is required to improve the quality of degraded speech in VoIP system.

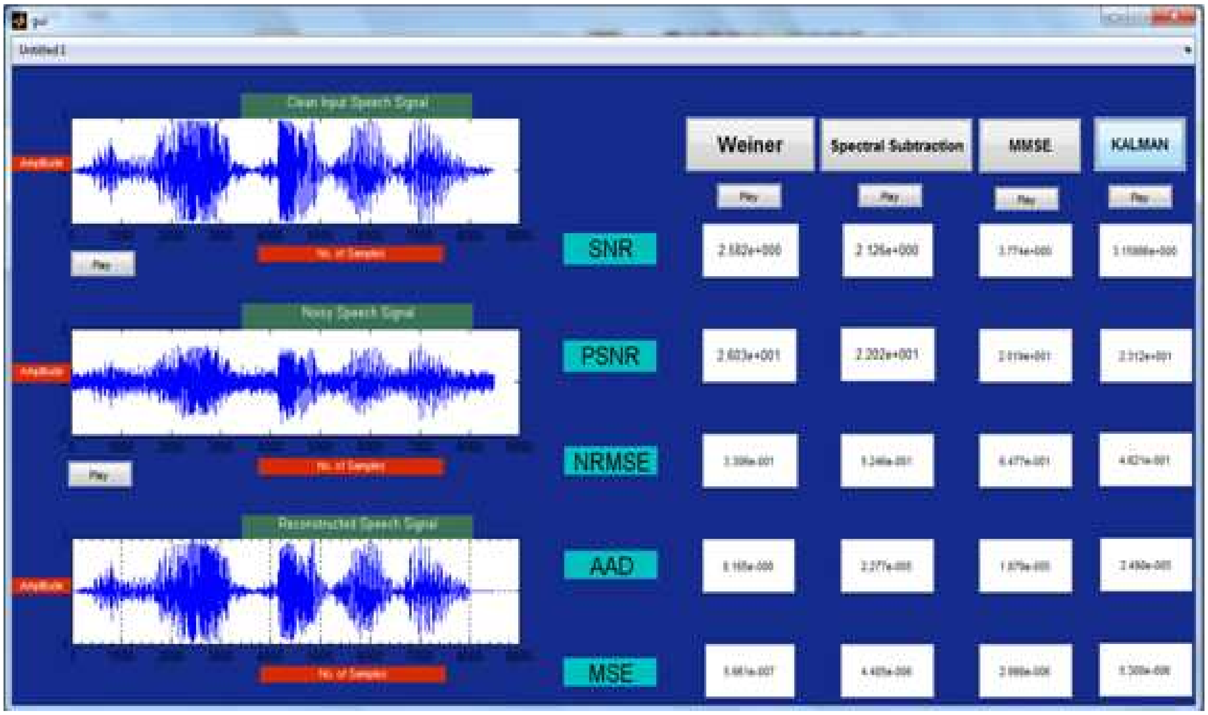
[2] Paper published on "**Performance comparison of Single Channel Speech Enhancement Techniques for Personal Communication**" *International Journal of Innovative Research in Computer and Communication (IJIRCCE),* Paper ID:V1I10C042, ISSN(Online): 2320-9801, ISSN(Print):2320-9798, Pages 67-76, Volume 1, Issue 1, March 2013.

Figure 2.6 shows the required flow of process in VoIP system for echo cancellation and speech enhancement. Spectral subtraction most widely used in single microphone algorithms for speech enhancement, but it produces difficulties in pause detection due to additional relic as musical noise. DONOHA [7] in 1995, presented approach for denoising signal degraded by additive white noise using wavelet thresholding technique. Different steps involved in implementation of speech enhancement using wavelet transform are shown in figure 2.7. Different steps involved in implementation of speech enhancement using wavelet transform are shown in figure 2.8.

Wiener filtering estimates noise free speech signal from that noisy speech signal corrupted by additive noise. Estimation is performed by minimizing the Mean Square Error (MSE) between the noise free signal $x(n)$ and its estimation $\hat{x}$(n). The problem with this method is that it has fixed frequency response at all frequencies and it also required estimation of power spectral densities of noise free and noise signal before filtering. To solve this problem, M.A. Abd E-Fattah [8] presented adaptive wiener filtering approach in 2008. According to this approach enhanced speech signal of small segment stationary noisy signal can be represented as

$$\widehat{x}(\mathrm{n}) = \mathrm{m_x} + (x(n) - \mathrm{m_x}) \text{ x } \frac{Vx}{Vx+Vd} \tag{2.1}$$

Where $\mathrm{m_x}$ is mean of noise free speech signal, Vx and Vd are variance of noise free speech and noise respectively. If Vx is smaller than Vd, input signal x(n) is attenuated due to filtering effect.



	Weiner	Spectral Subtraction	MMSE	KALMAN
SNR	2.582e+000	2.126e+000	3.774e+000	3.15886e+000
PSNR	2.603e+001	2.202e+001	2.019e+001	2.312e+001
NRMSE	3.306e-001	5.246e-001	6.477e-001	4.821e-001
AAD	8.165e-003	2.277e-003	1.879e-003	2.490e-003
MSE	5.661e-007	4.405e-006	2.088e-006	5.305e-006

Fig. 2.3 GUI based performance analysis of Wiener, SS, MMSE, Kalman techniques

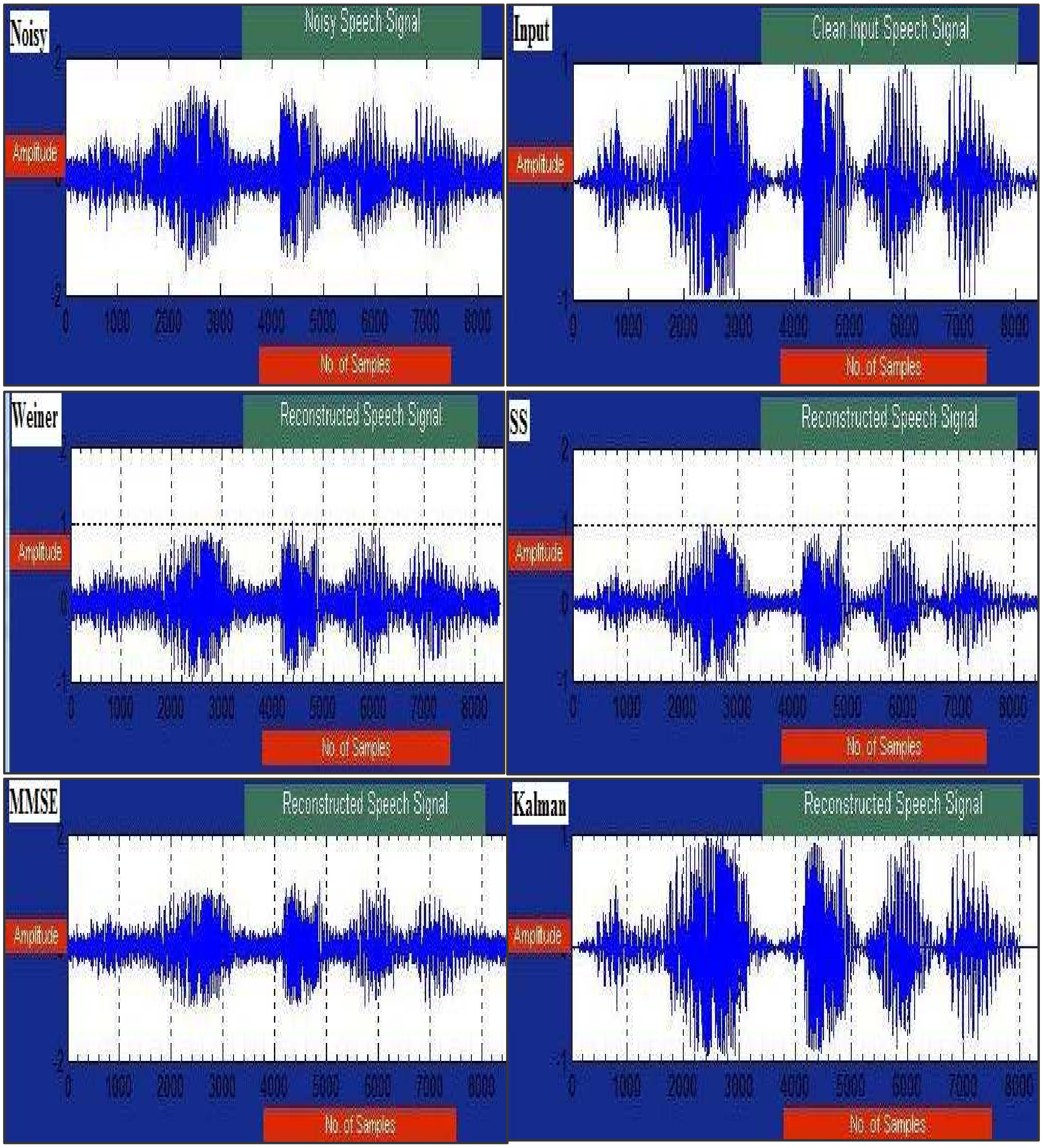


Fig. 2.4 Reconstructed speech signal for Wiener, SS, MMSE and Kalman Speech Enhancement methods

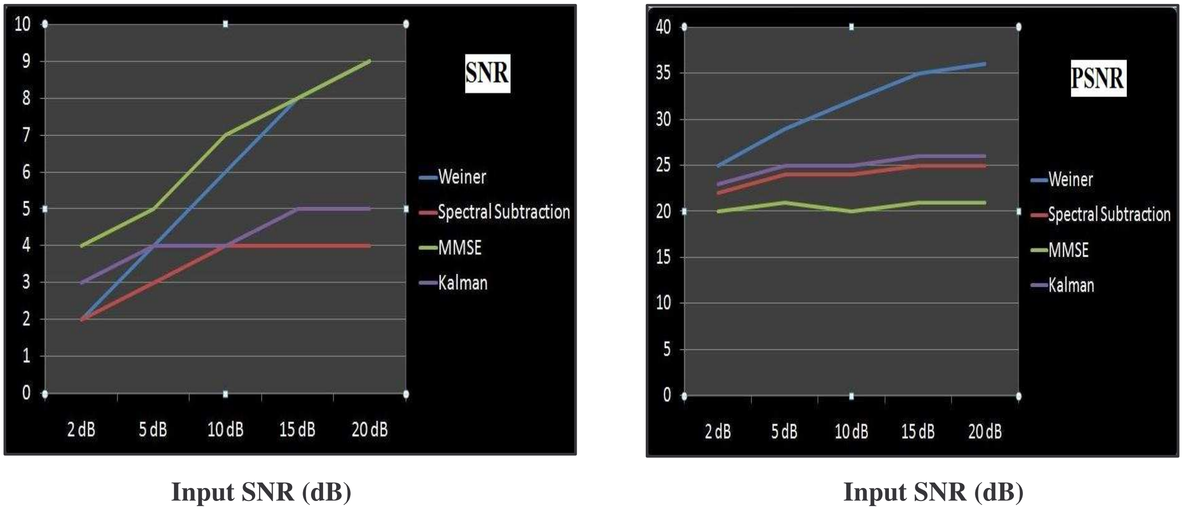


Fig. 2.5 Comparison of SNR and PSNR for Wiener, SS, MMSE and Kalman methods

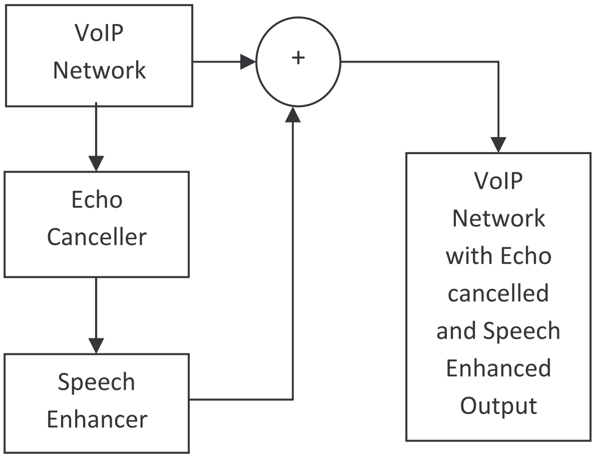


Fig.2.6 VoIP with Echo Canceller and Speech Enhancer

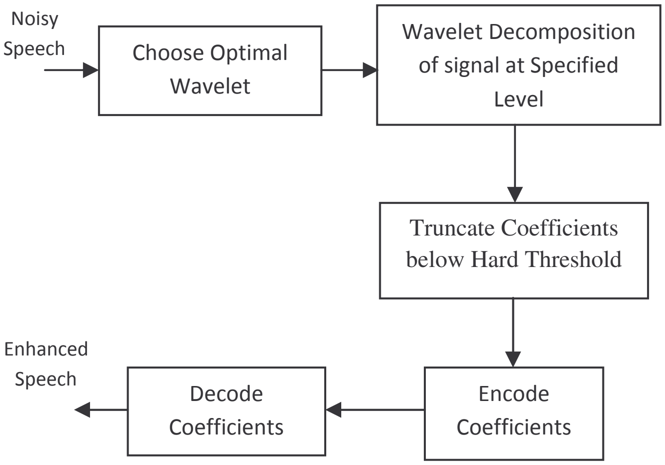

Noisy Speech
Choose Optimal Wavelet
Wavelet Decomposition of signal at Specified Level
Truncate Coefficients below Hard Threshold
Encode Coefficients
Decode Coefficients
Enhanced Speech

Fig.2.7. Wavelet Transform based Speech Enhancement

We have developed code for GUI based quantitative performance comparison of single channel speech enhancement techniques for personal communication in MATLAB. Different steps involved in implementation of speech enhancement using Adaptive Wiener Filtering are shown in figure 2.9. The experimental results that concerned to our single channel speech enhancement systems were compared to Wavelet Transform (WT), Adaptive Wiener Filtering (AWF) and Spectral Subtraction (SS) methods. Test for speech enhancement were performed using uncontaminated recorded “Hello” word, which have 11020 samples, one second length, data size of 22040 bytes and PCM 11.025KHz, 16 bit Mono audio format using sound recorder of PC. This word is then contaminated with white gaussian noise type SNR of 0,-10,-20,-30,-40,-50 and -60dB to show the ability of single channel speech enhancement techniques for improving SNR in noisy speech environment for personal communication. MATLAB GUI developed for speech enhancement techniques which help to be able to visualize the results shown in figure 2.9. Wavelet Transform based speech enhancement technique perform better due to good speech reconstruction quality. Table 2.1 shows Performance Measure of WT, SS and AWF based on Output SNR, PSNR and NRMSE.

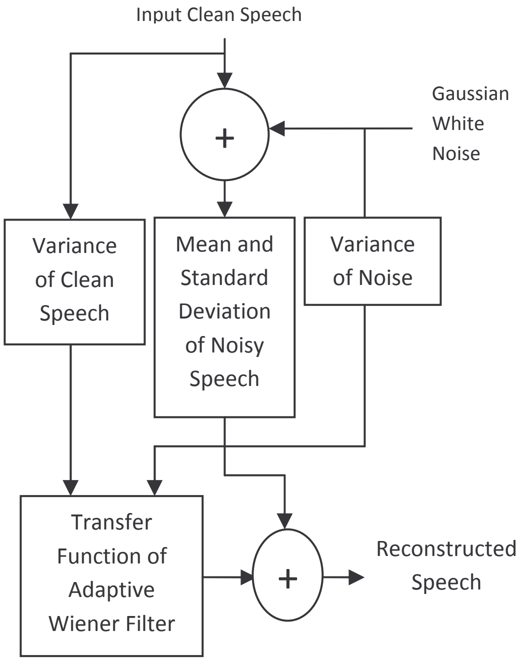
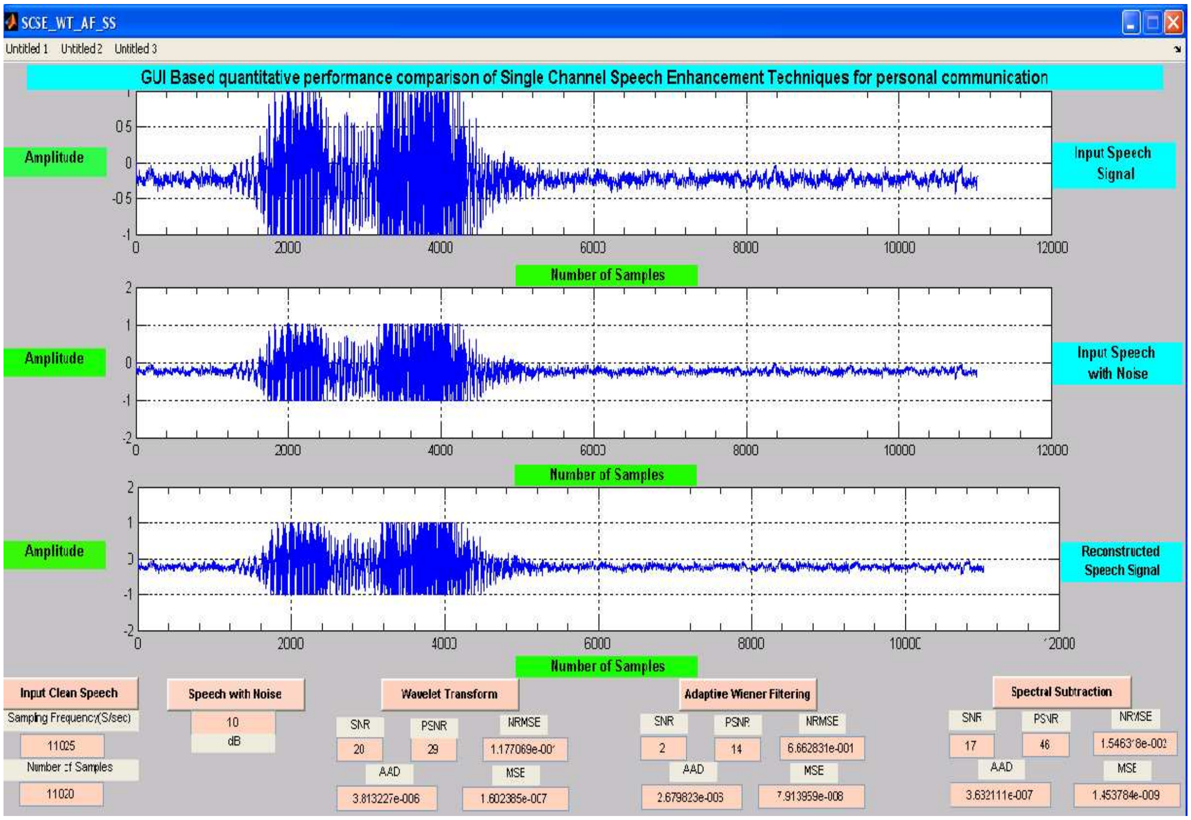


Fig.2.8. Adaptive Wiener Filtering based Speech Enhancement



Input Clean Speech
Sampling Frequency(S/sec): 11025
Number of Samples: 11020

Speech with Noise: 10 dB

Wavelet Transform
SNR: 20
PSNR: 29
NRMSE: 1.177069e-00
AAD: 3.813227e-006
MSE: 1.602385e-007

Adaptive Wiener Filtering
SNR: 2
PSNR: 14
NRMSE: 6.662831e-001
AAD: 2.679823e-006
MSE: 7.913959e-008

Spectral Subtraction
SNR: 17
PSNR: 46
NRMSE: 1.546378e-00
AAD: 3.632111e-007
MSE: 1.453784e-009

Fig.2.9 GUI for Comparison of SS, AWF and WT Speech Enhancement Techniques

Input SNR (dB)	Enhanced SNR (dB)		
	WT	SS	AWF
0	20	15`	2
-10	19	12	2
-20	13	8	2
30	6	3	1
-40	3	0	0
-50	3	0	0
-60	2	0	0

Input SNR (dB)	Output PSNR (dB)		
	WT	SS	AWF
0	29	40`	11
-10	29	32	12
-20	25	25	14
-30	24	23	19
-40	27	25	24
-50	38	35	35
-60	45	43	43

Input SNR (dB)	Output NRMSE		
	WT	SS	AWF
0	1.2E-001	3.2E-002	9.1E-001
-10	1.4E-001	9.4E-002	9.9E-001
-20	2.6E-001	2.7E-001	9.9E-001
-30	5.4E-001	6.7E-001	9.9E-001
-40	7.1E-001	9.2E-001	9.9E-001
-50	7.4E-001	9.7E-001	9.9E-001
-60	7.5E-001	9.8E-001	9.9E-001

Table 2.1 Performance Measure of WT, SS and AWF based on Output SNR, PSNR and NRMSE

2.4 Single Channel Estimation based Enhancement

The simplest form of speech enhancement primitive is the noise reduction from the noisy speech and is applicable for single channel based speech applications. In this type of speech enhancement techniques, algorithms are either / combine based on the model of noisy speech or and perceptual model of speech using masking threshold. The generalized diagram of single channel enhancement technique is shown in figure 2.10.

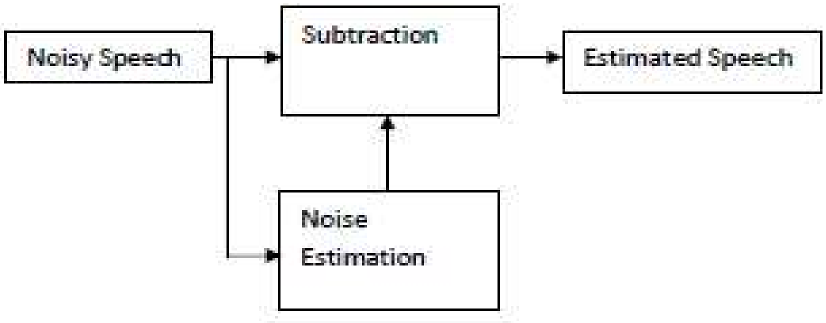


Fig. 2.10 Single channel enhancement technique

2.5 Multichannel Beamforming based Speech Enhancement

Another important class of speech enhancement methods is based on the beam-forming, where more than one speech channels (microphones) are used to process the speech. Speech signals are received simultaneously by all microphones and outputs of these sensors are then processed to estimate the clean speech signal. In adaptive beamforming, an array of antennas is exploited to achieve maximum reception in a specified direction by estimating the signal arrival from a desired direction (in the presence of noise) while signals of the same frequency from other directions are rejected. This is achieved by varying the weights of each of the sensors (antennas) used in the array. This kind of speech enhancement techniques can give better performance of the speech applications like automatic speech recognition (ASR) than single channel processing. Only disadvantage with this class of methods is higher cost of hardware, which can put restriction on using these methods in some speech applications. The basic block diagram of beamformer is shown in figure 2.11.

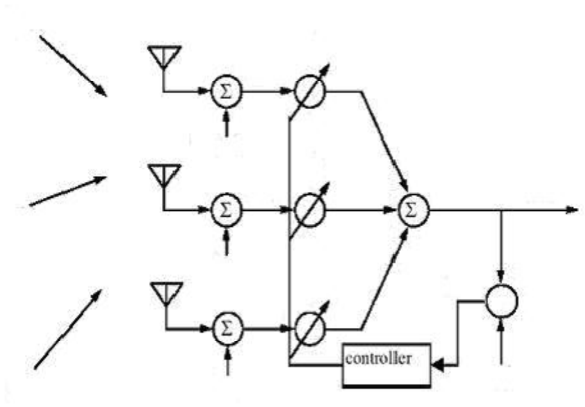


Fig. 2.11 Beamformer: An Adaptive array system

2.6 Basic Active Noise Cancellation

In Active Noise Cancellation (ANC) techniques, another source of noise is used to cancel out the existing noise present in the speech. The basic principle on which ANC works is when two sinusoidal signals of same frequency and equal in amplitude but out of phase (180 degree) are added, resultant signal yields no zero output and is shown in figure 2.12.

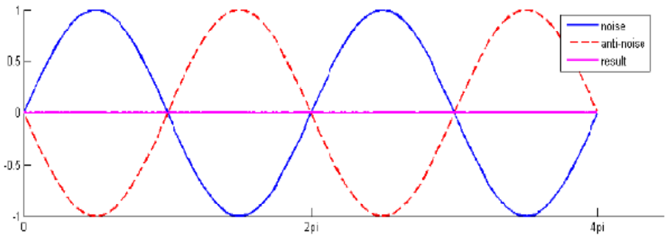


Fig. 2.12 A signal gets nullified by its "out phase signal"

The basic block diagram of ANC is shown in figure 2.13.

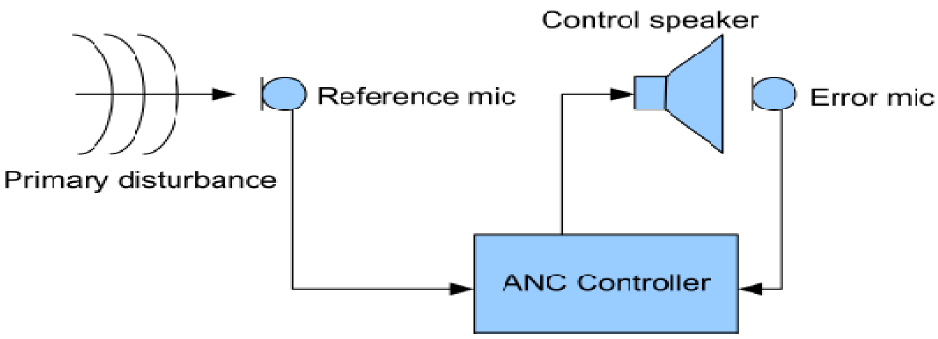


Fig. 2.13 Basic ANC system

2.7 Speech Recognition

There are various objectives for the development of automatic speech recognition (ASR). The application can be aimed at recognition to be performed either on isolated words or utterances or on continuous speech. There are various languages spoken in this world that makes to consider the one of the language for the recognition system. There are also situations, when recognition system should be speaker dependent or independent. The most difficult class of recognition system is to develop speaker independent recognition on continuous speech. This needs the inclusion of knowledge about the application for which system to be built in addition to the word recognition system. Typically, the first step in this kind of system is always word recognition for the limited number of words.

2.8 Performance Measures

The performance of any application depends heavily on the extent of purity present speech signal that is to be processed. Though, there are subjective ways to measure the amount of noise corruption, it is not always possible to accomplish it while application is on. Further, in large data based application, there would be constraint on the subjective measurement to be done manually. The objective measures are unbiased and can be determined online processing. Objective measures rely on mathematically based equations which include parameters of original signal and/or degrades speech signal. The suitability of these measures depends on the capability of correlation between measured values and subjective quality of the signal. The following are some of the objective measures for determining the quality of the speech.

A. Itakura-Saito Distortion Measure:

For an original clean frame of speech with linear prediction (LP) coefficient vector $\vec{a}_{\emptyset}$ and processed speech coefficient vector, $\vec{a}_d$, the Itakura-Saito distortion measure is calculated by equation 2.2.

$$\mathrm{d_{IS}}(\vec{\mathrm{a}}_{\mathrm{d}}, \vec{\mathrm{a}}_{\emptyset}) = \left[\frac{\sigma_{\emptyset}^2}{\sigma_{\mathrm{d}}^2}\right]\left[\frac{\vec{\mathrm{a}}_{\mathrm{d}} \mathrm{R}_{\emptyset} \vec{\mathrm{a}}_{\mathrm{d}}{}^{\mathrm{T}}}{\vec{\mathrm{a}}_{\emptyset} \mathrm{R}_{\emptyset} \vec{\mathrm{a}}_{\emptyset}{}^{\mathrm{T}}}\right] + \log\left(\frac{\sigma_{\mathrm{d}}^2}{\sigma_{\emptyset}^2}\right) - 1 \qquad (2.2)$$

Where σ_d^2 and $\sigma_{\emptyset}^2$ represent the all-pole gains for the processed and clean speech frame respectively.

B. Log-Likelihood Ration Measure:

The LLR measure is given in equation 2.3

$$d_{LLR}(\vec{a}_d, \vec{a}_\emptyset) = \log\left(\frac{\vec{a}_d R_\emptyset \vec{a}_d^{\,T}}{\vec{a}_\emptyset R_\emptyset \vec{a}_\emptyset^{\,T}}\right) \tag{2.3}$$

C. Log-Area-Ratio Measure:

The LAR is calculated from the dissimilarity of LP coefficients. The LAR parameters are calculated from the Pth order LP reflection coefficients for the original r_Φ (j) and processed r_d (j) signals for frame j. It is given by equation 2.4

$$d_{LAR} = \left|\frac{1}{M}\sum_{i=1}^{M}\left[\log\frac{1+r_\emptyset(j)}{1-r_\emptyset(j)} - \log\frac{1+\hat{r}_d(j)}{1-\hat{r}_d(j)}\right]^2\right|^{\frac{1}{2}} \tag{2.4}$$

D. Segmented SNR Measure:

It is formed by averaging frame level SNR estimates as follows,

$$d_{SEGSNR} = \frac{10}{M}\sum_{m=o}^{M-1}\log\frac{\sum_{n=Nm}^{Nm+N-1} S_\emptyset^2(n)}{\sum_{n=Nm}^{Nm+N-1}[S_d(n)-S_\emptyset(n)]^2} \tag{2.5}$$

E. Weighted Spectral Slope Measure:

Weighted spectral slope measure per frame is calculated by equation 2.6

$$d_{WSS}(j) = K_{spl}\left(K-\widehat{K}\right) + \sum_{k=1}^{36}\omega_a(k)\left(S(k)-\hat{S}(k)\right)^2 \tag{2.6}$$

Where K, $\widehat{K}$ are related to overall sound pressure level of the original and enhanced utterances and K_{spl} is a parameter which can be varied to increase overall performance.

The objective comparison of speech enhancements techniques is carried by evaluating performance of parameters such as, Mean Square Error (MSE), Normalized Mean Square Error (NRMSE), Signal to Noise Ratio (SNR), Peak Signal to Noise Ratio (PSNR) and Average Absolute

Distortion (AAD). It is based on mathematical comparison of the original and processed speech signals.

A. Signal to Noise Ratio (SNR):

It is most widely used and popular method to measure the quality of speech. It is ratio of signal to noise power in decibels. It is calculated by equation 2.7

$$\text{SNR (dB)} = 10\log_{10}(\frac{\sigma_x^2}{\sigma_d^2}) \quad (2.7)$$

Where σ_x^2 the mean square of speech signal and σ_d^2 is the mean square difference between the original and reconstructed speech.

B. Peak Signal to Noise Ratio (PSNR)

$$\text{PSNR (dB)} = 10\log_{10}(\frac{NX^2}{\|x-r\|^2}) \quad (2.8)$$

Where N is the length of the reconstructed signal, X is the maximum absolute square value of signal ‘x’ and $\|x-r\|^2$ is the energy of the difference between the original and reconstructed signal.

C. Normalized Root Mean Square Error (NRMSE)

$$\text{NRMSE}= \sqrt{\frac{[X(n)-r(n)]^2}{[x(n)-\mu x(n)]^2}} \quad (2.9)$$

Here $X(n)$ is input speech signal and $r(n)$ reconstructed speech signal.

D. Mean Square Error (MSE) and Average Absolute Distortion (AAD)

Mean square error is calculated by equation 2.10

$$\text{MSE}=\frac{1}{N}[(r(n)-x(n))^2] \quad (2.10)$$

Average absolute distortion is calculated by equation 2.11

$$\text{AAD}=\frac{1}{N}\left|(r(n)-x(n)\right| \quad (2.11)$$

Where N is length of input speech signal, x(n) is input speech signal and r(n) is reconstructed speech signal.

2.9 Conclusion

The effect of speech purity is visible at the performance of any speech based applications. The speech recognition, being a multimedia application under consideration here, has been a very important system in almost every area of life. The intelligent speech enhancement techniques can raise the outcome of speech recognition and hence it is very important to know the basics involved in it. This chapter is devoted to know the fundamental concepts behind the various classes of speech image enhancements techniques. The short description of speech recognition is also included in this chapter. Finally, brief note on various performance measures in speech enhancements and recognition has been incorporated.

CHAPTER 3

LITERATURE REVIEW

3.1 Introduction

In general, there exists a need for voice based communications, human-machine/ machine-machine interfaces, and automatic speech recognition systems to increase the reliably of these systems in noisy environments. In many cases, these systems work well in nearly noise-free conditions, but their performance deteriorates rapidly in noisy conditions. Therefore, improvement in existing pre-processing algorithms or introducing entire new class for algorithm for speech enhancement is always the objective of research community. The main requirement for speech enhancement systems varies according to specific applications, such as to boost the overall speech quality, to increase intelligibility, and to improve the performance of voice communication devices.

3.2 Estimation based Filtering Techniques

One of the early papers [1] in speech enhancement considers the problem of estimation of speech parameters from the speech, which has been degraded by additive background noise. In this work they propose the two suboptimal procedures which have linear iterative implementations in order to suppress the non-linear effect on the speech parameters due to background noise. In another similar problem [2] of enhancing the speech in presence of additive acoustic noise, spectral decomposition of frame of noisy speech was adopted. The attenuation of particular spectral component was determined based o n how much the measured speech plus noise power exceeds an estimation of background noise leading an importance of proper choice of the suppression or subtraction factors. The short-time spectral amplitude (STSA) was used to model the speech and noise spectral components in [3]. The parametric estimation techniques, where parameters of underlying model, consist of small set of parameters, is determined and then numerical process is used to modify the parameters, can be contrasted by the non-parametric method which can be used as in [4] where no model is assumed and uses non-parametric spectrum estimation techniques.

In application point of view, there is work described in [5], where noisy speech enhancement algorithm has been discussed and implemented to compare its performance against the various levels of LPC (Linear Predictive coefficient) perturbation. Various speech enhancement techniques have been considered here such as spectral subtraction, spectral over subtraction with use of a spectral floor, spectral subtraction with residual noise removal and time and frequency domain adaptive MMSE filtering. The speech signal sued here for recognition experimentation was a typical sentence with additive normally distributed white noise distortion.

The single channel speech enhancement algorithm at very low SNR has been presented in [6], which uses masking properties of human auditory system. This algorithm is the subtractive type in its nature and subtraction parameter is adapted as per the levels of rough estimate of the background noise and the added musical residual noise and thus making this algorithm adaptable to noise present in every frame of speech. In another interesting research [7], speech was enhanced from noise along with coding using discrete wavelet packet transform decomposition. Two stages of subtractive-type algorithm used, once estimating noise and subtracting it from noisy speech to have rough estimate of speech later, this estimate is further used to determine the time-frequency masking threshold assuming high-energy frames of speech will partially mask the input noise and hence reducing the need for a strong enhancement process. The both of these work used Noisex-92 database to evaluate the performance of their proposed algorithms. In yet another similar work [8], the noise autocorrelation function is estimated during non-speech activity periods and it is used in deciding the masking threshold for the speech enhancement. Here, author also uses frequency to Eigen-domain transformation to provide the upper bound estimate of residual noise to be introduced in the speech.

It is believed that the time distribution of speech samples is much better modelled by a Laplacian or a Gamma density functions rather than a Gaussian density function. The same is valid for short time DFT domain, typically, frame size less than 100ms [9]. Optimal estimators for speech enhancement in the Discrete Fourier Transform (DFT) domain is used for estimating complex DFT coefficients in the MMSE sense when the clean speech DFT coefficients are Gamma distributed and the DFT coefficients of the noise are Gaussian or Laplace distributed. When the noise model is a Laplacian density, this estimator outperforms other estimators in the sense it show less annoying random fluctuations in the residual noise than for a Gaussian density noise. In [10] and [11], adaptive estimation of non-stationary noise present in the speech has been presented.

3.3 Beamforming based Speech Enhancement

Frost [12] has suggested constrained minimum power adaptive beamforming, which deals with the problem of a broadband signal received by an array, where pure delay relates each pair of source and sensor. Each sensor signal is processed by a tap delay line filter after applying a proper time delay compensation to form delay-and-sum beamformer. The algorithm is capable of satisfying some desired frequency response in the look direction while minimizing the output noise power by using constrained minimization of the total output power. This minimization is realized by adjusting the taps of the filters under the desired constraint using constrained LMS-type algorithm. Griffiths and Jim [13] reconsidered Frost's algorithm and introduced the generalized sidelobe canceller (GSC) solution. The GSC algorithm is comprised of three building blocks. The first is a fixed beamformer, which satisfies the desired constraint. The second is a blocking matrix, which produces noise-only reference signals by blocking the desired signal (e.g., by subtracting pairs of time-aligned signals). The third is an unconstrained LMS-type algorithm that attempts to cancel the noise in the fixed beamformer output. In [13], it is shown that Frost algorithm can be viewed as a special case of the GSC. The main drawback of the GSC algorithm is its delay-only propagation assumption.

In another work [14], switching adaptive filters were used to form the beamformer. This beamformer has two sections and interconnected with switch. The first section determines the adaptive look direction and cues in on the desired speech and is adapted only when speech is present. Second section which adapted during silence-only periods is implemented as multichannel adaptive noise canceller. In [15], authors have proposed the solution to GSC algorithm by estimating ratio of transfer functions (TFs), otherwise it is based on TFs which relates source signal and the sensors. The TF ratios are estimated by exploiting the non-stationarity characteristic of the desired signal. This algorithm can be used normally in reverberating room having acoustic environment. One interesting paper [16], describes how optimal finite-impulse response subband beamforming can be used by including coherent multipath propagation into optimality criterion for speech enhancement in multipath environment.

In application point of view, a constrained switched adaptive beamforming (CSA-BF) [17] was used for speech enhancement and recognition in real moving car environment. This algorithm consists of a speech/noise constraint section, a speech adaptive beamformer and noise adaptive beamformer. The performance obtained with this algorithm was compared with classic delay-and-sum beamforming (DASB) using CU-Move corpus and found decrease in word-error-rate (WER) by 31% in speech recognition. The computational complexity of DASB is very low and can be easily implemented for real-time requirement. It is also effective when direction of desired source is known and can be applied in the car as driver's head position is restricted based on seat position. However, as there is possibility of change in drivers head direction, DASB algorithm could be inconsistent and this inconsistency can be solved by employing CSA-BF algorithm which can improve the SNR by up to +5.5 dB on the average. For the application of hands-free speech recognition, one of the works [18] uses sequence of features to be used for speech recognition itself, to optimize a filter-and-sum beamformer instead of separating the beamformer, to be used for speech enhancement, from speech recognition system. In this work, they used Mel Frequency Cepstral Coefficient (MFCC) and applied to the HMM based classifier for speech recognition.

Optimizing beamformer without knowledge of source or acoustic characteristic of environment is termed as "blind beamforming". One of the papers [19] proposes blind speech enhancement using beamformer which consist of subband soft-constrained adaptive filter using recursive least square (RLS) algorithm, combined with subband weighted time-delay estimator (TDE). Estimation of propagation time difference of arrival of a dominate speech source received by sensor array is based the steered response power with phase transform (SRP-PHAT) algorithm, which was modified to work in subband structure. One recent paper [20] presents phase-based dual-microphone speech enhancement technique based on prior speech model. In this work, it is claimed that around 23% improvement achieved using this algorithm as compared to the delay-and-sum beamformer, where experiments were conducted on the CARVUI database.

In application point of view, the study presented in [21] addresses the problem of distant speech acquisition in multiparty meeting s using multiple cameras and microphones. The camera, used as a multi-person tracker, was used to give the more precise location of each person to the microphone array beamformer. They evaluated the performance of speech recognition using data recorded in a real meeting room for stationary speaker, moving speaker and overlapping speech scenarios. The result obtained with audio-video speech enhancement was better than that with only audio. In one of the recent work [22], adaptive beamformer based on estimation of power spectral density (PSD) and noise statistics update was proposed. An inactive-source detector based on minimum statistics is developed to detect the speech presence and to acquire the noise statistics. The performances of this beamformers were tested in a real hands-free in-car environment. One of the most recent papers [23] uses GSC based speech enhancement using the location of speaker obtained via localization module. This algorithm relies on time delay compensation, DFT computations, fixed channel compensator, adaptive channel compensator.

3.4 Active Noise Cancellation

It is believed that if error sensor output is measured properly and mapped to the control speaker even with some propagation delay, then essential problem for the active noise cancellation structure is to predict the future values and/or components of noise. The work presented in [24], deals this problem with single sensor and predicts the noise model parameters with Kalman filter using non-gradient algorithm and gradient search algorithm. These two algorithms were applied on noise generated by a propeller aircraft, a helicopter and jet aircraft. The noise was reduced significantly, though; it was most in propeller case and least in jet noise. The gradient algorithm has also less computational complexity over the non-gradient algorithms. Another way of having active noise cancellation is to employ adaptive filter to characterize the transfer function between error mic to control noise in frequency domain and it is described in the [25]. The least mean square (LMS) algorithm was applied in each band of frequency decomposition using DFT to model the control signal. Further, a frequency-domain periodic active noise equalization (ANE) system, which reshapes the residual noise by controlling the output of the adaptive comb filter at each frequency bin, is also presented in this work.

In real-life application point of view, one of the interesting papers [26] presents the integrated feedback based approach for noise reduction headset for the audio and communication purposes. This system uses single microphone per ear cup which makes it simple in its applicability with existing audio and communication devices. In another work [27] of designing ANC headset, filtered-x least means square algorithm (FXLMS), which introduces secondary path for synthesizing the reference signal was implemented. One of the study [28] based on FXLMS analysed the algorithm and extended the new algorithm to reduce the multi-tonal noise. Similar work was developed in [29] and evaluated with narrowband and broadband noise against the instability in convergence of adaptation algorithm. The modified FXLMS algorithm based on efficient affine projection technique is compared with conventional FXLMS algorithm in [30]. It is claimed that modified filtered-x structure provides better convergence speed over conventional filtered-x algorithm but need more computations comparatively as it requires additional filtering channel.

3.5 Noise Types and Sources

The noise reduction is any active area of speech processing where noise is being reduced to achieve the noise free speech. This becomes critical in so many applications including speech recognition where main objective is to reduce the word error rate (WER) and at the same time providing flexibility to use system in anywhere irrespective of which kind of noise present in acquired speech. To design the efficient and generalized algorithm for speech enhancement, it is critical to have knowledge about the noise that could be present in speech. Mathematically, noise can be sometimes modelled by its probability distribution of values and represented in some work [31] by Gaussian distribution or Laplacian types or Gama types distributed speech in case speech modeling. The level of SNR can be important indicator to the strategy to be employed in enhancement algorithms. In the presence of high level SNR and low level SNR, algorithm can be switched from one enhancement scheme to another after estimating the SNR present in speech.

Depending of the stationarity characteristics of noise, the technique is to be designed to handle stationary and non-stationary noise separately.

There are some noises, whose source are known and are easy to be modelled and analyzed. Despite of known sources of noise, it becomes difficult to reduce the noise when more than one type of noises is present. In applications like noise free headset, it may become easy to overcome the problem of noise as ears are isolated from external region in large extent. In car application, noises generated by engine, wind and rain wipers are to be handled intelligently. Hands free devices have echo voice as most dominant noise. In close room meeting application, acoustic characteristics of walls can generate multiple echoes of voice. In crowd or outdoor meetings, lot many people may talk simultaneously and can overlap with principle speech. In traffic area, noise may be raised from other vehicles and impulsive honking. In car or at home sometimes background music can become source of noise for the speech of interest. In hands-free operation of cellular phones in vehicles, the speech signal to be transmitted may be contaminated by reverberation and background noise.

3.6 Real Time Speech Processing

Some of the definitions of real time speech processing [32] found from the different dictionaries and researchers are given below,

- In real-time speech processing system the correctness of the computations not only depends upon the logical correctness of the computation but also upon the time at which the reconstructed speech is produced. There is occurrence of system failure if the timing constraints of the system are not met. (gillies @ ee.ubc.ca).
- It is the time in which the occurrence and the reporting or recording of an event are almost simultaneous.
- The actual time used by a computer to solve and control speech processing problems effectively at the same time (Webster's dictionary) (Agnes and Guralnik, 2000)
- Real time speech processing requires a fast enough data processing to maintain with an outside process, it is a form of transaction processing in which each transaction is executed as soon as complete data becomes available for the transaction. (http://www.wordwebonline.com).

There are basically two types of real-time systems 'soft' and 'hard ' as shown in figure 3.1. Soft real-time system means a system which has reduced constraints on 'tardiness' but still must operate very quickly. In hard real-time system the type of a typical real-time system which requires a stringent deadline (http://media. wiley.com).

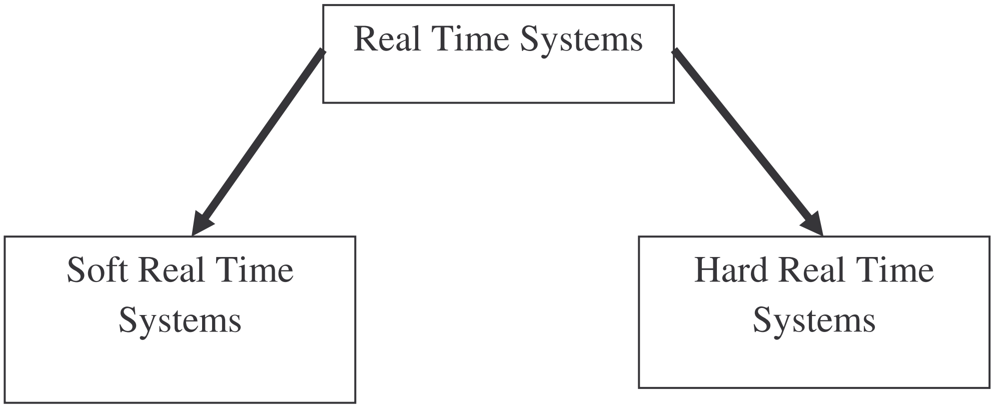


Fig. 3.1 Types of real time systems

Considering several definitions mentioned above, we can say a common basis of a real-time speech processing system requires explicit bounded response time constraints without system failure, with the logical correctness based on both the correctness of the reconstructed speech outputs and their timeliness. The response time is called the time between the presentation of a set of speech inputs and the appearance of all the associated reconstructed speech outputs. In one of the paper real time speech processing strategies[3] are explores for different applications.

3.7 Performance measurement of Real Time Systems

A real time system performance is a measure of the percentage of non-idle processing often measured by CPU utilization. A system is said to be time overloaded if it is 100% or more time-loaded. Systems that are time-overloaded are unstable and exhibit missed deadlines and unpredictable response times.

[3] Paper presented and published on "**Exploration of Real Time Speech Processing Strategies: A Review of Applications**", *National Conference on Emerging Trends in Electronics and Telecommunication Engineering (ETETE)*, Watumull Institute of Electronics Engineering and Computer Technology, Worli, Pages 205-208, Sept. 16-17, 2011.

Table 3.1 shows CPU utilization for real-time systems. Utilization factors in the 0%-69% range are generally considered as safe. Beyond 70% they have a high risk of missing deadlines, and above 100% are potentially terrible.

Table 3.1

CPU utilization for real-time systems

Utilization Percentage (%)	**Type of Zone**	**Applications Support**
0 - 25	Excess	Different
26 - 50	More safer	Different
51 - 68	Safe	Different
69	Theoretical limit	Different
70 - 99	Dangerous	Embedded systems
100 and above	Overload	Strained systems

3.8 Conclusion

The speech, being a fundamental way of communication for the humans, has been embedded in various essential applications like speech recognition, voice-distance-talk and other forms of personal communications. There are so many applications of speech still to be far from reality just because of lack of efficient and reliable noise removal mechanism and preserving or improving the intelligibility for the speech signals. The broad categories of speech enhancement techniques can be listed as speech filtering techniques, beam forming techniques and active noise cancellation methods. In this chapter, an attempt has been stepped towards surveying the methodologies for speech improvement. It was also interesting to discuss, how these techniques affect the performance of various application systems like speech recognition and speech communication. Essentially, we also discussed about the types and sources of noise that can be considered in speech enhancements. In last phase we also discussed about definitions of real time speech processing and the performance measures for real time systems.

CHAPTER 4

SOFT-COMPUTING: EVOLUTIONARY ALGORITHM

Applied Computational Intelligence and Soft Computing will focus on the disciplines of computer science, engineering, and mathematics. The scope includes developing applications related to all aspects of natural and social sciences by employing the technologies of computational intelligence and soft computing. The new applications of using computational intelligence and soft computing are still in development. Although computational intelligence and soft computing are established fields, the new applications of using computational intelligence and soft computing can be regarded as an emerging field.

Conventional techniques have successfully been applied for the solution of many complex real world problems in diverse areas, but solving a problem using traditional approach requires understanding and development of an algorithm. The algorithmic requirement limits their usefulness in applications such as automobile autopilot, intelligent robotics, computer vision, recognition of speech, hand written graphics, machine translation, learning through experience etc. where no exact mathematical relationships between input-output variables are available. Therefore a non-algorithmic approach to deal with such situations is required. Soft computing/Computational Intelligence is an engineering discipline that provides an alternative to algorithmic programming. The terms Soft computing and Computational Intelligence are generally used interchangeably in engineering literature.

The term soft computing was first coined by Zadeh in 1990s when there was intense competition between various methodologies linked to artificial intelligence. His perception was that more could be gained by cooperation than by claims and counterclaims of superiority. The principal constituents of soft computing are fuzzy logic, neurocomputing and probabilistic reasoning with the latter subsuming genetic algorithms, belief networks, chaotic systems, and parts of learning theory. In many cases a problem can be solved most effectively by using fuzzy logic, neural networks, and probabilistic reasoning in combination rather than exclusively. The main paradigms of Computational Intelligence are neurocomputing, evolutionary computing, swarm intelligence, and fuzzy logic. Soft Computing, in addition to the paradigms of Computational Intelligence, also includes probabilistic methods.

4.1 Neural Network

Artificial Neural Networks (ANNs) are of major research interest at present, involving researchers of many different disciplines. Subjects contributing to this research include biology, computing, electronics, mathematics, medicine, physics, and psychology. The approaches to this topic are very diverse, as are the aims. The basic idea is to use the knowledge of the nervous system and the human brain to design intelligent artificial systems. On one side biologists and psychologists are trying to model and understand the brain and parts of the nervous system and searching for explanations for human behavior and reasons for the limitations of the brain. On the other, computer scientists and electronic engineers are searching for efficient ways to solve problems for which conventional computers are currently used. Biological and psychological models and ideas are often the resource of inspiration for these scientists. In the computing environment the term Neural Network (NN) is usually used as synonym for artificial neural network.

Artificial NN draw much of their inspiration from the biological nervous system. It is therefore very useful to have some knowledge of the way this system is organized. Most living creatures, which have the ability to adapt to a changing environment, need a controlling unit which is able to learn. Higher developed animals and humans use very complex networks of highly specialized neurons to perform this task. The control unit or brain can be divided in different anatomic and functional sub-units, each having certain tasks like vision, hearing, motor and sensor control. The brain is connected by nerves to the sensors and actors in the rest of the body.

The brain consists of a very large number of neurons, about 1011 in average. These can be seen as the basic building bricks for the central nervous system (CNS). The neurons are interconnected at points called synapses. The complexity of the brain is due to the massive number of highly interconnected simple units working in parallel, with an individual neuron receiving input from up to 10000 others. The neuron contains all structures of an animal cell. The complexity of the structure and of the processes in a simple cell is enormous. Even the most sophisticated neuron models in artificial neural networks seem comparatively toy-like. Structurally the neuron can be divided in three major parts: the cell body (soma), the dendrites, and the axon, see figure 4.1 for an illustration.

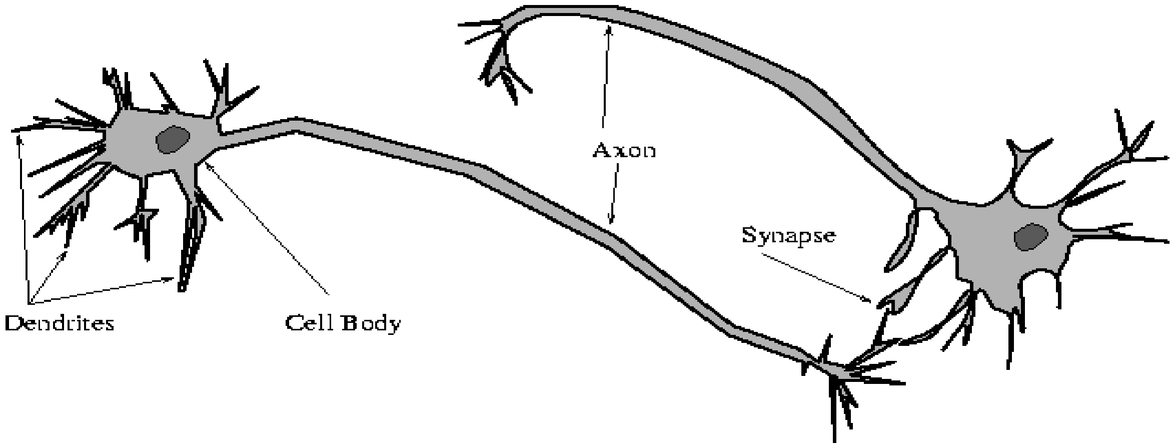


Fig. 4.1 Simplified Biological Neurons

The cell body contains the organelles of the neuron and also the `dentrites' are originating there. These are thin and widely branching fibers, reaching out in different directions to make connections to a larger number of cells within the cluster.

Input connections are made from the axons of other cells to the dentrites or directly to the body of the cell. These are known as axondentrititic and axonsomatic synapses. There is only one axon per neuron. It is a single and long fiber, which transports the output signal of the cell as electrical impulses (action potential) along its length. The end of the axon may divide in many branches, which are then connected to other cells. The branches have the function to fan out the signal to many other inputs. There are many different types of neuron cells found in the nervous system. The differences are due to their location and function. The neurons perform basically the following function: all the inputs to the cell, which may vary by the strength of the connection or the frequency of the incoming signal, are summed up. The input sum is processed by a threshold function and produces an output signal. The processing time of about 1ms per cycle and transmission speed of the neurons of about 0.6 to 120 {ms} are comparing slow to a modern computer [1-2].

The brain works in both a parallel and serial way. The parallel and serial nature of the brain is readily apparent from the physical anatomy of the nervous system. That there is serial and parallel processing involved can be easily seen from the time needed to perform tasks. For example a human can recognize the picture of another person in about 100ms. Given the processing time of 1 ms for an individual neuron this implies that a certain number of neurons, but less than 100, are involved in serial; whereas the complexity of the task is evidence for a parallel processing, because a difficult recognition task cannot be performed by such a small number of neurons, example taken from [1]. This phenomenon is known as the 100-step-rule.

Biological neural systems usually have a very high fault tolerance. Experiments with people with brain injuries have shown that damage of neurons up to a certain level does not necessarily influence the performance of the system, though tasks such as writing or speaking may have to be learned again. This can be regarded as re-training the network.

The artificial neuron shown in figure 4.2 is a very simple processing unit. The neuron has a fixed number of inputs n; each input is connected to the neuron by a weighted link wi. The neuron sums up the net input according to the equation: net = $\sum_{i=1}^{n}$ xi wi or expressed as vectors net = xT w. To calculate the output activation function f is applied to the net input of the neuron. This function is either a simple threshold function or a continuous non linear function. Two often used activation functions are:

$$f_C(\text{net}) = \{11\text{-e-net}\} \tag{4.1}$$

$$f_T(\text{net}) = \{\{\text{ if } a > \theta \text{ then } 1 \text{ else } 0\ \} \tag{4.2}$$

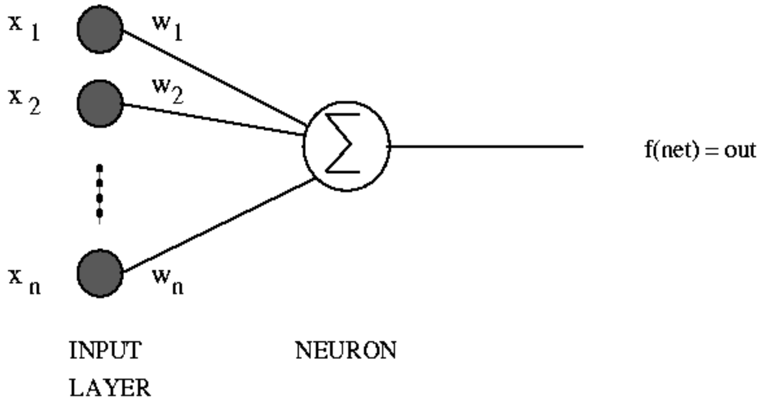


Fig.4.2 An Artificial Neuron

The artificial neuron is an abstract model of the biological neuron. The strength of a connection is coded in the weight. The intensity of the input signal is modeled by using a real number instead of a temporal summation of spikes. The artificial neuron works in discrete time steps; the inputs are read and processed at one moment in time. There are many different learning methods possible for a single neuron. Most of the supervised methods are based on the idea of changing the weight in a direction that the difference between the calculated output and the desired output is decreased. Examples of such rules are the Perceptron Learning Rule, the Widrow-Hoff Learning Rule, and the Gradient descent Learning Rule. The Gradient descent Learning Rule operates on a differentiable activation function. The weight updates are a function of the input vector x, the calculated output f(net), the derivative of the calculated output f'(net), the desired output d, and the learning constant η.

$$net = xT\ w \qquad (4.3)$$

$$\Delta w = \eta f'(net)\ (d-f(net))\ x \qquad (4.4)$$

The delta rule changes the weights to minimize the error. The error is defined by the difference between the calculated output and the desired output. The weights are adjusted for one pattern in one learning step. This process is repeated with the aim to find a weight vector that minimizes the error for the entire training set. A set of weights can only be found if the training set is linearly separable [3]. This limitation is independent of the learning algorithm used; it can be simply derived from the structure of the single neuron.

Multilayer networks solve the classification problem for non linear sets by employing hidden layers, whose neurons are not directly connected to the output. The additional hidden layers can be interpreted geometrically as additional hyper-planes, which enhance the separation capacity of the network. Figure 4.3 shows typical multilayer network architectures.

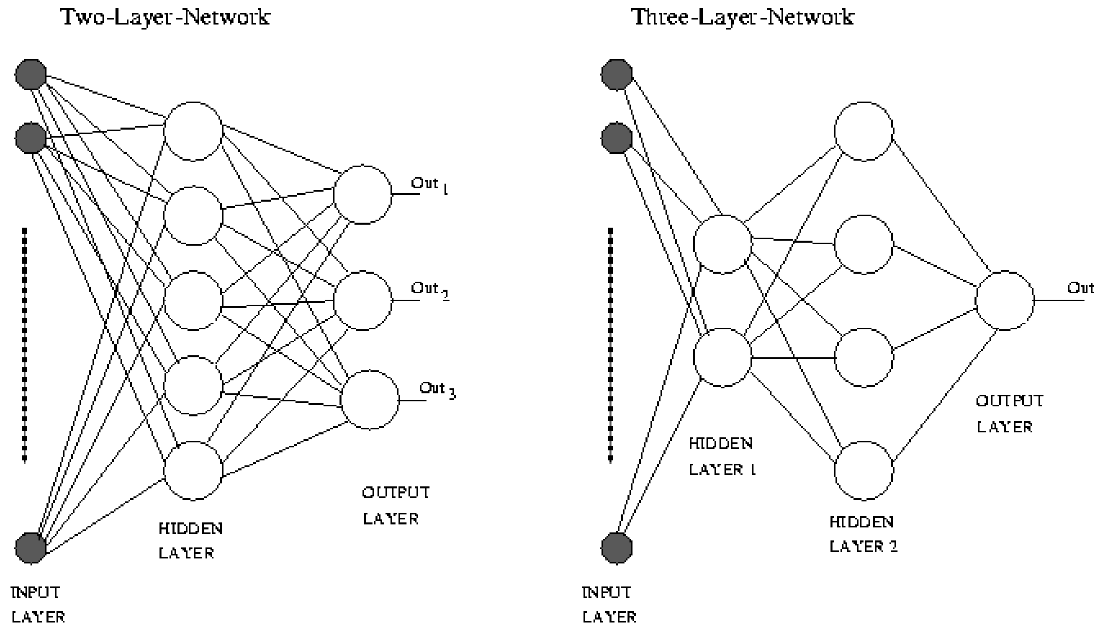


Fig.4.3 Examples of Multilayer Neural Network Architectures

This new architecture introduces a new question: how to train the hidden units for which the desired output is not known. The Back Propagation algorithm offers a solution to this problem.

4.2 Fuzzy logic

Since the development of the theory of fuzzy sets, started with the 1965 paper “Fuzzy Sets” [4], and the introduction of the concept of a linguistic variable, that is, a variable whose values are words rather than numbers [5], the concept of a linguistic variable has played and is continuing to play a pivotal role in the development of fuzzy logic and its applications [6]. Fuzzy logic is a precise logic of imprecision and approximate reasoning and it may be viewed as an attempt at formalization/mechanization of two remarkable human capabilities. First is the capability to converse, reason and make rational decisions in an environment of imprecision, uncertainty, incompleteness of information, conflicting information, partiality of truth and partiality of possibility - in short, in an environment of imperfect information. And second, the capability to perform a wide variety of physical and mental tasks without any measurements and any computations [7].

A fuzzy system is a control system that utilizes the fundamental principles of fuzzy logic to deliver a definitive conclusion to a problem that is characterized by vague, ambiguous, imprecise, noisy, or even missing information. Systems of this nature are often referred to as fuzzy systems (FS), fuzzy knowledge based systems (FKBS) and fuzzy inference system (FIS); all of which are relatively interchangeable and amount to the same thing. Fuzzy systems use fuzzy sets and fuzzy if-then rules as a part of a computer systems decision making process in order to draw conclusions.

Normally, a fuzzy system has specific steps fundamental to the design procedure. The diagram below, figure 4.4, illustrates the steps taken during any application system. The steps are listed and discussed as follows:

1. Pre-processing
2. Fuzzification
3. Rule Base
4. Inference Engine
5. Defuzzification
6. Post-processing

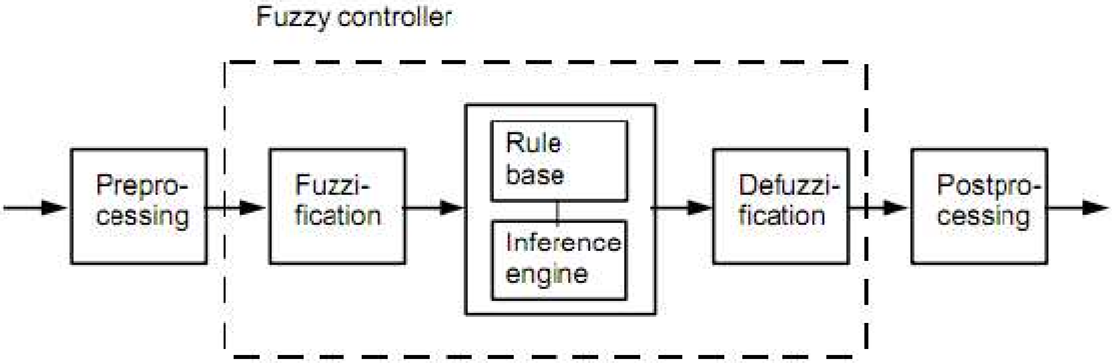


Fig. 4.4 Structure of fuzzy system

4.3 Evolutionary Computations

To tackle complex real world problems, scientists have been looking into natural processes and creatures both as model and metaphor for years. Optimization is at the heart of many natural processes including Darwinian evolution, social group behavior and foraging strategies. Over the last few decades, there has been remarkable growth in the field of nature-inspired search and optimization algorithms. Currently these techniques are applied to a variety of problems, ranging from scientific research to industry and commerce. The two main families of algorithms that primarily constitute this field today are the evolutionary computing methods and the swarm intelligence algorithms. Although both families of algorithms are generally dedicated towards solving search and optimization problems, they are certainly not equivalent, and each has its own distinguishing features. Reinforcing each other's performance makes powerful hybrid algorithms capable of solving many intractable search and optimization problems. In the last few decades, the continuing advance of modern technology has brought about something new. Evolution is now producing practical benefits in a very different field, and this time, the creationists cannot claim that their explanation fits the facts just as well. This field is computer science, and the benefits come from a programming strategy called genetic algorithms.

Concisely stated, a genetic algorithm (or GA for short) is a programming technique that mimics biological evolution as a problem-solving strategy. Given a specific problem to solve, the input to the GA is a set of potential solutions to that problem, encoded in some fashion, and a metric called a fitness function that allows each candidate to be quantitatively evaluated. These candidates may be solutions already known to work, with the aim of the GA being to improve them, but more often they are generated at random. The GA then evaluates each candidate according to the fitness function. In a pool of randomly generated candidates, of course, most will not work at all, and these will be deleted. However, purely by chance, a few may hold promise-they may show activity, even if only weak and imperfect activity, toward solving the problem.

These promising candidates are kept and allowed to reproduce. Multiple copies are made of them, but the copies are not perfect; random changes are introduced during the copying process. These digital offspring then go on to the next generation, forming a new pool of candidate solutions, and are subjected to a second round of fitness evaluation. Those candidate solutions which were worsened, or made no better, by the changes to their code are again deleted; but again, purely by chance, the random variations introduced into the population may have improved some individuals, making them into better, more complete or more efficient solutions to the problem at hand. Again these winning individuals are selected and copied over into the next generation with random changes, and the process repeats. The expectation is that the average fitness of the population will increase each round, and so by repeating this process for hundreds or thousands of rounds, very good solutions to the problem can be discovered.

As astonishing and counterintuitive as it may seem to some, genetic algorithms have proven to be an enormously powerful and successful problem-solving strategy, dramatically demonstrating the power of evolutionary principles. Genetic algorithms have been used in a wide variety of fields to evolve solutions to problems as difficult as or more difficult than those faced by human designers. Moreover, the solutions they come up with are often more efficient, more elegant, or more complex than anything comparable a human engineer would produce. In some cases, genetic algorithms have come up with solutions that baffle the programmers who wrote the algorithms in the first place also it is interesting to note that soft computing techniques soft computing based speech recognition techniques can be used for speech enhancement for multimedia applications[4].

4.4 Conclusion

The continuing advance of computational technology such as availability of large memories in small space and parallel GPUs have changed the paradigm of the way the problem used to be solved. The soft-computing is the new paradigm for the computationally solvable complex problems and has been heavily relied on the computational power of devices. It includes the probabilistic theory in addition to the elements covered by computational intelligence. In this chapter, we have given brief description of main elements of the soft-computing, such as neural network, fuzzy logic and genetic algorithm.

[4] Paper published on "**Survey of Soft Computing based Speech Recognition Techniques for Speech Enhancement in Multimedia Applications**", *International Journal of Advance Research in Computer and Communication Engineering (IJARCCE)*, Paper ID:V25105, Certificate No: V215C008-1/2, ISSN(Online):2278-1021, ISSN(Print):2319-5940, Pages 2039-2043, Volume 2, Issue 5, May 2013.

CHAPTER 5

SPEECH ENHANCEMENT AND BEAMFORMING

In general, the digital voice communications, human-machine interfaces, and automatic speech recognition systems have been so integral part of the human life that it should also perform reliably in noisy environments. In hands-free operation of cellular phones in car, the speech signal to be transmitted may be contaminated by vibration, engine and background noise. Most of the speech based system works well in noise-free situations, but their performance becomes intolerable in noisy atmosphere. Thus, researchers across the world have been working in the speech enhancement algorithms that need to be placed before the application processing module. Therefore, development of real-time preprocessing algorithms for speech enhancement is critical step for making any portable and mobile device. The goal of speech enhancement varies according to specific applications, such as to boost the overall speech quality, to increase intelligibility[5], and to improve the performance of voice communication devices. Similarly, speech image enhancement techniques can be used in medical devices and healthcare applications. Beamforming is the one of the effective techniques for the speech enhancement[6].

Beamforming techniques can be broadly classified as being either fixed or adaptive. Fixed beamformers are so named because their parameters are fixed during operation. Adaptive beamformer is continuously updated based on the received signals. As different beamforming techniques are suitable for different noise conditions, hence, having knowledge of noise levels and types is essential before applying speech beamforming. We present here detailed explanation about basic beamformer.

5.1 Delay-Sum Beamformer

The simplest microphone array beamforming technique is delay-sum beamforming. In time domain beamforming, a finite impulse response (FIR) filter like structure is applied to each microphone signal, and the filter outputs combined to form the beamformer output [1]. Beamforming can be performed by computing resultant multichannel filters output and it is given by equation 5.1

[5] Paper presented and Published on "**Speech Enhancement Techniques: Quality vs. Intelligibility**", in *Fourth International Conference on Electronics Computer Technology (ICECT)*, Kanyakumari, India, Paper ID: E2121, Published in International Journal of Computer and Communication Volume 3, Number 3 , April 6-8, 2012.

[6] Paper published on "**Improvement in Speech Recognition Performance using Beamforming based Speech Enhancement**", in *International Journal of Electronics Communication and Computer Engineering (IJECCE),* Paper ID:730, ISSN:2249-071X, Volume 3, Issue 4, Pages 745-751, July 2012.

$$\hat{s}(t) = \sum_{i=1}^{N} \sum_{p=0}^{P-1} w_{i,p} x_i(t-p) \tag{5.1}$$

Where ‘P–1’ is the number of delays in each of the ‘N’ filters. In frequency domain beamforming, the microphone signal is separated into narrowband frequency bins (for example using a STFT), and the data in each frequency bin is processed separately. Using LMS algorithm, weights of each filter are estimated using training phase. For deterministic optimal condition, delay and sum elements are used to converge LMS algorithm. For statistical optimality, wiener filter is used.

5.2 Filter-Sum Beamformer

The delay-sum beamformer belongs to a more general class known as filter-sum beamformers, in which both the amplitude and phase weights are frequency dependent. The output of a filter-sum beamformer is given by equation 5.2

$$y(f) = \sum_{n=1}^{N} w_n(f) x_n(f) \, . \tag{5.2}$$

The typical filter-sum beamformer is shown in figure 5.1.

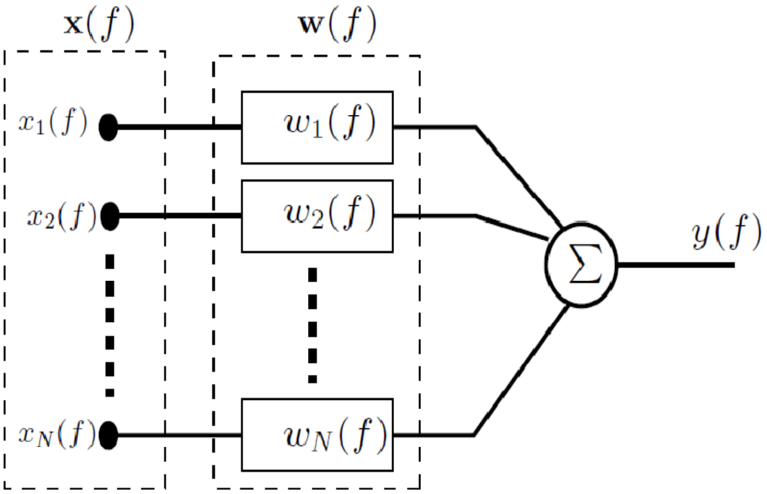


Fig. 5.1 Filter-Sum Beamformer structure

5.3 Sub-Array Beamformer

The directivity pattern of a uniform linear array (ULA) depends on the frequency of interest, the inter-element spacing (or effective length, as L = Nd), and the number of elements in the array. The dependency on the operating frequency means that the response characteristics (beam-width and side lobe level) will only remain constant for narrow-band signals. Speech, however, is a broad-band signal, meaning that a single ULA is inadequate if a frequency invariant beam-pattern is desired. One simple method of covering broadband signals is to implement the array as a series of sub-arrays, here they are themselves ULAs. These sub-arrays are designed to give desired response characteristics for a given frequency range.

As the frequency increases, a smaller array length is required to maintain constant beam-width. To ensure the side lobe level remains the same across different frequency bands, the number of elements in each sub-array should remain the same. The sub-arrays are generally implemented in a nested fashion, such that any given sensor may be used in more than one sub-array. Each sub-array is restricted to a different frequency range by applying band-pass filters, and the overall broad-band array output is formed by recombining the outputs of the band-limited sub-arrays. An example of such a nested sub-array structure for delay-sum beamforming is shown in figure 5.2.

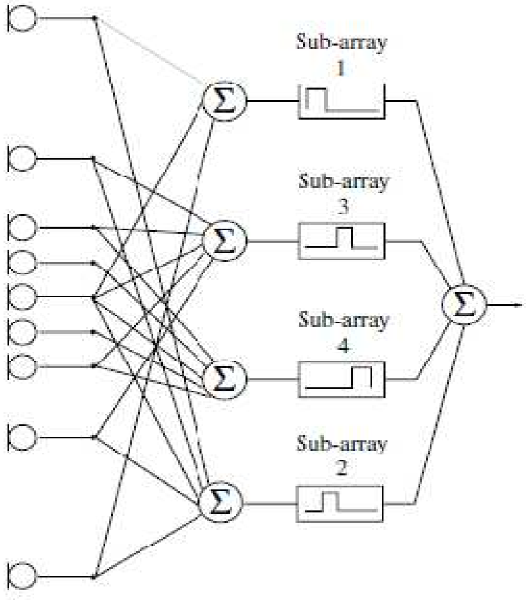


Fig. 5.2 Sample Nested Sub-Array structure

5.4 Super-Directive Beamformers

Conventional delay-sum beamformers have directivity that is approximately proportional to the number of sensors, N. Super-directive beamformers are designed to maximize the array gain, or the directivity, for one (or a few) principle desired direction, while suppressing noise coming from all other directions. Low frequency performance is problematic for conventional beamforming techniques because large wavelengths give negligible phase differences between closely spaced sensors, leading to poor directive discrimination. Delay-weight-sum beamformers can roughly cover the octave band $0.25 < d/\lambda < 0.5$ (where d is the inter-element spacing) before excessive loss of directivity occurs [2-3].

A frequency of 100 Hz corresponds to a wavelength of 3.4 m for sound waves, so this frequency range requires that $0.85m < d < 1.7m$. For a sub-array of 5 elements, this would give an array dimension of $3.4m < L < 6.8m$, which is impractical for many applications. For example, in the context of a multimedia workstation, it is desirable that the array dimension does not exceed the monitor width, which will be approximately 17 inches, or 40 cm.

Thus methods providing good low frequency performance with realistic array dimensions are required. One such method is a technique called near-field super-directivity. As its name implies, near-field super-directivity is a modification of the standard super-directive technique, in which the propagation vector d is replaced by one formulated for a near-field source.

5.5 Conclusion

A new generation of speech acquisition applications is emerging as a result of advances in technology and the prevalence of mobile and broadband communication. Thus, it becomes essential to have reliable speech processing based applications. The speech is corrupted with so many different types of noises and by cross voices. This presents the need of cleaning out the speech so that applications can perform without any flaws. In this chapter, we have described in detail about the speech enhancement theory and especially with beamforming techniques.

CHAPTER 6

SPEECH EXPERIMENTAL DATASET

6.1 Dataset

The dataset for any research experiment is a very critical component. It is also important to have control on the parameters of data and to know the ground truth about it. In this research we considered the isolated word, as objective of our research was to examine the performance of speech recognition. For analysing the performance of speech enhancement based speech recognition, we have considered here four speaker's 20 number of spoken words. Since these words are regularly used in every human's life, we have chosen these words. These words are listed below and can be categorised on the basis of their use, as numbers and commands. The spoken word from speaker has a length of 2 sec in time. The every speech was recorded with 16 KHz data rate.

Table 6.1

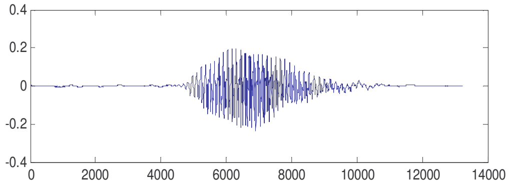
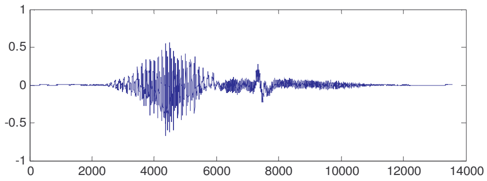
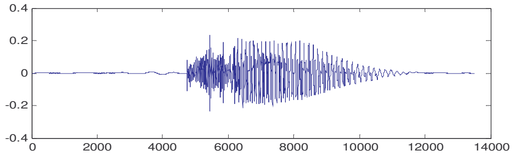
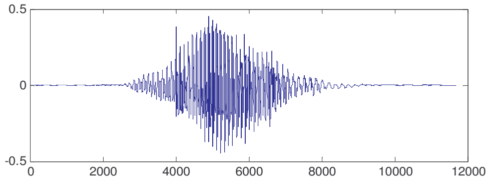
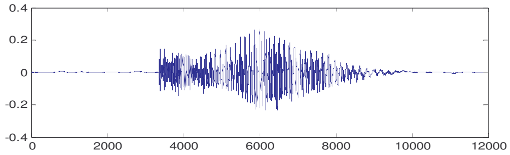
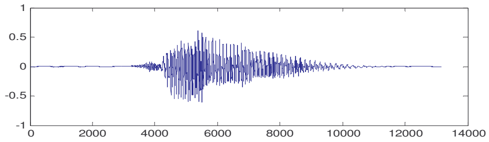
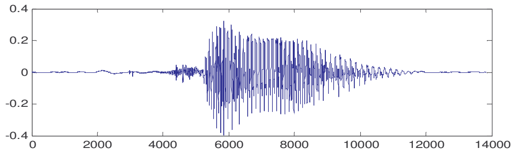
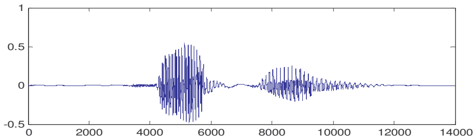
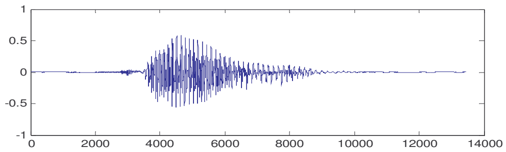
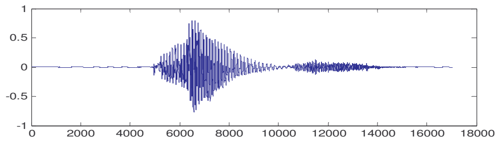
List of Spoken Words

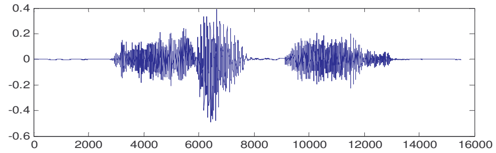
Spoken Words (each for 2 sec)	
Numbers	**Commands**
one	yes
two	no
three	hello
four	open
five	close
six	start
seven	stop
eight	dial
nine	on
ten	off

In table 6.1, the speech waveforms of these words are shown.

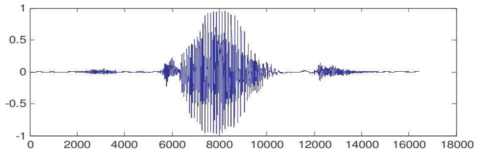
Table 6.2

Waveforms of Spoken Words

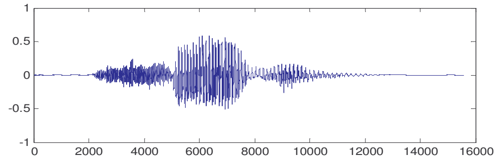
Spoken word's waveform (each for 2 sec)	
Numbers	**Commands**
 **One**	 **Yes**
 **Two**	 **No**
 **Three**	 **Hello**
 **Four**	 **Open**
 **Five**	 **Close**



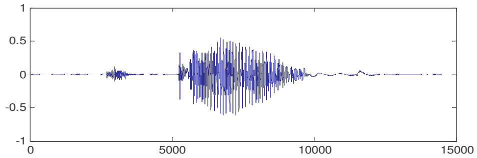
Six



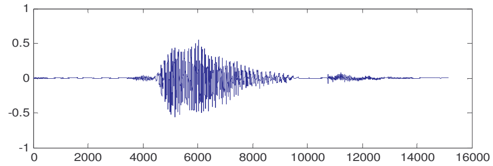
Start



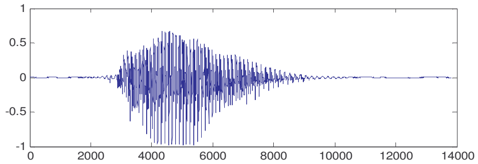
Seven



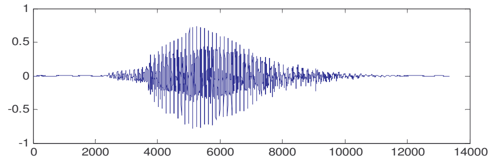
Stop



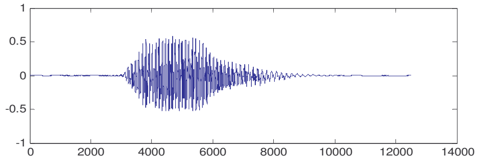
Eight



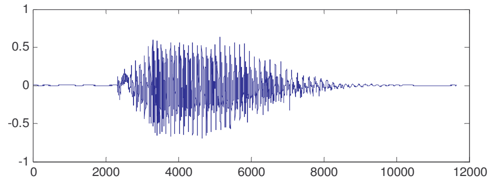
Dial



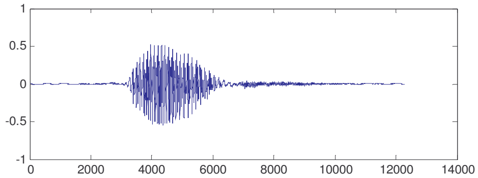
Nine



On



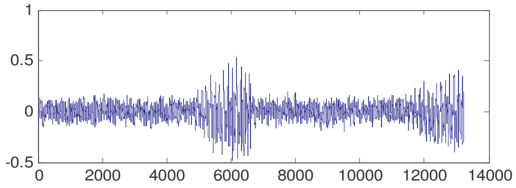
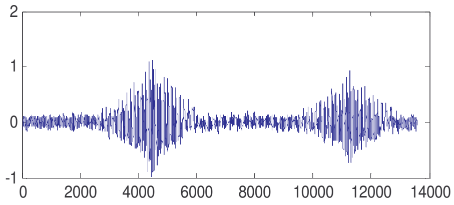
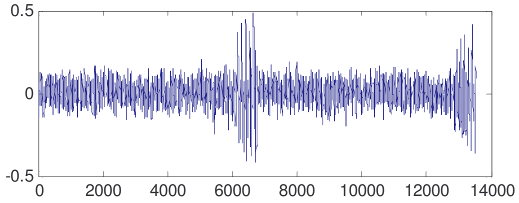
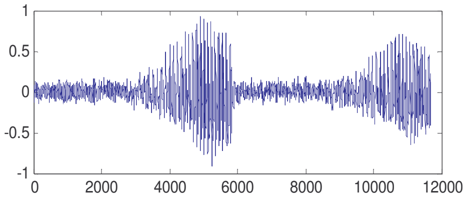
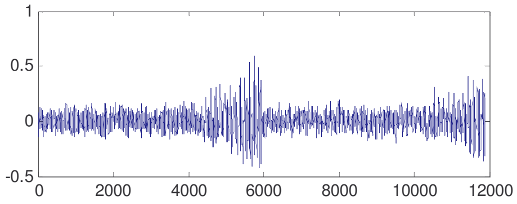
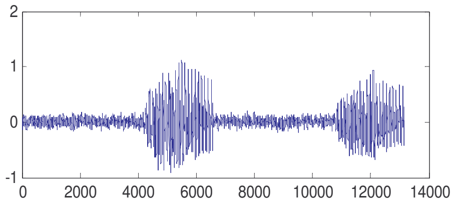
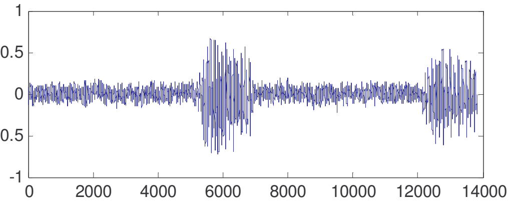
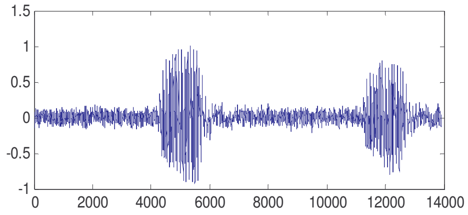
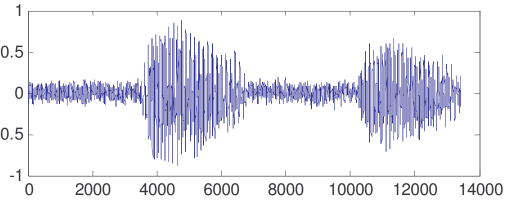
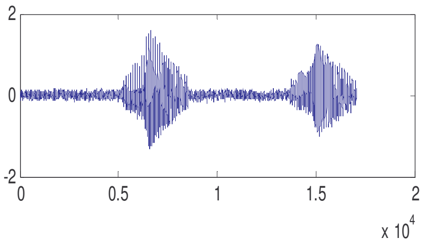
Ten

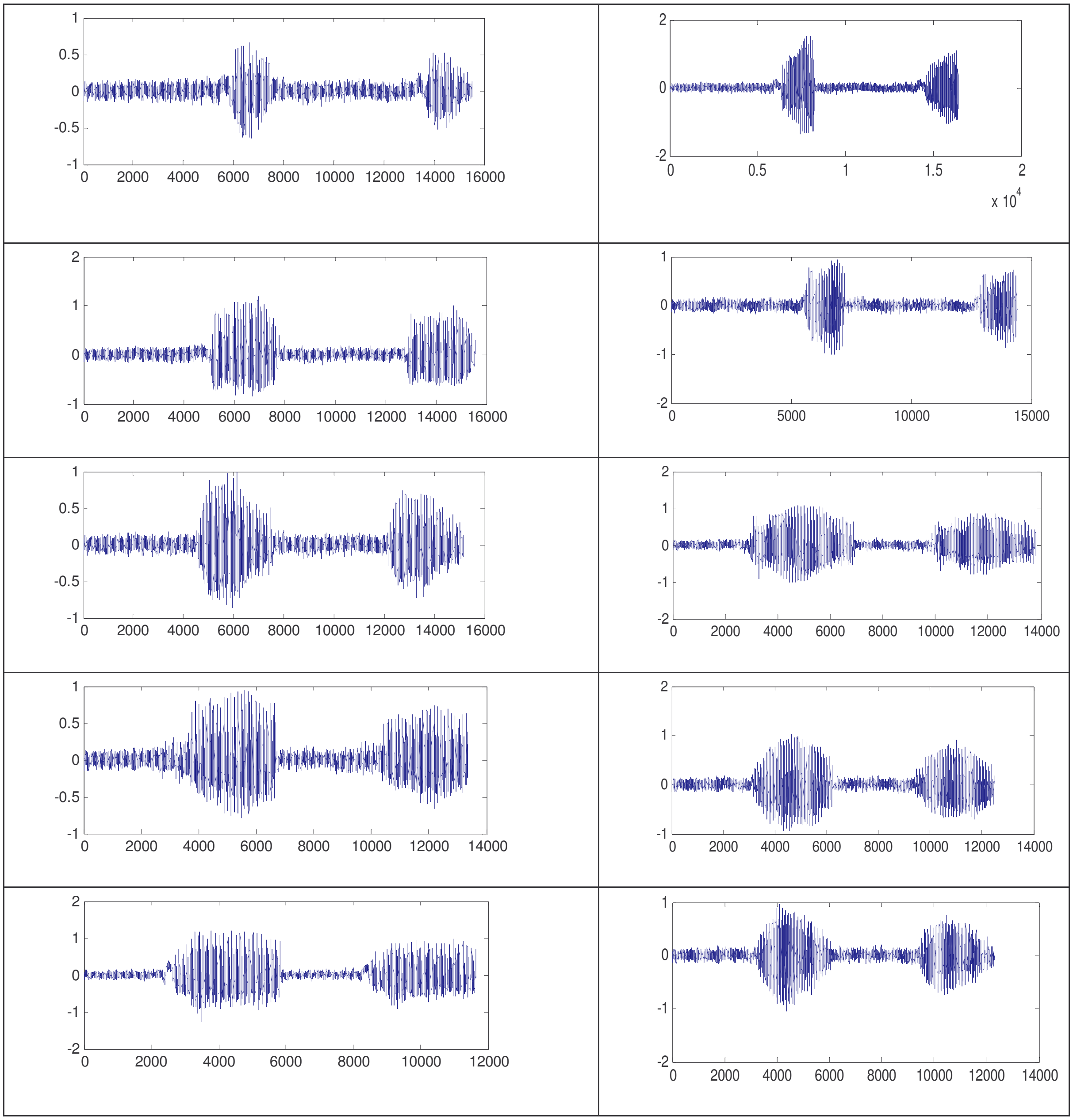


Off

Table 6.3

Simulated Multi-source waveforms of Spoken Words

Spoken words after Multi-mic and Echo (Multi-sources)	
Numbers	**Commands**
	
	
	
	
	



The multi-mic or multi source simulated voice showing table 6.3 can be observed to see the difference with raw voice waveforms shown in table 6.1. The multi-source simulated voice are made up with resultant speech obtained from six different delayed FIR filters, white noise source and echo component of raw voice.

6.2 Conclusion

While experimenting with speech enhancement using beamforming technique and later for the speech recognition experiments, there is a need of dataset with ground truth. There are not many datasets available in public across the world. It was necessary to construct the dataset using beamforming parameters. In doing so, we have attempted to do the simulation of speech database to be used for speech recognition experiments with beamforming parameters. In this chapter, the detail of this simulation has been provided about this dataset.

CHAPTER 7

SPEECH RECOGNITION USING BEAMFORMING TECHNIQUE

7.1 Introduction

With advancement in hardware technologies, there are so many electronic and mobile personal communication based devices available, today in market and that too in cheaper cost and with easy availability. The applications like speech recognition, mobile and personal communication, public address system are few of the applications from long list of speech based systems. However, undesired noises in environment like sound from heavy machines, vehicles are also present in one or other form everywhere. These noises cause undesired effects in speech transmission and acquiring systems. Recently restricted or usable vicinity of applications is moving from one place and close room to more open and multiple locations, leading to several types of undesired signals of mixing with desired speech signal making speech more corrupt with noise. Not only human communications but intelligent machines which trying to automate the things and sometimes also takes decision based on what it receives as a speech, also suffers from the degraded performance.

Since last five decades, various approaches for noise reduction and speech enhancements have been investigated and developed. Among, very early and fundamental approach of noise reduction was introduced to use the theory of the optimum Wiener filter. Given a desired signal and an input signal, the Wiener filter produces an estimate of the desired signal that is optimal, i.e. the squared mean error or difference between the signals is minimized. The Wiener filter can also be adaptively estimated used in an environment where the surrounding noise has time-varying characteristics. Adaptive algorithms such as Least Mean Square (LMS) and Recursive Least Squares (RLS) are well known examples and also widely used.

Recent advances in CPU and multi-core hardware has provided ample amount of computational power and thus, need for today is to design the complex but yet efficient and realistic approach for noise reduction to achieve speech enhancement. The speech enhancement is not only useful for storage and transmission of speech data but it can play vital role in improving much need system based speech recognition where accurate identification of words and sentences can provide automation in most of the human-machine based interface and also be useful in machine-machine interaction based automation. Robotics is a familiar example where speech recognition systems can become boon for today's advanced society at social level in addition to during natural calamities and on war fields.

It is obvious that speech enhancement can boost up the performance of speech recognition systems by keeping low word error rate (WER). There are various types of advanced speech enhancement algorithms in literature and they can be classified in main three categories, namely; filtering/estimation based noise reduction, beam forming and active noise cancellation (ANC) techniques. In partly implementation of this thesis, our work has two-fold objective. First is to improve the speech recognition performance in multi-microphone environment. Second, we attempted to analyze the performance of speech recognition against the filter-bank parameters; filter length and number of subbands. The experiments were performed for 20 words including numbers and commands, 10 words of numbers only and 10 words of commands only for different values of filter bank parameters. The results obtained have proved the speech enhancing capability of the beamforming technique in multi-microphone network where noise and echo-interference can degrade the original speech signal.

7.2 Existed Work Related to Beamforming Technique based Speech Recognition

One of important class of speech enhancement methods is based on the beam-forming, where more than one speech channels (microphones) are used to process the speech. Speech signals are received simultaneously by all microphones and outputs of these sensors are then processed to estimate the clean speech signal. In adaptive beamforming, an array of antennas is exploited to achieve maximum reception in a specified direction by estimating the signal arrival from a desired direction (in the presence of noise) while signals of the same frequency from other directions are rejected. This is achieved by varying the weights of each of the sensors (antennas) used in the array. This kind of speech enhancement techniques can give better performance of the speech applications like automatic speech recognition (ASR) than signal channel processing. Only disadvantage with this class of methods is higher cost of hardware, which can put restriction on using these methods in some speech applications.

Frost [1] has suggested constrained minimum power adaptive beamforming, which deals with the problem of a broadband signal received by an array, where pure delay relates each pair of source and sensor. Each sensor signal is processed by a tap delay line filter after applying a proper time delay compensation to form delay-and-sum beamformer. The algorithm is capable of satisfying some desired frequency response in the look direction while minimizing the output noise power by using constrained minimization of the total output power. This minimization is realized by adjusting the taps of the filters under the desired constraint using constrained LMS-type algorithm. Griffiths and Jim [2] reconsidered Frost’s algorithm and introduced the generalized sidelobe canceller (GSC) solution. The GSC algorithm is comprised of three building blocks. The first is a fixed beamformer, which satisfies the desired constraint. The second is a blocking matrix, which produces noise-only reference signals by blocking the desired signal (e.g., by subtracting pairs of time-aligned signals). The third is an unconstrained LMS-type algorithm that attempts to cancel the noise in the fixed beamformer output. In [2], it is shown that Frost algorithm can be viewed as a special case of the GSC. The main drawback of the GSC algorithm is its delay-only propagation assumption.

In another work [3], switching adaptive filters were used to form the beamformer. This beamformer has two sections and interconnected with switch. The first section determines the adaptive look direction and cues in on the desired speech and is adapted only when speech is present. Second section which adapted during silence-only periods is implemented as multichannel adaptive noise canceller. In [4], authors have proposed the solution to GSC algorithm by estimating ratio of transfer functions (TFs), otherwise it is based on TFs which relates source signal and the sensors. The TF ratios are estimated by exploiting the non-stationarity characteristic of the desired signal. This algorithm can be used normally in reverberating room having acoustic environment. One interesting paper [5], describes how optimal finite-impulse response subband beamforming can be used by including coherent multipath propagation into optimality criterion for speech enhancement in multipath environment.

In application point of view, a constrained switched adaptive beamforming (CSA-BF) [6] was used for speech enhancement and recognition in real moving car environment. This algorithm consists of a speech/noise constraint section, a speech adaptive beamformer and noise adaptive beamformer. The performance obtained with this algorithm was compared with classic delay-and-sum beamforming (DASB) using CU-Move corpus and found decrease in word-error-rate (WER) by 31% in speech recognition. The computational complexity of DASB is very low and can be easily implemented for real-time requirement. It is also effective when direction of desired source is known and can be applied in the car as driver's head position is restricted based on seat position. However, as there is possibility of change in drivers head direction, DASB algorithm could be inconsistent and this inconsistency can be solved by employing CSA-BF algorithm which can improve the SNR by up to +5.5 dB on the average. For the application of hands-free speech recognition, one of the works [7] uses sequence of features to be used for speech recognition itself, to optimize a filter-and-sum beamformer instead of separating the beamformer, to be used for speech enhancement, from speech recognition system. In this work, they used frequency cepstral coefficient (MFCC) and applied to the HMM based classifier for speech recognition.

Optimizing beamformer without knowledge of source or acoustic characteristic of environment is termed as "blind beamforming". One of the papers [8] proposes blind speech enhancement using beamformer which consist of subband soft-constrained adaptive filter using recursive least square (RLS) algorithm, combined with subband weighted time-delay estimator

(TDE). Estimation of propagation time difference of arrival of a dominate speech source received by sensor array is based the steered response power with phase transform (SRP-PHAT) algorithm, which was modified to work in subband structure. One recent paper [9] presents phase-based dual-microphone speech enhancement technique based on prior speech model. In this work, it is claimed that around 23% improvement achieved using this algorithm as compared to the delay-and-sum beamformer, where experiments were conducted on the CARVUI database.

In application point of view, the study presented in [10] addresses the problem of distant speech acquisition in multiparty meeting s using multiple cameras and microphones. The camera, used as a multi-person tracker, was used to give the more precise location of each person to the microphone array beamformer. They evaluated the performance of speech recognition using data recorded in a real meeting room for stationary speaker, moving speaker and overlapping speech scenarios. The result obtained with audio-video speech enhancement was better than that with only audio. In one of the recent work [11], adaptive beamformer based on estimation of power spectral density (PSD) and noise statistics update was proposed. An inactive-source detector based on minimum statistics is developed to detect the speech presence and to acquire the noise statistics. The performances of this beamformers were tested in a real hands-free in-car environment. One of the most recent papers [12] uses GSC based speech enhancement using the location of speaker obtained via localization module. This algorithm relies on time delay compensation, DFT computations, fixed channel compensator, adaptive channel compensator.

7.3 Beamforming Filter Structure

In order to analyse the performance of the speech recognition in the speech corrupted by noise and echo-interference, the analysis frame work used here is depicted in figure 7.1. The noisy speech is simulated using multi-microphone speech environment is shown in figure 7.2. The main section beamforming based speech enhancement, filter bank design is explained in next subsection. In later subsection, multi-microphones speech generation is explained in details.

7.3.1 Beamform Filter Design

Adaptive Filtering is an important technique in the field of speech processing including speech enhancement, echo and interference cancellation and speech coding. Filter banks have been introduced in order to improve the performance of time domain adaptive filters with additional benefits like faster convergence and the reduction of computational complexity with shorter filters in the subbands being processed at reduced sampling rate [13]. Due to inappropriate structure of filter bank in subband processing and improper design of filters, filter bank may yield degraded performance. The subband FIR filter bank scheme [14] to be used for beamforming is shown in figure 7.3. The design of filters used here is adapted from and given in detail in references [14-17]. The design includes the prototype analysis and synthesis filter. The filter bank is obtained by using cosine modulation of prototype filter. The analysis-synthesis filter bank structure is shown in figure 7.4.

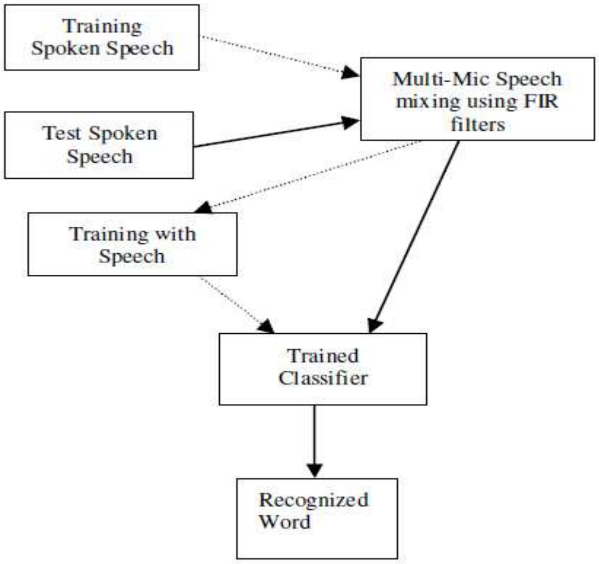


Fig. 7.1 Analysis Framework for Speech Recognition Performance using Beamforming

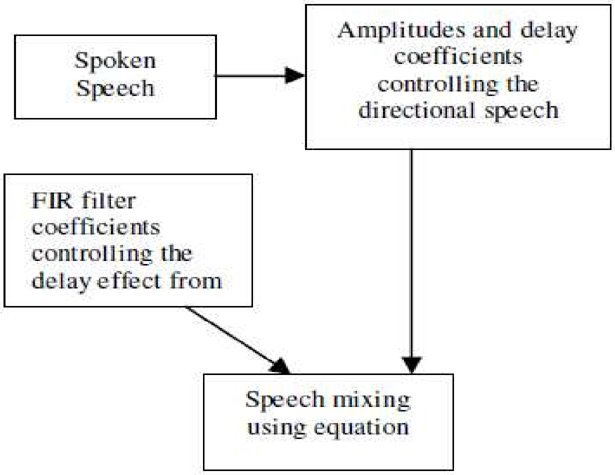

Spoken Speech
Amplitudes and delay coefficients controlling the directional speech
FIR filter coefficients controlling the delay effect from
Speech mixing using equation

Fig. 7.2 Speech-Splitting Scheme for simulating multi-microphones speech environment

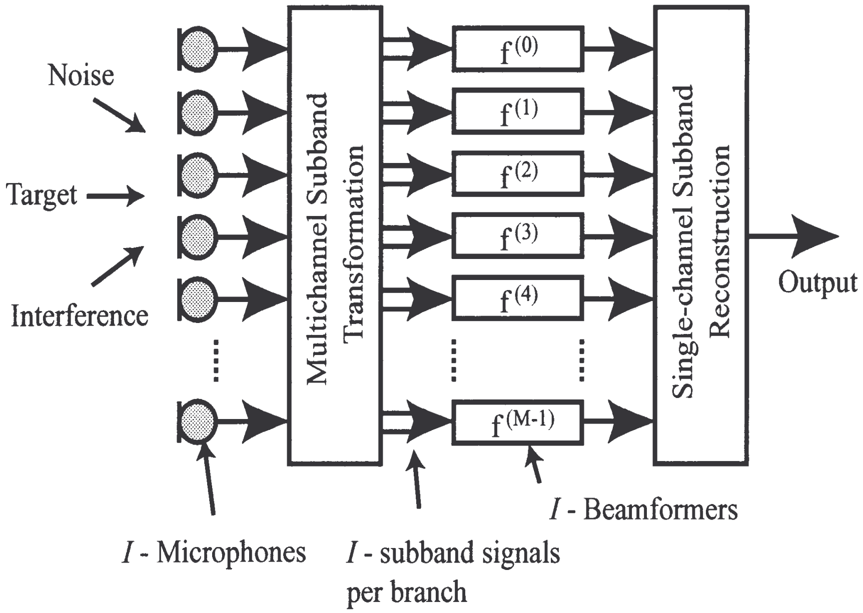

Noise
Target
Interference
Multichannel Subband Transformation
f (0)
f (1)
f (2)
f (3)
f (4)
f (M-1)
Single-channel Subband Reconstruction
Output
I - Microphones
I - subband signals per branch
I - Beamformers

Fig. 7.3 Subband FIR Beamforming Structure

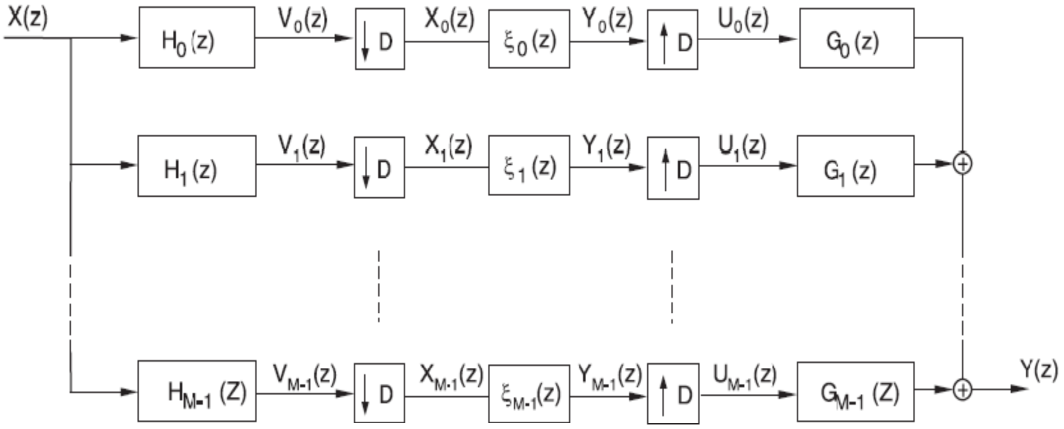


Fig. 7.4 Analysis and Synthesis Filter Banks with Subband Filtering

7.3.2 Multi microphone Environment

The source of spoken word is from the speaker (person). This speech will travel to all the microphones with different delays and gains depending on the distance between the speaker and microphone. The spoken speech s(n) is simulated to produce N directional sources such that they will be acquired by N different microphones. This is achieved using the amplitude coefficients A(s, k) and filter coefficients b(s, n). The objective of amplitude coefficients is to control the gain of speech sources to be added in speech received by particular microphone. Filter coefficients controls the delay and gain of particular directional source to be mixed with speech being acquired by particular microphone.

Here

'n' is number of microphones.

'k' is number of speech sources to be mixed with speech, being acquired by microphone, where speech sources are target speech, echo (interference) and noise.

's (n)' is spoken speech by speaker.

A(d, k) is k^{th} directional source, to be added with s^{th} microphone speech.

b(s, n) is speech to be filtered with coefficients set b, n=1: L coefficients to produce s^{th} directional speech. Thus, the speech Y_s received by sth microphone is given by equation 7.1

$$Y_s = f(s(n),A(d,k),b(s,n)) \tag{7.1}$$

7.4 Methodology for designing Beamforming based Speech Recognition

Methodology for designing beamforming based speech recognition is explained in next sub sections.

7.4.1 Beamforming

The signal obtained in each of the microphone is passed through the subband filter bank. The beamformers are formed by using the FIR adaptive filters, whose coefficients are determined by using the LMS algorithm. The beamformer filter is placed between each of analysis subband filter bank and each of microphone branch. This control the gain of each of the subband output from each microphone branch to be passed through the synthesis filter bank for each of the microphone line. The output of entire synthesis filter bank from each of the microphone line is added to form the reconstructed speech output.

7.4.2 Recognition

First of all, the features are extracted from the speech of spoken words. The feature Mel frequency cepstral coefficients (MFCC) have been proved to give better performance in case of speech recognition and hence widely used in speech recognition applications [18-20]. In speech processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a speech, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. The recognition process consists of training the classifier and testing the spoken words with trained classifier. The classier used here is nearest neighbor classifier (NN) based on Euclidian distance metric.

7.5 Experimental Results for Beamforming based Speech Recognition

For analysing the performance of speech recognition, we have considered here four speaker's 20 number of spoken words. These words are listed below and can be categorized on the basis of their use, as numbers and commands. The spoken word from speaker has a length of 2 sec in time. The speech to be used in the experiment is created using multi-microphone mixing environment as described earlier.

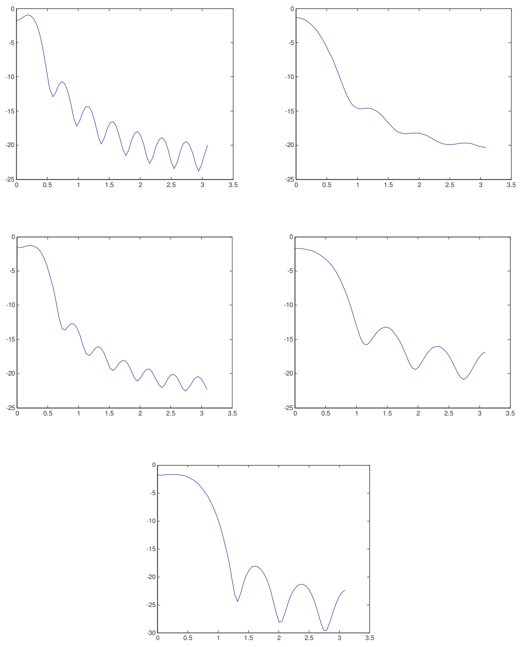


Fig.7.5 Frequency Response of Prototype filter with different specifications (no of subbands – filter length): Row-1-Col 1) 16 -16; Row-1-Col-2) 16-8; Row-2-Col-1) 8-16; Row-2-Col-2) 8-8 ; Row-3) 4-8.

The prototype filter was designed to construct the filter bank. The frequency spectrums of prototype filter for different length of filters and different numbers of sub-bands are shown in figure 7.5.

For training classifier, 2 speakers's spoken words used and for testing we used 4 speakers, wherein 2 speakers are unknown and 2 speakers are same as they were in training phase. Each person (speaker) has 20 spoken words, which includes 10 words for numbers and 10 words for commands as listed in table 7.1. The experiments are performed separately with following class of words:

- Numbers and Commands together (20 words)
- Numbers only (10 words)
- Commands only (10words)

The recognition accuracy is calculated as the ratio of correctly recognised words and total words used for recognition test experiment. We have used MATLAB® environment for performing all experiments.

For each class of experiment, the recognition accuracy is calculated in three scenarios. First when pure speech is feed to the recognition experiment without any noise and interference. Secondly, speech was prepared with multi-mic environment with an inclusion of noise and interference (echo). Finally, using beamforming multi-mic speech is enhanced with beamforming-filter bank structure and then fed to the recognition experiment. The last three columns of each of the following tables showing recognition accuracy represents the performance obtained in these three situations. In order to analyse the speech recognition performance against the parameters of filter-bank, we selected the two parameters: filter length and number of subbands in filter bank. The experiments were repeated for different values of these two parameters as mentioned in the first two columns of following tables.

Table 7.1

Recognition Accuracy with and without Beamforming for numbers and commands together

Filter length	Number of subbands	Recognition Accuracy in Percentage		
		Pure speech	Multi-Microphone speech	Multi-mic Beamformed speech
16	16	75	27.5	30
8	16	75	27.5	32.5
16	8	75	27.5	31.25
8	8	75	27.5	21.25
8	4	75	27.5	25

Table 7.2

Recognition Accuracy with and without Beamforming for numbers only

Filter length	Number of subbands	Recognition Accuracy in Percentage		
		Pure speech	Multi-Microphone speech	Multi-mic Beamformed speech
16	16	72.5	35	40
8	16	72.5	35	45
16	8	72.5	35	52.5
8	8	72.5	35	27.5
8	4	72.5	35	40

Table 7.3

Recognition Accuracy with and without Beamforming for commands only

Filter length	Number of subbands	Recognition Accuracy in Percentage		
		Pure speech	Multi-Microphone speech	Multi-mic Beamformed speech
16	16	82.5	50	52.5
8	16	82.5	50	50
16	8	82.5	50	37.5
8	8	82.5	50	62.5
8	4	82.5	50	35

7.6 Discussion

The recognition accuracy obtained in various experiments is shown in tables 7.1, 7.2 and 7.3. The main objective of the speech enhancement is to bring up the speech recognition performance in the presence of noise and echo-interference to the performance obtained with pure speech signals, which is the ideal case. Thus our aim was to boost up the performance of beamforming based speech enhancement to that in the case of ideal signal. It can be observed that from table 7.1, the speech recognition performance can be improved using the beam forming based speech enhancement. This is also visible in other two experiments, table 7.2 and 7.3, where only numbers and only commands were used for speech recognition. These three cases observations are depicted as below;

1. Numbers + Commands: In the case of numbers and commands together in recognition experiment, accuracy reduces to 27.5% from ideal value 75% due to noise and echo mixing. Using beam forming, the degraded performance can be improved to the optimized performance 32.5%. This is a significant improvement in the recognition accuracy.
2. Numbers: In this case, interference due to noise and echo mixing causes decrease in recognition performance to 35% from ideal value 72.5%. Using beam forming, the degraded performance can be boost up to the optimized performance of 52.5%. This is a much significant improvement in the recognition accuracy.
3. Commands: In this case, due to noise and echo mixing, recognition performance degrades to 50% from ideal value 82.5%. Using beam forming, the degraded performance can be brought up to the optimized performance 62.5%. This is a much significant improvement in the recognition accuracy.

These results are in consistent with the fact that proper beamforming can improve the recognition performance. Since, this improvement can be achieved with less computational parameters of sub-band filtering; this technique is suitable for real-time application of speech recognition.

The performance is also dependent on the parameters of filter bank used for sub-band filtering. Another objective of this work is to analyse the effect of the filter-bank parameters on the speech recognition. It can be seen easily that there is an undesired effect of improper selection of the parameters like filter length and number of subbands on the recognition accuracy. In general, more the number of subbands, more parallelism can be achieved by the system and larger the filter length, better the frequency response but with higher computational complexity. Thus, it is important to design the system with proper selection of these parameters so that system yield can be improved. For the all word experiments, it can be seen that best performance we get, more precisely, is with filter length 8 and number of subbands 16. However, roughly, it can be seen that the filter length and number of subbands with values 8 or 16 giving better results. The same conclusion can be inferred from other two experiments also.

Another important point that can be observed here is that for numbers type speech and commands type of speech, different parameters of sub-band filtering are required. This is due to the fact that numbers are normally pronounced with short duration support and commands are comparatively long duration speech words. Hence, having optimized parameters in general application is required. The highest performance in both individual experiments (table 7.2 and 7.3), has different lengths, 16 and 8. However, second highest performance with beamforming is with filter length 8 and number of sub-bands 16, in both types of experiments. Thus these parameters can be selected as optimized parameters for general application.

7.7 Conclusion

The main objective of the speech enhancement is to bring up the speech recognition performance in the presence of noise and echo-interference to the performance obtained with pure speech signals, which is the ideal case. Thus, our aim was to approach the performance of beamforming based speech enhancement to that in the case of ideal signal. Another objective of this work is to analyse the effect of the filter-bank parameters on the speech recognition. It can be seen easily that there is an undesired effect of improper selection of the parameters like filter length and number of subbands on the recognition accuracy. The experiments were performed for 20 words including numbers and commands, 10 words of numbers only and 10 words of commands only for different values of filter bank parameters. The results obtained have proved the speech enhancing capability of the beamforming technique in multi-microphone network where noise and echo-interference can degrade the original speech signal. Since, this improvement can be achieved with less computational-intensive parameters of sub-band filtering; this technique is suitable for real-time application of speech recognition.

CHAPTER 8

EVOLUTIONARY COMPUTATIONS BASED REAL TIME SPEECH BEAMFORMING FOR MULTIMEDIA APPLICATIONS

8.1 Introduction

In this chapter, we have presented the approach of evolutionary computation in form of genetic algorithm to select the features that are responsible for discriminating the different words. In doing so, the amount of feature elements to be used also gets reduced and hence system can be made to recognise the word-speech with real-time performance. The system is made to be working in real-time as time required for classifier has been reduced dramatically. This is particularly achieved by including the zeros at random places and in random amount in initial population chromosomes, which were generated randomly in the range of 0 to 1. This results in the reduction of feature elements in feature descriptor and have feature vector length. This is especially an important requirement in the mobile devices where power, memory and processing power are available with large constraints. The in-car infotainment and mobile devices are the potential examples of real-time constraint requirements. The experiments were carried out different filter-bank parameters; filter length and number of subbands. The experiments were performed separately for 20 words including numbers and commands, 10 words of numbers only and 10 words of commands only for different values of filter bank parameters. The results obtained have proved the speech enhancing capability in the beamforming based speech recognition system using genetic algorithm. In beamforming simulated speech, multi-microphone network was generated with noise and echo-interference, which can degrade the original speech signal. The performance of multimedia systems is greatly improved if beamforming based speech enhancement is used for speech recognition using soft computing techniques in real time mode.

8.2 Existed Work Related to Genetic based Optimization in Speech Recognition

There are attempts made by researchers for employing computational intelligence and evolutionary computations in the speech recognition system. Especially, genetic algorithm has been applied in various techniques of speech recognition. In [1], author proposes a genetic algorithm (GA) based beamformer in which beamformer weights are optimized with the help of GA operators, crossover and mutation. It is claimed that this GA based optimization is successful in tackling the non-differentiable and non-linear natures of speech recognition in normal and noisy environment.

Another category of GA application in speech recognition is to select the optimized feature such that it will improve the performance of recognition. The generic sound recognition system that exploits evolutional algorithms for a selection of discriminative acoustic features has been presented in [2]. Similar kinds of objectives have been achieved in [3-5]. In [6] and other works, feature set itself, in form of codebook dictionary such as in vector quantization, have been optimized.

Apart from optimizing feature set, classifier like multilayer perceptron based neural network was optimized using genetic algorithm in [7-8]. Another classifier HMM was also made to give optimal performance with the help of GA in [9].

8.3 Methodology for designing GA based Optimisation Technique

In order to analyse the performance of the speech recognition in the speech corrupted by noise and echo-interference, the analysis frame work used here is taken from [10]. The subband FIR filter bank scheme and the analysis-synthesis filter bank structure explained in previous chapter is used for beamforming. The multi-microphone network for beamforming based speech recognition was simulated as described in [11]. The methodology for beamforming and recognition explained in previous chapter is used with GA based optimization for speech recognition. Next paragraph explain the GA based optimisation for speech recognition.

There are numerous attempts made by researchers for evolutionary model in the speech recognition system. Especially, genetic algorithm has been applied in various techniques of speech recognition. The genetic algorithm can be used at two levels. First, the feature elements selection level, where important features are preserved while ignoring remaining, can employ the genetic algorithm. In some cases, the dictionary of features is generated using genetic algorithm as in case of vector quantization codebook [6]. Secondly, the GA can be used at classifier level to determine its optimized parameters. For examples, in neural network the number of hidden layers and number of nodes in each of them can be determined effectively using the evolutionary computations. In case of Hidden Markov Model (HMM), states and state transition parameters can be decides using GA.

The standard MFCC features are very sensitive to additive noise and channel mismatch, therefore the recognition accuracy deteriorates drastically in noisy environments. Thus it is important to remove or suppress the effect of the feature elements which are sensitive to the noise and echo to optimize the recognition performance in high noisy environment. This optimization problem can be handled using genetic algorithm.

The genetic algorithm is applied to the recognition problem of speech words with the objective to find important feature elements that contributes more to classifier for distinguishing one word from others. Additionally, the number of feature elements is also reduced eland into the reduction in the length of feature vector for further step of classification. The chromosome of GA was of length as same as that of feature vector. This chromosome has real value between 0 and 1, randomly generated at each position, in its first form and then it is modified by making 0 to the positions that has lesser value than some randomly generated value between 0 and 1. This modification helped in bringing the wide range variations in usable percentage of total feature elements. This enabled to evaluate the chromosomes' performance of recognition with having even small percentage of elements. This chromosome was multiplied (element wise) with feature vector to be optimized, before using it for recognition.

8.4 Experiment Results for GA based OptimisationTechnique

Four speaker's 20 number of spoken words is considered for experiment conducted for GA based Optimisation. List of spoken words is given in previous chapter. The speech to be used in the experiment is created using multi-microphone mixing environment as described in [10]. Experiment set up used here is same as explained earlier in beamforming based speech recognition implementation.

For training classifier, 2 speakers's spoken words used and for testing we used 4 speakers, wherein 2 speakers are unknown and 2 speakers are same as they were in training phase. Each person (speaker) has 20 spoken words, which includes 10 words for numbers and 10 words for commands. The experiments are performed separately with following class of words:

- Numbers and Commands together (20 words)
- Numbers only (10 words)
- Commands only (10words)

The recognition accuracy is calculated as the ratio of correctly recognised words and total words used for recognition test experiment. The above set of experiments was performed twice; firstly without optimization and secondly with GA based optimization.

In first set of experiments, for each class of experiment, the recognition accuracy is calculated in three scenarios. First when pure speech is feed to the recognition experiment without any noise and interference. Secondly, speech was prepared with multi-mic environment with an inclusion of noise and interference (echo). Finally, using beamforming multi-mic speech is enhanced with beamforming-filter bank structure and then fed to the recognition experiment. Experiments were conducted separately with and without GA optimization and results are presented in the table 8.1, 8.2 and 8.3.

Table 8.1

Multi-mic Beamformed Speech Recognition Accuracy for numbers and commands together

Filter length	**Num of subbands**	**Multi-mic Beamformed Speech Recognition Accuracy in Percentage**			
		Without GA[85]	**Best Solution with GA**	**Percentage improvement with GA**	**With Least Features**
16	16	46.25	51.25	10.8	50.00
8	16	40.00	46.25	15.6	46.25
16	8	42.50	50.00	17.6	48.75
8	8	48.75	56.25	15.4	53.75
8	4	45.00	51.25	13.9	50.00

Table 8.2

Multi-mic Beamformed Speech Recognition Accuracy for numbers only

Filter length	**Num of subbands**	**Multi-mic Beamformed Speech Recognition Accuracy in Percentage**			
		Without GA[85]	**Best Solution with GA**	**Percentage improvement with GA**	**Least Features GA**
16	16	55.00	65.00	18.2	62.50
8	16	47.50	60.00	26.3	57.50
16	8	50.00	62.50	25.0	62.50
8	8	50.00	62.50	25.0	60.00
8	4	60.00	67.50	12.5	67.50

Table 8.3

Multi-mic Beamformed Speech Recognition Accuracy for commands only

Filter length	Num of subbands	Multi-mic Beamformed Speech Recognition Accuracy in Percentage			
		Without GA[85]	Best Solution with GA	Percentage improvement with GA	Least Features GA
16	16	52.50	62.50	19.0	60.00
8	16	50.00	57.50	15.0	57.50
16	8	45.00	55.00	22.2	50.00
8	8	60.00	70.00	16.6	70.00
8	4	52.50	62.50	19.0	60.00

The two parameters: filter length and number of subbands in filter bank is selected to analyse the speech recognition performance against the parameters of filter-bank. The experiments were repeated for different values of these two parameters as mentioned in the first two columns of tables 8.1, 8.2 and 8.3. While performing the experiments with GA based optimization, the feature vector was modified by the each chromosome from the population using inner product operator. Then fitness value as calculated for each of the modified feature vector. In fitness function calculation, recognition ratio is calculated with kNN classifier with two subjects' words in training set and remaining two for testing. The parameters for GA algorithm are:

- Initial Population, 200
- Selected Population in each iteration, 100
- Generated Population using operators, 200
- Elitism, 2 %
- Mutation Rate, 2%
- Uniform crossover rate, 98%

Table 8.4
Performance in both criteria in terms of recognition accuracy and Feature Vector length for numbers and commands together

Filter Length	Number of Sub-band	Command + Numbers			
		Best Recognition Solution		Least Feature Elements Solution	
		Recognition	FV Size	Recognition	FV Size
16	16	51.25	8.04	50.00	2.24
8	16	46.25	0.78	46.25	0.78
16	8	50.00	5.79	48.75	1.46
8	8	56.25	7.01	53.75	1.85
8	4	51.25	5.55	50.00	1.07

Table 8.5
Performance in both criteria in terms of recognition accuracy and Feature Vector length for numbers only

Filter Length	Number of Sub-band	Numbers			
		Best Recognition Solution		Least Feature Elements Solution	
		Recognition	FV Size	Recognition	FV Size
16	16	65.00	1.60	62.50	1.07
8	16	60.00	2.87	57.50	0.63
16	8	62.50	1.31	62.50	0.68
8	8	62.50	6.14	60.00	0.68
8	4	67.50	5.70	67.50	2.04

Table 8.6
Performance in both criteria in terms of recognition accuracy and Feature Vector length for commands only

Filter Length	Number of Sub-band	Commands			
		Best Recognition Solution		Least Feature Elements Solution	
		Recognition	FV Size	Recognition	FV Size
16	16	62.50	3.43	60.00	1.34
8	16	57.50	4.47	57.50	0.49
16	8	55.00	2.45	50.00	0.49
8	8	70.00	6.37	70.00	1.40
8	4	62.50	9.06	60.00	0.98

8.5 Discussion

It has been observed that GA based optimization was converging in each of the iteration in the sense that mean fitness value of the population in each of the iterations was monotonically increasing. This proves the fact that GA optimization was approaching to find out the optimal solution with the help of GA operations like crossover, mutation in addition to the elitism property of evolution. The graph of fitness value for each of the iteration is shown in figure 8.1.

The recognition accuracy obtained in various experiments is shown in tables 8.1, 8.2 and 8.3. The main objective of the speech enhancement is to bring up the speech recognition performance in the presence of noise and echo-interference to the performance obtained with pure speech signals, which is the ideal case. Thus our aim was to boost up the performance of beamforming based speech recognition. Not just that but, performance can be further improved by using GA based optimization. Thus, these tables show the performance with and without GA based optimization.

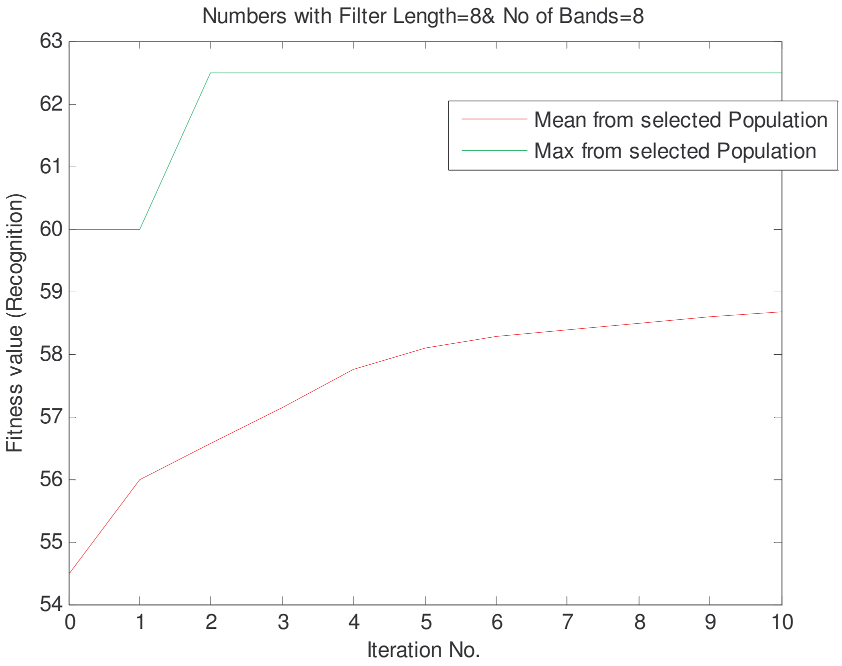


Fig. 8.1.The graph of fitness value for each of the iteration

From the observation of first and second columns of recognition accuracy in each of these tables, it is clear that recognition performance with GA optimized feature vector has been significantly improved. It can be observed that from table 8.1, the speech recognition performance for commands and numbers both can be improved using GA based optimization in the beam forming based speech enhancement. This is also visible in other two experimental results in table 8.2 and 8.3, where only numbers and only commands were used for speech recognition. In addition to this, the best solution with least features elements in last iteration is given in third column of recognition accuracy in each of the tables. The important inference in this point of view is that the recognition performance with least feature elements is not deterrent from the best solution. In other words, we can say that optimal recognition performance with best solution and least amount of feature elements (shortest feature vector) can be easily obtained by the GA optimization, which can be complemented for the real-time computation of classification. These three cases observations in particular are depicted as below:

1. Numbers + Commands: In the case of numbers and commands together in recognition experiment, GA optimization based speech recognition is improved by an average of 15%. The least feature vector length solution also gives similar performance with that of best of solution.

2. Numbers: In case of number recognition, the GA optimization gives very optimal performance and it improves recognition by a factor as high as 26%. This is significant improvement with the additional fact that this improvement can be obtained with shortest feature vector.
3. Commands: In this case, average improvement with GA optimization was around average of 20% with all parameters of filter-bank based beamforming.

The results with percentage of feature elements required to get optimal solution are presented in tables 8.4, 8.5 and 8.6. The length of feature vector can be reduced as low as 0.5%, saving almost 99% computational power and memory with optimized solution. It is interesting to observe that even with least number of feature elements optimal solution best recognition can be obtained. This fact proves that there so many unnecessary feature elements that are redundant to representation of the word-speech signal. Additionally, there are also feature elements that are sensitive to the noise and echo, removing which performance gets boost up. Another most important shorter feature vector is that in memory required for gallery samples required less and in classification stage computational complexity reduces.

The amount of feature elements that can be used in classification is an important factor, especially in the case of devices with low power(battery operated), low memory and less computational power as in case of mobile hand-held devices. With the smaller size feature vector and yet optimal in recognition performance will take less gallery features. The classifier will take lesser number of computations that is required with full feature vector. This will also increase the speed of application processing with cheaper hardware, leading to the economical cost of the embedded product.

The speech based applications have been always important in communication for the humans. Most recently, speech based interface has been tried to be employed in almost all the mobile and stationary devices. However, these attempts could not give ultimate response due to variations in surrounding noises, changes in person to person speech and also intra person variation. This scenario leads to further research that will make speech recognition more robust and general. It can be applied upcoming electronic devices to be sued for various multimedia applications like gaming, entertainment and cellular phones.

The performance of multimedia systems is greatly improved if beamforming based speech enhancement is used for speech recognition using soft computing techniques in real time mode[7] [10-11].

8.6 Conclusion

In this chapter, we have presented the approach of evolutionary computation in form of genetic algorithm to select the features that are responsible for discriminating the different words. In doing so, the amount of feature elements to be used also gets reduced and hence system can be made to recognise the word-speech with real-time performance. The system is made to be working in real-time as time required for classifier has been reduced dramatically. This is particularly achieved by including the zeros at random places and in random amount in initial population chromosomes, which were generated randomly in the range of 0 to 1. This results in the reduction of feature elements in feature descriptor and have feature vector length. This is especially an important requirement in the mobile devices where power, memory and processing power are available with large constraints. The in-car infotainment and mobile devices are the potential examples of real-time constraint requirements. The experiments were performed for 20 words including numbers and commands, 10 words of numbers only and 10 words of commands only for different values of filter bank parameters. The results show the effectiveness of the GA optimization in all the subsets of experiments with different parameters of beamforming. The length of feature vector can be reduced as low as 0.5%, saving nearly 99% computational power and memory with optimized solution, leading to a one of the approach to be used in real-time embedded system devices for speech recognition applications.

[7] Paper published on " **Beamforming based Speech Recognition using Genetic Algorithm for Real Time systems**", *International Journal of Recent Technology and Engineering (IJRTE)*, Paper ID:B0618052213, ISSN:2231-2307, Pages 96-104, Volume 2, Issue 2, May 2013.

CHAPTER 9

REAL TIME IMPLEMENTATION OF SPEECH RECOGNITION

Speech recognition is an important field of digital signal processing. Automatic Speaker Recognition (ASR) objective is to extract features, characterize and recognize speaker. Mel Frequency Cepstral Coefficients (MFCC) is most widely used feature vector for ASR. MFCC is used for designing a text dependent speaker identification system. Here the DSP processor TMS320C6713 from Texas Instruments with Code Composer Studio (CCS) has been used for real time speech recognition. DSK 6713 from Spectrum Digital Incorporation is used for implementing algorithm on the TMS320C6713 DSP. The Code Composer Studio Integrated Development Environment version 3.3 (CCS IDE V3.3) from Texas Instruments is used as compiler and debugger. For analysing the performance of speech recognition, we have considered here four speaker's 20 number of spoken words as numbers and commands of length 2 sec in time. It is also investigates, how MFCC algorithm extracts features and how ED from all training vectors is calculated using GMM. MFCC algorithm calculates cepstral coefficients of Mel frequency scale. Each Euclidian Distance (ED) from all training vectors is calculated using Gaussian Mixture Model (GMM) as it gives better recognition for the speaker features. The command/voice having minimum ED is applied as similarity criteria.

9.1 Introduction

Speech recognition is an important field of digital signal processing. There are various objectives for the development of Automatic Speech Recognition (ASR). Main objective of ASR is to extract features, characterize and recognize speaker. The application can be aimed at recognition to be performed either on isolated words or utterances or on continuous speech. There are various languages spoken in this world that makes to consider the one of the language for the recognition system. There are also situations, when recognition system should be speaker dependent or independent. The most difficult class of recognition system is to develop speaker independent recognition on continuous speech. This needs the inclusion of knowledge about the application for which system to be built in addition to the word recognition system. Typically, the first step in this kind of system is always word recognition for the limited number of words.

L. Rabiner, B.H. Juang and B. Yegnanarayana presented the approach of simple speech recognition system [1]. It consists of four main building blocks speech analysis, feature extraction, language translation and message understanding. Speech analysis stage consists of noise removal, silence removal and end point detection. End point detection and removal of noise, silence is required to improve the performance of speech recognition system. Noisy speech processes along the basilar membrane in the inner ear, which provides spectrum analysis of noisy speech. The speech analysis also deals with suitable frame size for segmenting speech signal for further analysis using segmentation, sub segmental and supra segmental analysis techniques [2].

Feature extraction and coding stage reduces the dimensionality of the input vector and maintain discriminating power of the signal. We need feature extraction because the number of training and test vector needed for the classification problem grows with the dimension of the given input. Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC) are the most widely used methods for feature extraction. MFCC preferred over LPC because it is less prone to noise. The spectral signal output of speech analysis converted to activity signals on the auditory nerve using neural transduction method. Then activity signal converted into a language code within the brain, and finally message understanding is achieved.

Mel Frequency Cepstral Coefficients (MFCC) is most widely used feature vector for ASR [3] and this feature has been used in this paper. MFCC is used for designing a text dependent speaker identification system. Gaussian Mixture Model (GMM) [4] has been widely used as speaker model because it gives better recognition for the speaker features. DSP starter kit TMS320C6713 has been used in various applications and the features, which make it suitable, are faster data access, data transfer to and from real world, computation, execution control and numerical fidelity.

9.2 Hardware Implementation Tools

Specific hardware implementation tools is require for testing and embedding speech processing algorithm on dedicated DSP platform. Real time digital signal processing made considerable advancements after the introduction of specialized DSP processors. Suitable starter kits with a specific DSP processor and related software tools such as compilers, assemblers, simulators, debuggers, and so on, are provided in order to make system design and application development easier. The 32-bit floating point processor TMS320C6713 from Texas Instruments is very powerful for real time speech and audio processing algorithm implementations. This DSP processor is based on the VLIW (Very Large Instruction Word) technology, which allows fast parallel computing jointly using its optimized "C" compiler. For a rapid evaluation of the TMS320C6713 processor a Developer Starter Kit 6713 (DSK 6713) is available from Spectrum Digital Incorporation; comprises a board and the software tools. The board must be connected to a standard PC running under its integrated development environment- Code Composer Studio (CCS IDE).

9.2.1 TMS320C6713 DSK

The TMS320C6000 platform of digital signal processors (DSPs) is part of the TMS320 family of DSPs. The TMS320C67xx (C67x) devices are floating-point DSPs in the TMS320C6000 platform. The TMS320C67x DSPs (including the TMS320C6713 device) compose the floating-point DSP generation in the TMS320C6000 DSP platform [5-6]. The C6713 device is based on the high-performance, advanced very-long-instruction-word (VLIW) architecture developed by Texas Instruments (TI), making this DSP an excellent choice for multichannel and multifunction applications. Operating at 225 MHz, the C6713 delivers up to 1350 million floating-point operations per second (MFLOPS), 1800 million instructions per second (MIPS), and with dual fixed-/floating-point multipliers up to 450 million multiply-accumulate operations per second (MMACS). Operating at 300 MHz, the C6713 delivers up to 1800 million floating-point operations per second (MFLOPS), 2400 million instructions per second (MIPS), and with dual fixed-/floating-point multipliers up to 600 million multiply-accumulate operations per second (MMACS). The TMS320C6713 device has two boot modes: from the HPI or from external asynchronous ROM.

The C6713 has a rich peripheral set that includes two Multichannel Audio Serial Ports (McASPs), two Multichannel Buffered Serial Ports (McBSPs), two Inter-Integrated Circuit (I^2C) buses, one dedicated General-Purpose Input/Output (GPIO) module, two general-purpose timers, a host-port interface (HPI), and a glue less external memory interface (EMIF) capable of interfacing to SDRAM, SBSRAM, and asynchronous peripherals. The two McASP interface modules each support one transmit and one receive clock zone. Each of the McASP has eight serial data pins, which can be individually allocated, to any of the two zones. The serial port supports time-division multiplexing on each pin from 2 to 32 time slots. The C6713B has sufficient bandwidth to support all 16 serial data pins transmitting a 192 kHz stereo signal. Serial data in each zone may be transmitted and received on multiple serial data pins simultaneously and formatted in a multitude of variations on the Philips Inter-IC Sound (I^2S) format. In addition, the McASP transmitter may be programmed to output multiple S/PDIF, IEC60958, AES-3, CP-430 encoded data channels simultaneously, with a single RAM containing the full implementation of user data and channel status fields. The McASP also provides extensive error checking and recovery features, such as the bad clock detection circuit for each high-frequency master clock, which verifies that the master clock is within a programmed frequency range. The two I^2C ports on the TMS320C6713 allow the DSP to easily control peripheral devices and communicate with a host processor. In addition, the standard multichannel-buffered serial port (McBSP) may be used to communicate with serial peripheral interface (SPI) mode peripheral devices. The TMS320C6713 device has two boot modes: from the HPI or from external asynchronous ROM. The TMS320C67x DSP generation is supported by the TI eXpressDSP - set of industry benchmark development tools, including a highly optimizing C/C++ Compiler, the Code Composer Studio-Integrated Development Environment (IDE), JTAG-based emulation and real-time debugging, and the DSP/BIOS kernel.

DSK 6713 key features includes

- A TI TMS320C6713 DSP operating at 225 MHz.
- An AIC 23 stereo codec.
- 4 user LEDs and 4 DIP switches.
- 16 MB SDRAM and 512 KB non-volatile Flash memory.
- Software board configuration through registers implemented in CPLD.
- JTAG (Joint Test Action Group) emulation through on-board JTAG emulator with USB host interface or external emulator.
- Single voltage power supply (+5V).

The block diagram describing the board is shown in figure 9.1

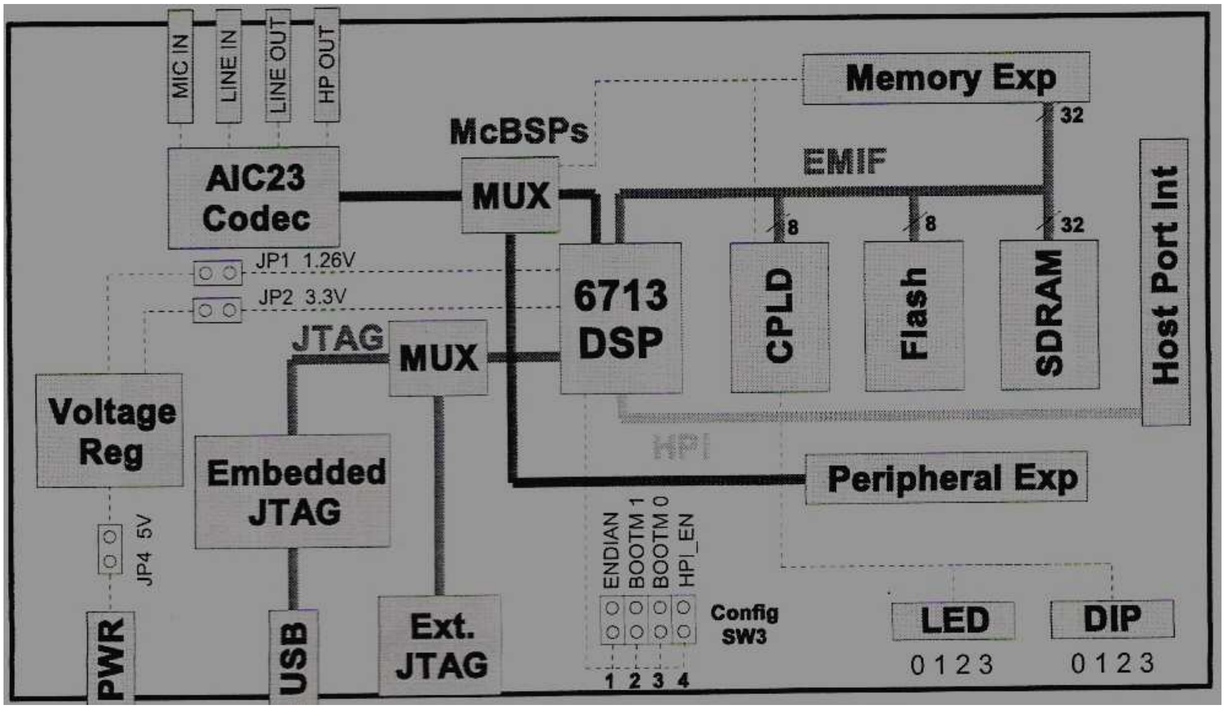


Fig.9.1 TMS320C6713DSK block diagram

9.2.2 Code Composer Studio (CCS)

The Code Composer Studio (CCS) application provides an integrated environment with the following capabilities [7]:

- Integrated development environment (IDE) with an editor, debugger, project manager, profiler, etc.
- ‘C/C++’ compiler, assembly optimizer and linker (code generation tools).
- Simulator.
- Real-time operating system (DSP/BIOS).
- Real-Time Data Exchange (RTDX) between the Host and Target.
- Real-time analysis and data visualization.

The CCS Project Manager organizes files into folders for source files; include files, libraries and DSP/BIOS configuration files. Once the files are added to the project any changes in any of source files will be reflected automatically in the project files. This allows multi user system development. CCS also provides the ability to debug mixed, multi-processor designs simultaneously. It also includes new emulation capabilities with Real Time Data Exchange (RTDX), plus advanced DSP code profiling capabilities. An improved Watch Window monitors the values of local and global variables and C/C++ expressions. Users can quickly view and track variables on the target hardware. It has ability to share C and C++ source and libraries in a multi-user project. Figure 9.2 shows working model of code composer studio.

The TMS320C67x DSP generation is supported by the TI eXpressDSP-set of industry benchmark development tools, including a highly optimizing C/C++ Compiler, the Code Composer Studio Integrated Development Environment (IDE), JTAG-based emulation and real-time debugging, and the DSP/BIOS kernel. CCS offers robust core functions with easy to use configuration and graphical visualization tools for system design. Programming in C/C++ for application is complied, linked and executed by the CCS. Figure 9.3 shows the programming interface of CCS.

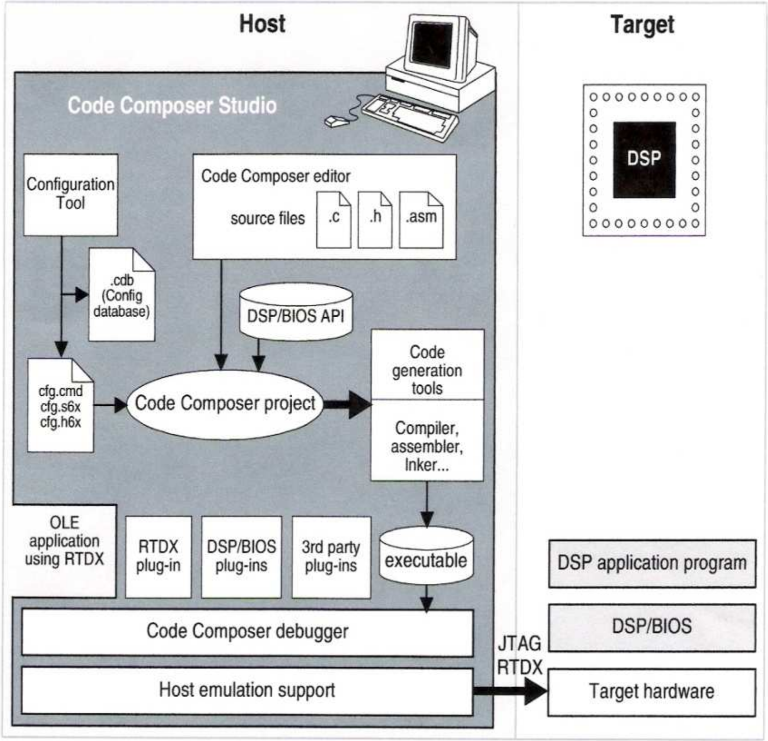
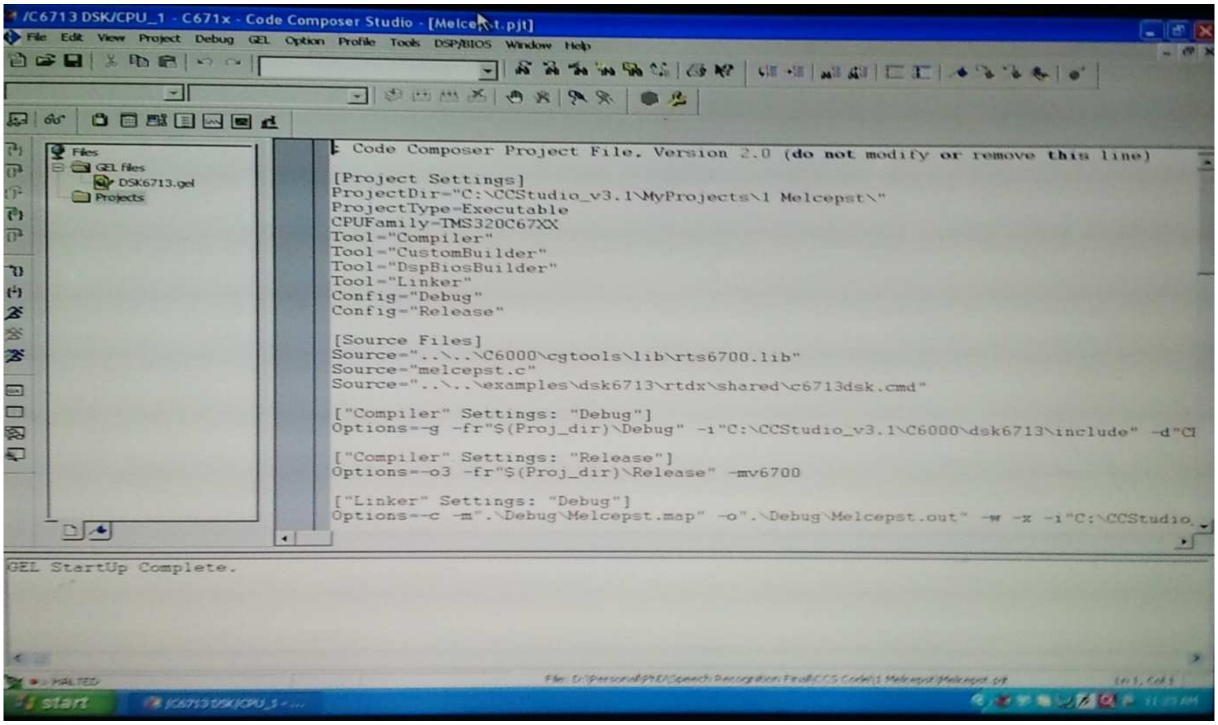


Fig. 9.2 Code Composer Studio working Model


/C6713 DSK/CPU_1 - C671x - Code Composer Studio - [Melcepst.pjt]
File Edit View Project Debug GEL Option Profile Tools DSP/BIOS Window Help
Files
GEL files
DSK6713.gel
Projects
start
/C6713 DSK/CPU_1 - ...
HALTED

```
Code Composer Project File, Version 2.0 (do not modify or remove this line)

[Project Settings]
ProjectDir="C:\CCStudio_v3.1\MyProjects\1 Melcepst\"
ProjectType=Executable
CPUFamily=TMS320C67XX
Tool="Compiler"
Tool="CustomBuilder"
Tool="DspBiosBuilder"
Tool="Linker"
Config="Debug"
Config="Release"

[Source Files]
Source="..\..\C6000\cgtools\lib\rts6700.lib"
Source="melcepst.c"
Source="..\..\examples\dsk6713\rtdx\shared\c6713dsk.cmd"

["Compiler" Settings: "Debug"]
Options=-g -fr"$(Proj_dir)\Debug" -i"C:\CCStudio_v3.1\C6000\dsk6713\include" -d"CH

["Compiler" Settings: "Release"]
Options=-o3 -fr"$(Proj_dir)\Release" -mv6700

["Linker" Settings: "Debug"]
Options=-c -m".\Debug\Melcepst.map" -o".\Debug\Melcepst.out" -w -x -i"C:\CCStudio
```

```
GEL StartUp Complete.
```

Fig. 9.3 The programming interface of CCS

9.3 Hardware Setup for Developing Models

Figure 9.4 shows hardware set up block diagram for developing model [8]. Figure 9.5 shows actual connection between host PC and target board. The audio input is applied to Line-in of DSK TMS320C6713, which is taken from VLC player via host PC. For simplicity of operation mono channel is used. Signal processing performed in DSK with the help of "C" code downloaded in it which generates required speech recognized output. The output is taken from the "Headphone out" of the DSK and then it is applied to speaker.

For analysing the performance of speech recognition, we have considered here four speaker's 20 number of spoken words. Since these words are regularly used in every human's life, we have chosen these words. These words are listed in chapter 6 and can be categorized on the basis of their use, as numbers and commands.

MATLAB is used to convert recorded WAV files to DAT file. These data files are then used as input to DSK board. In next operation features are extracted from the data files and then average features are chosen for real time speech recognition experiment.

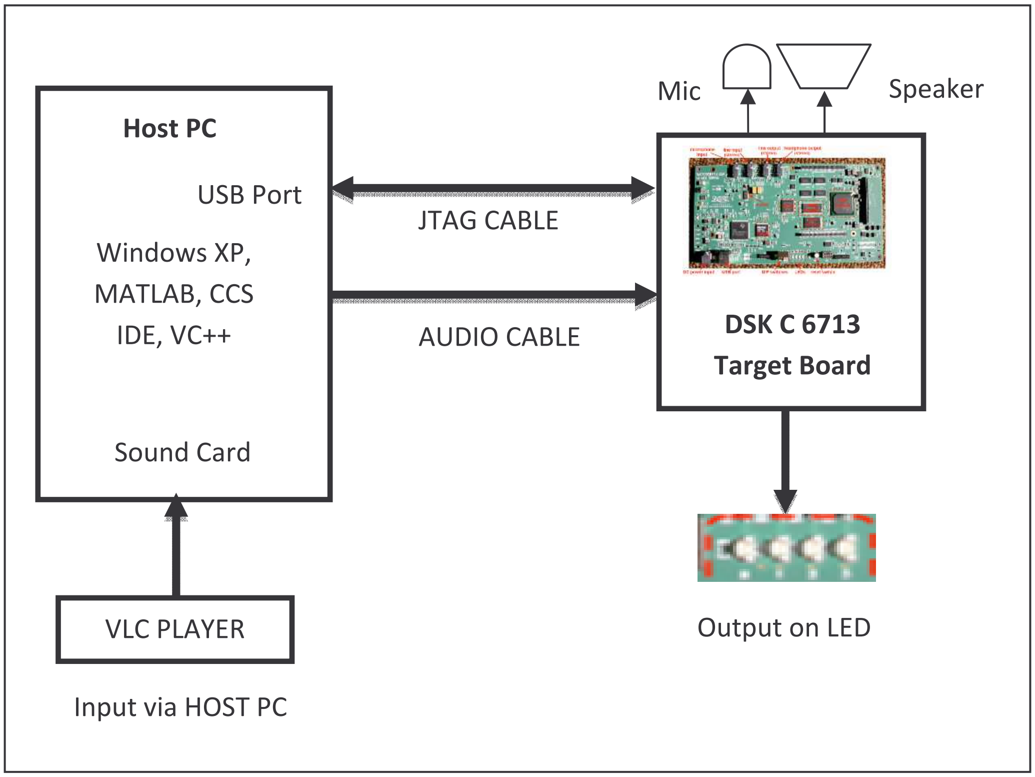


Fig. 9.4 Hardware set up block diagram for developing model

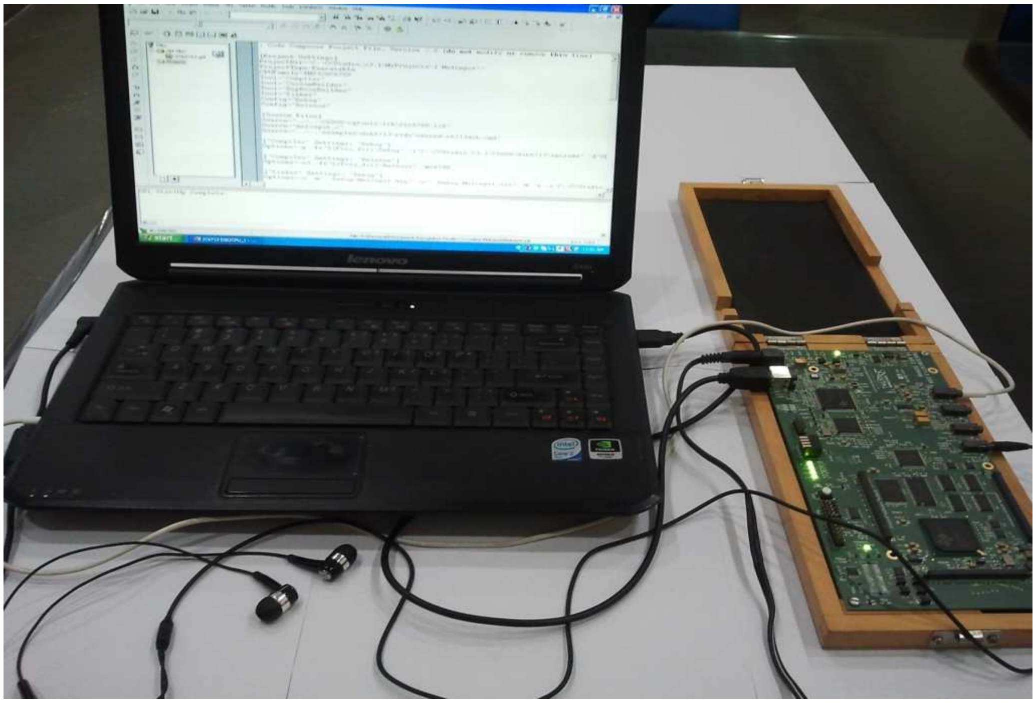


Fig. 9.5 Actual connection between host PC and target board

9.4 Feature Extraction using Mel Frequency Cepstral Coefficients

Many experiment has shown that the ear’s perception to the frequency components in the speech does not follow the linear scale but the Mel-frequency scale, which should be understood as a linear frequency spacing below 1kHz and logarithmic spacing above 1kHz . So filters spaced linearly at low frequency and logarithmic at high frequencies can be used to capture the phonetically important characteristics of the speech. The aim of this work is classification of voice given by the user into the predefined commands in training set. This classification is done with the help features which are in the form of Mel-Cepstral Coefficients.

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behaviour. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear’s critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech [9]. The overall process of the MFCC is shown in figure 9.6.

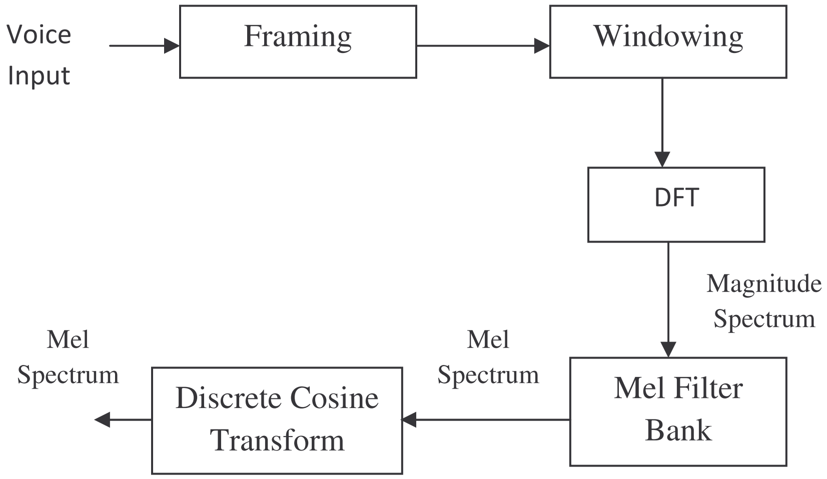


Fig. 9.6 MFCC Process

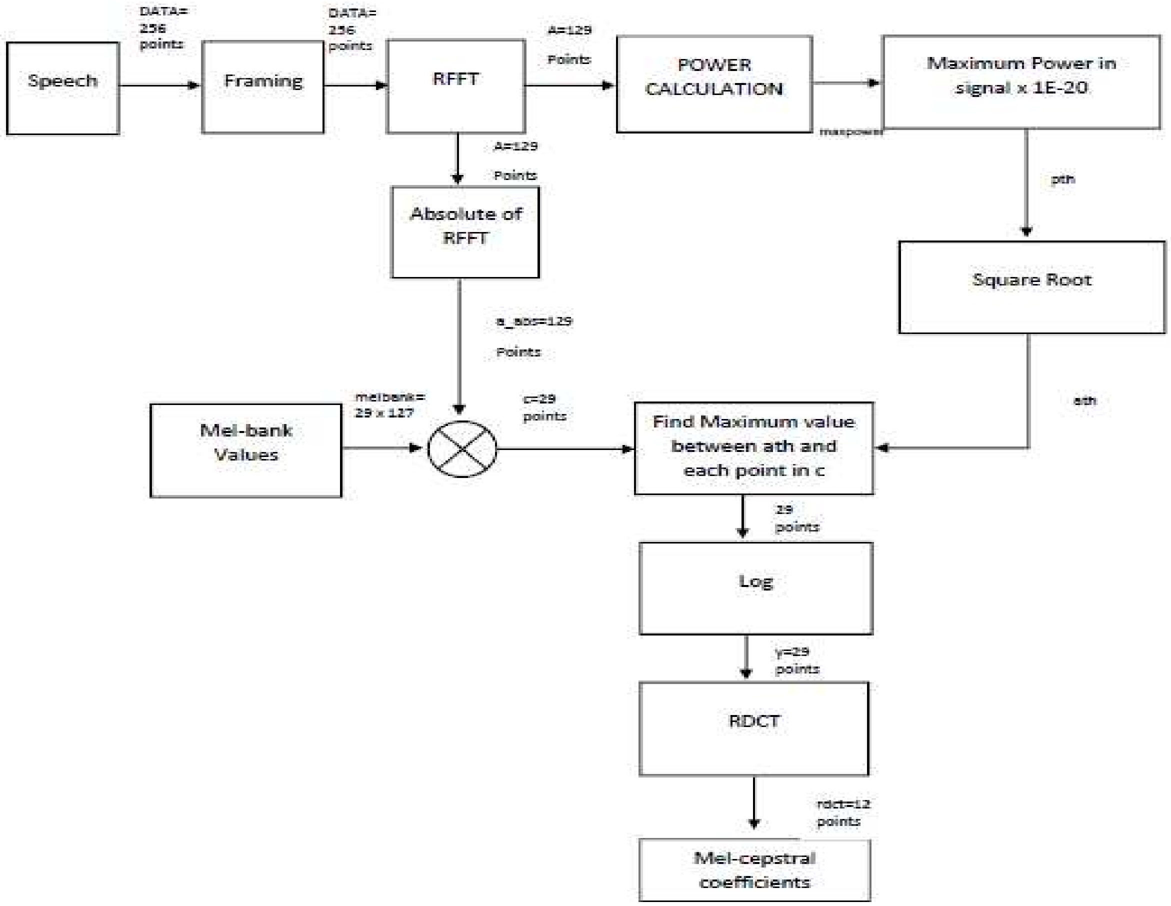


Fig. 9.7 Dataflow Diagram for Mel-cepstral feature extraction

The dataflow diagram for the extraction of mel-cepstral coefficients is given in figure 9.7.At first speech signal given to system where it's framing and windowing process is done. We are using hamming window here.

Let

$$\mathrm{Y(n) = X(n)\ \times W(n)} \tag{9.1}$$

Where

$$\mathrm{W(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right)} \tag{9.2}$$

After that real fast fourier transform is used having output samples as,

$$\text{Noutput= (Ninput /2) +1} \tag{9.3}$$

Where, 'Noutput' is number of output samples and 'Ninput' is number of input samples

$$\text{Noutput} = (256/2) + 1 = 129$$

$$\mathrm{Y(w) = FFT[h(t) * X(t)] = H(w) * X(w)} \tag{9.4}$$

Where, X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively. RFFT gives the frequency domain representation of signal. After performing the RFFT operation we find maximum power in the signal. This is simply done by using equation 9.5

$$\text{Power} = \sqrt[2]{(\text{real part})^2 + (\text{imaginary part})^2} \tag{9.5}$$

Maximum power is calculated from the whole signal. Maximum power is then multiplied for using with log filter.

$$\text{Ath} = \sqrt[2]{(\text{Maximum Power})^{-20}} \tag{9.6}$$

We have used 29 Mel-bank filters in this work. This number is decided by using sampling frequency of signal. After filtering of the signal the each frame will be converted to 29 points of data. The Mel-filters are shown in figure 9.8. Log spectrum of this data is found by taking the LOG of maximum number in Ath and filtered data.

$$\mathrm{Y(n)} = \log\left(\mathrm{x} = \begin{cases} \mathrm{c(n)}, & Ath < c(n) \\ \mathrm{Ath}, & Ath \geq c(n) \end{cases}\right) \tag{9.7}$$

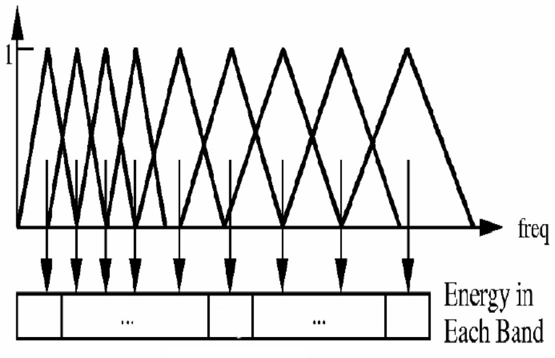


Fig. 9.8 Mel Filters

9.5 Program Flowchart for Real Time Implementation of Speech Recognition

The flowchart of the program for real time implementation of speech recognition is shown in figure 9.9, at first we load the training dataset extracted features into the memory of processor. After loading the features we go for recording the sound from user. Here in our work we have taken input from AIC23 Mic input. For recording the sound the user will have to press the DIP Switch No 3.DSK will record sound until switch is pressed. After releasing the switch processor will extract the features from it. Once Features are extracted from recording, the classification task will start. In classification we have used GMM classifier in this each Euclidian Distance from all the training vectors is calculated. The Command/Voice having minimum ED is result of the process.

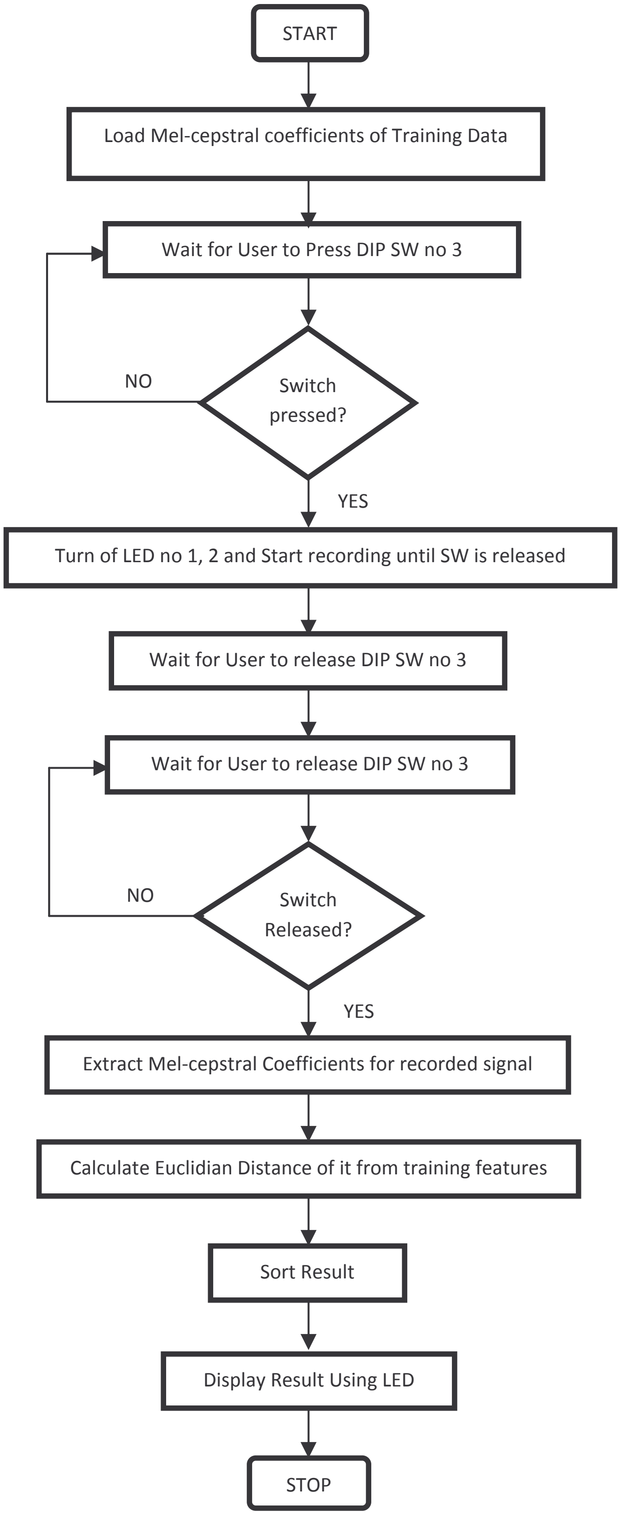

START
Load Mel-cepstral coefficients of Training Data
Wait for User to Press DIP SW no 3
NO
Switch pressed?
YES
Turn of LED no 1, 2 and Start recording until SW is released
Wait for User to release DIP SW no 3
Wait for User to release DIP SW no 3
NO
Switch Released?
YES
Extract Mel-cepstral Coefficients for recorded signal
Calculate Euclidian Distance of it from training features
Sort Result
Display Result Using LED
STOP

Fig. 9.9 Flowchart for real time implementation of speech recognition

Figure 9.10 shows speech waveform of single frame. Figure 9.11 shows RFFT plot of single frame. Figure 9.12 shows plot of the Mel-cepstral coefficients extracted from frame. Table 9.1 shows status of LED when command is detected on TMS320C6713 DSK board. The minimum Euclidean distance between gallery and probe speech (Euclidean distance are highlighted in different following tables). Table 9.2 shows the Euclidean distance between gallery and speech (for speaker Amir [Row: Command1-10] [Column: Command1-10]). Table 9.3 shows the Euclidean distance between gallery and speech (for speaker Amir [Row: Command1-20] [Column: Command11-20]). Table 9.4 shows the Euclidean distance between gallery and speech (for speaker Ayo [Row: Command1-10] [Column: Command1-10]). Table 9.5 shows the Euclidean distance between gallery and speech (for speaker Ayo [Row: Command1-20] [Column: Command11-20]). Table 9.6 shows timing analysis of work in MATLAB and CCS. Hardware configuration of Host PC CPU (Core2Duo), clock frequency is 2.93GHZ+2.93GHZ and that of TMS320C6713DSK clock frequency is 225MHZ.

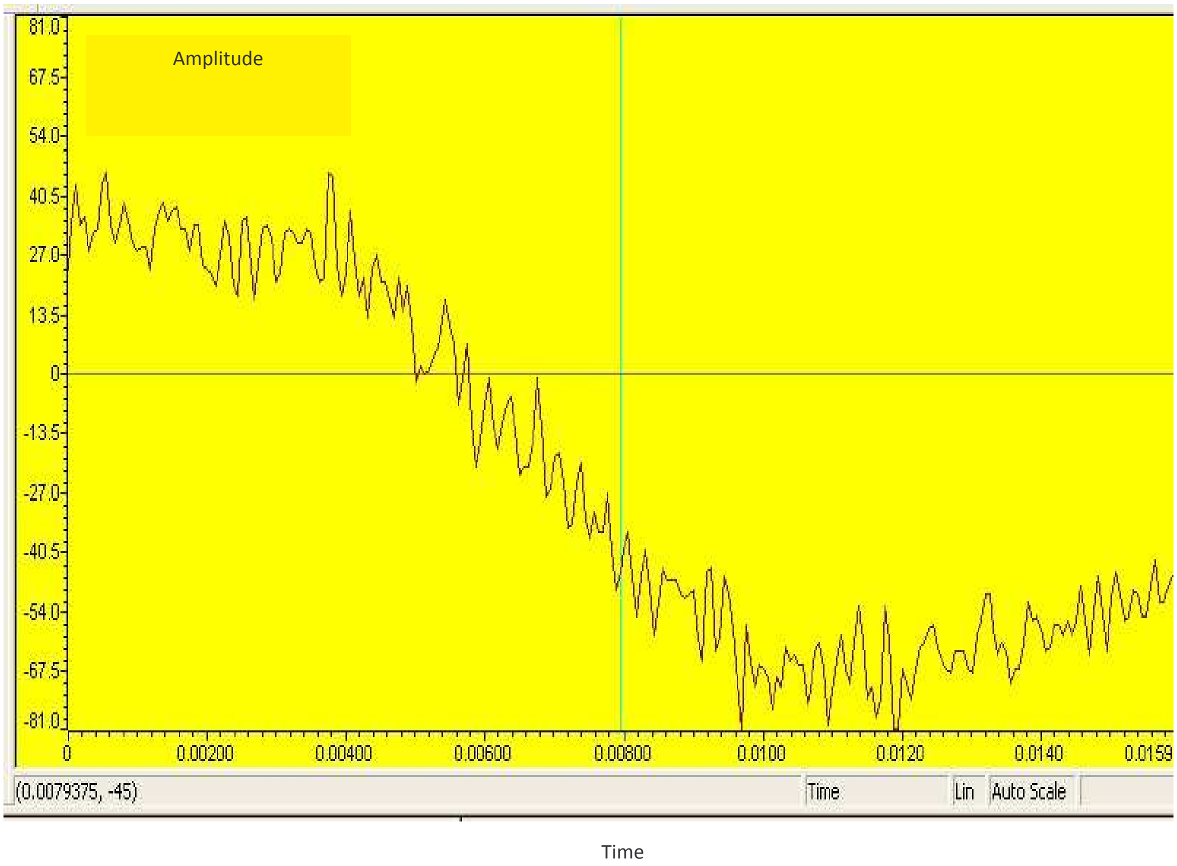


Fig. 9.10 Speech waveform of single frame

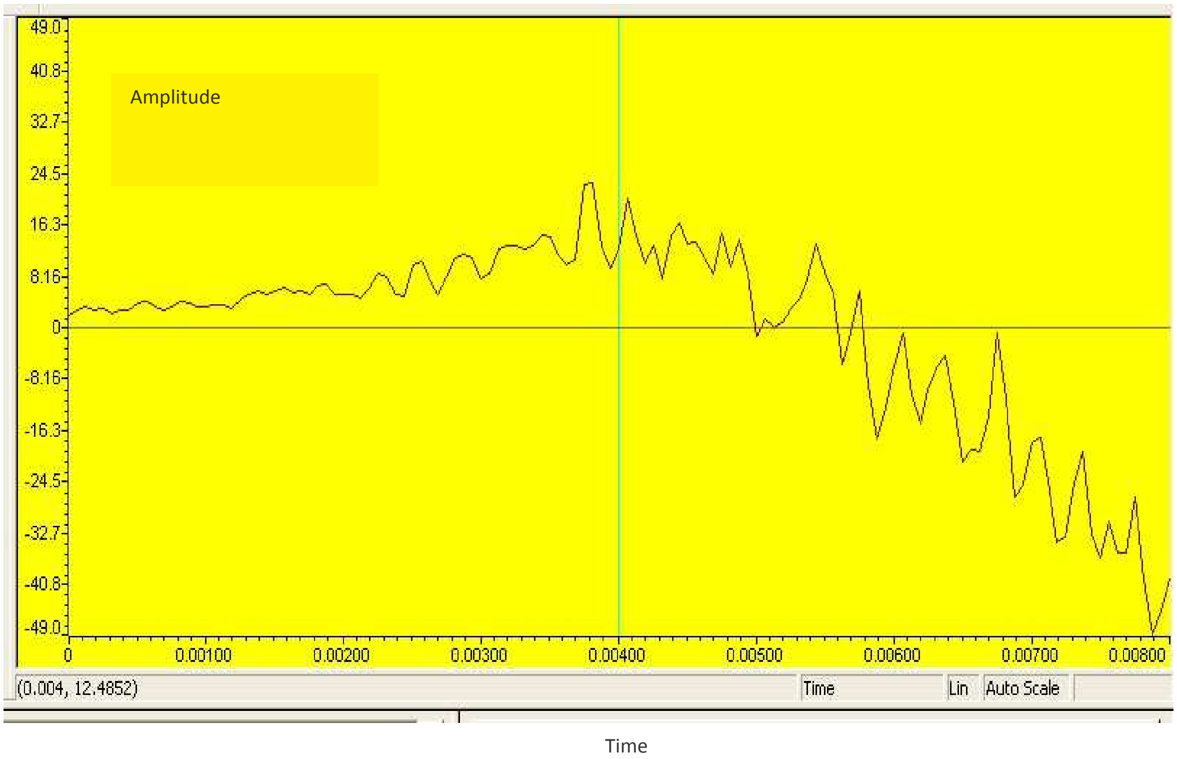


Fig. 9.11 RFFT plot of single frame

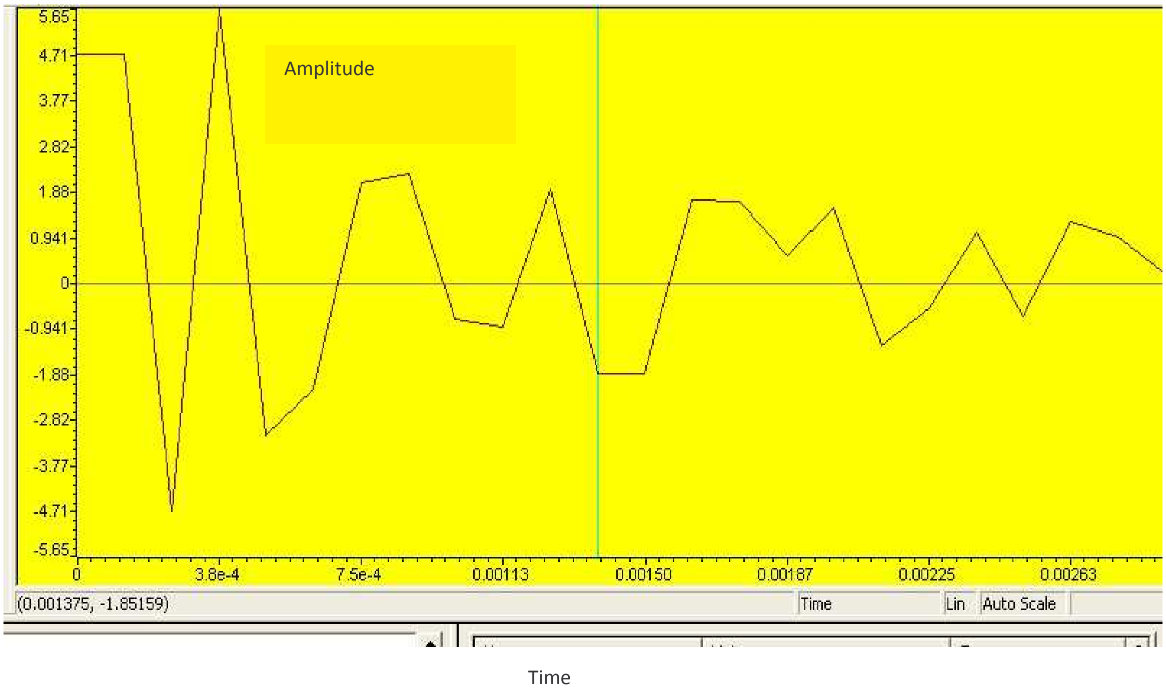


Fig. 9.12 Mel-cepstral coefficients extracted from frame

Table 9.1
Status of LED when command is detected

LED 4	**LED3**	**LED2**	**LED1**	**Detected Command**
0	0	0	0	Command 0
0	0	0	1	Command 1
0	0	1	0	Command 2
0	0	1	1	Command 3
0	1	0	0	Command 4
0	1	0	1	Command 5
0	1	1	0	Command 6
0	1	1	1	Command 7
1	0	0	0	Command 8
1	0	0	1	Command 9
1	0	1	0	Command 10
1	0	1	1	Command 11
1	1	0	0	Command 12
1	1	0	1	Command 13
1	1	1	0	Command 14
1	1	1	1	Command 15

Table 9.2

The Euclidean distance between gallery and speech (Amir [Row: Command1-10] [Column: Command1-10])

Speakers	**Command 1**	**Command 2**	**Command 3**	**Command 4**	**Command 5**	**Command 6**	**Command 7**	**Command 8**	**Command 9**	**Command 10**
Amir Command 1	1674.312499	1875.050547	3178.88766	3592.846642	2273.735253	6907.774004	3706.708939	3092.08818	2194.335329	1551.1374
Amir Command 2	3684.475283	2704.480586	3357.878816	3018.237211	2678.199599	6643.808643	3395.232959	3267.159372	3340.739663	3198.195493
Amir Command 3	4269.400087	3064.379716	3348.30567	3595.968632	2597.26989	6149.472404	3301.459185	3611.340354	3164.776557	3679.014352
Amir Command 4	3104.03495	3058.137651	4523.038401	1718.335241	3170.846878	8011.035428	2917.374275	4531.377495	2710.310743	3588.070601
Amir Command 5	4202.792069	3729.86706	4304.215047	2886.947037	3591.607798	9388.937325	3491.197349	5373.075039	3522.152728	4399.267324
Amir Command 6	4606.999741	2767.348088	3238.490979	4036.240573	2763.584566	4194.57894	3938.060733	3328.119568	3740.260443	3801.471284
Amir Command 7	2235.809737	2904.427294	2866.034057	2363.365046	2555.767294	5954.058064	2168.269791	3272.860584	2418.793007	2807.184325
Amir Command 8	2935.406256	2483.95839	2495.864586	3240.818707	2500.99255	5218.858417	3132.136052	2307.447702	2630.868224	2411.616672
Amir Command 9	2556.943066	3117.276941	4216.310899	3889.919054	2989.207553	8142.894461	4578.429548	3823.534299	2696.956888	2588.509263
Amir Command 10	2184.583271	2121.121296	2525.998308	3645.515934	2237.192074	5617.255457	3544.525437	2482.563713	2620.250118	1851.696401
Amir Command 11	4359.423996	2754.473695	3747.625463	5119.694863	3072.021463	7042.737423	4836.412547	3848.730357	3590.788099	2976.198129
Amir Command 12	2592.785413	2989.150016	4391.817278	4623.913551	3581.580004	7856.706308	5268.84058	4125.095504	2724.526868	2610.189714
Amir Command 13	2163.778086	3234.960689	4253.645318	2326.874906	2868.663982	8346.705651	3420.947611	4143.675491	2359.122491	2860.276738
Amir Command 14	2745.338309	3103.571815	3619.22796	3040.092379	2981.657842	7491.69585	2983.970326	4042.660939	2273.20371	3083.712654
Amir Command 15	1826.441225	2751.345419	3246.743502	2896.170525	2484.229945	6321.484505	3504.956994	2802.973255	2372.273945	2292.755623
Amir Command 16	2466.272101	3630.936673	3418.054428	2451.169957	3498.840945	6258.449027	2480.20257	4211.785257	2877.398906	3729.838852
Amir Command 17	3353.813056	3125.078471	3376.043283	2568.106263	3457.110128	5152.680978	2351.780276	4374.624603	3147.261699	4140.324621
Amir Command 18	4411.748267	3896.675539	3849.978979	3343.688694	3376.542654	8048.454731	3463.121197	5120.382381	3203.519015	4575.228519
Amir Command 19	3914.037306	3592.003165	4364.52191	5564.217185	3879.129038	9189.298182	5879.812137	4855.55744	3495.154634	3342.700499
Amir Command 20	4269.285908	3784.153607	3701.64974	5614.424044	3408.453015	7559.145325	5305.450907	4229.206925	3203.107229	3430.986577

Table 9.3

The Euclidean distance between gallery and speech (Amir [Row: Command1-20], [Column: Command 11-20])

Speakers	**Command 11**	**Command 12**	**Command 13**	**Command 14**	**Command 15**	**Command 16**	**Command 17**	**Command 18**	**Command 19**	**Command 20**
Amir Command 1	5375.545373	15871830.791137	1810.676667	2263.732327	3517.850374	4722.978651	3681.059034	2035.177567	1759.339943	3047.721002
Amir Command 2	6657.063059	4058.564258	2827.318496	3360.045409	3778.413603	4424.559594	3181.067932	3015.971646	4229.371156	4697.170997
Amir Command 3	5953.476397	4688.927923	3913.486132	3601.762901	4584.978661	4241.165588	3270.721459	3566.581719	4930.969381	4670.949511
Amir Command 4	8625.236351	3129.505991	2350.307818	3542.38858	4180.445931	4036.402348	2854.938062	2572.108879	4291.05321	5511.590042
Amir Command 5	7690.017633	3916.75724	3799.69925	3289.526906	6056.613644	4972.554571	3242.160158	3538.707503	5351.287884	5304.024697
Amir Command 6	5117.678441	4531.458679	4334.262668	4252.588395	4102.778666	3864.774762	3385.737315	4071.632853	4611.932494	4017.565777
Amir Command 7	6886.919037	2749.82378	2228.222437	2773.694807	3145.314731	2841.705126	2660.902341	2827.260581	3634.501924	4452.89549
Amir Command 8	5440.371108	3160.183911	2881.399446	3563.944937	3528.84062	4032.2794	3345.187421	2555.040056	3275.160959	3725.493072
Amir Command 9	6321.683314	2339.336539	2580.731843	3139.561457	3838.794847	5456.452942	4550.120065	2153.919134	2537.598115	3952.098534
Amir Command 10	5024.443264	2335.530778	2400.623955	2834.78477	3205.548218	3973.660844	3668.44683	2500.915684	2308.492206	3104.357923
Amir Command 11	2297.578821	3489.666957	4001.051419	3115.276227	5448.11967	5761.409007	4196.930573	3461.759372	3722.066304	2582.220585
Amir Command 12	5256.52629	2119.352801	3323.655479	3371.20066	4949.953762	6100.734842	5281.426742	2662.630526	2457.121966	2843.081049
Amir Command 13	8107.057066	2241.18729	1900.905869	3184.079834	3388.557034	3894.068837	3394.910957	2253.260926	3309.338789	4951.694486
Amir Command 14	6903.189916	3185.847389	2834.916529	2956.902765	4598.779414	4332.488349	3444.479812	2868.039691	3988.921322	4017.670487
Amir Command 15	6514.930815	2149.585256	2022.127672	2938.409593	2400.21905	3688.960539	3487.204298	2343.813018	2518.881644	3895.339834
Amir Command 16	8062.903235	3034.908419	3025.138082	3713.823299	3929.527028	2592.740075	2696.582897	3744.485986	4243.684067	5206.24303
Amir Command 17	7698.346879	3891.170845	3693.17512	4205.91315	4491.170631	2671.091591	2670.498075	3980.052153	4694.860801	4736.990347
Amir Command 18	7932.580412	4484.90646	4413.837598	4009.638789	5684.192788	4512.087498	3719.290427	3932.523148	5545.701595	5465.283625
Amir Command 19	6645.269099	3206.41091	3891.605404	3971.159271	5984.745273	7381.577672	6356.537439	3493.957054	2795.889839	3991.70819
Amir Command 20	4303.097474	3730.441523	4048.140532	3202.757112	5559.759351	5956.409624	5127.252373	3936.042209	3658.715591	2851.903959

Table 9.4

The Euclidean distance between gallery and speech (Ayo [Row: Command 1-10] [Column: Command 1-10])

Speakers	Command 1	Command 2	Command 3	Command 4	Command 5	Command 6	Command 7	Command 8	Command 9	Command 10
Ayo Command 1	2392.555236	2827.845261	3663.986496	4183.469297	2580.622069	7704.065659	4306.790614	3571.002236	2416.237695	2100.719669
Ayo Command 2	2835.343813	2355.700079	2817.563798	4252.404365	2554.623893	6524.570013	3815.239314	3379.088544	2955.795095	2194.20241
Ayo Command 3	3653.402161	3137.16429	1892.127295	3760.36562	3060.91479	4672.837743	3303.731939	2193.169148	3451.105247	3195.470597
Ayo Command 4	2609.68389	3130.852815	3701.91877	4101.789766	2970.06705	7415.149143	4362.987429	3690.001803	3264.240817	2419.59652
Ayo Command 5	2387.747469	2749.063366	3162.221382	3156.663219	2263.047125	6003.20043	2641.924119	3168.792763	2013.754711	2555.351639
Ayo Command 6	6982.393572	5630.212199	5156.874104	7615.789142	6146.505596	1328.899234	6173.58542	4413.408054	7342.748104	6424.962495
Ayo Command 7	3264.532527	3394.807312	3344.149949	2602.174341	3326.153558	5628.407916	1261.077649	4372.684899	2977.757889	4322.565006
Ayo Command 8	4260.979414	2715.513253	2861.410549	5484.464848	3171.00181	4800.678827	5212.094547	2472.505901	4049.885164	2560.373707
Ayo Command 9	2510.343524	2444.288863	3388.529698	3633.559437	2226.071253	6625.308843	3635.087717	3141.064451	1472.43906	1945.994348
Ayo Command 10	2452.705146	2030.362215	2213.6324	3237.462618	1767.096421	4977.743412	2414.434872	2159.61298	2511.092847	1669.228736
Ayo Command 11	5983.290063	5431.07488	6338.626621	7688.485149	5895.506871	3890.756353	7218.62035	4528.264949	6661.589419	5436.699437
Ayo Command 12	2948.034957	3336.728646	5091.340242	3905.841664	3780.549004	7727.764768	4617.246695	4253.884438	3674.587346	3099.664414
Ayo Command 13	2528.847633	2776.278202	3709.228703	2598.47702	2877.391008	7116.897991	3094.585358	3354.39313	2499.622847	2592.169608
Ayo Command 14	2583.970266	2411.940946	3954.003281	3575.430742	2456.529609	7851.244341	3658.692899	4027.212041	2592.110997	2774.288863
Ayo Command 15	4357.42925	4723.156646	5259.918958	6325.162819	4358.351537	5068.030196	6613.994732	3722.853045	5096.116201	4050.880373
Ayo Command 16	7802.366959	5272.428779	6481.59266	6044.34477	5829.555128	8806.505196	5575.018672	7907.411472	6668.851611	7185.599192
Ayo Command 17	5177.571053	4526.412899	4607.885337	2450.454233	4254.25409	7165.61134	2996.884056	6033.316563	4153.217813	6213.409649
Ayo Command 18	2734.285687	2768.87521	3804.549502	3262.639979	2257.59775	7364.963973	3464.875695	3622.648402	1986.591561	2338.292631
Ayo Command 19	1998.462716	2647.17247	3405.264416	3755.002801	2082.580977	7015.770462	4038.480729	2889.748352	2370.683187	2073.649414
Ayo Command 20	3205.456576	3480.242304	3767.554239	5270.319401	3330.680969	4817.4001	5074.817113	3211.595454	3602.180139	3004.951605

Table 9.5
The Euclidean distance between gallery and speech (Ayo [Row: Command1-20] [Column: Command11-20])

Speakers	**Command 11**	**Command 12**	**Command 13**	**Command 14**	**Command 15**	**Command 16**	**Command 17**	**Command 18**	**Command 19**	**Command 20**
Ayo Command 1	5726.356738	2106.065751	2673.80487	2898.061618	4256.18017	5138.38711	4566.019003	2160.927765	1741.834149	3360.178005
Ayo Command 2	4587.132242	2404.755496	3045.86341	3154.238929	4315.188765	4864.786225	4335.751136	3135.541768	2665.227538	3475.875149
Ayo Command 3	5931.522745	4437.473555	3821.729965	4020.766285	4007.999397	4437.059426	3816.151721	3877.718591	4444.010972	4471.985099
Ayo Command 4	6085.803257	2267.706932	2730.898715	3330.93658	4342.149097	4886.498161	4546.664812	3258.756905	2541.958999	4238.885291
Ayo Command 5	5658.275718	2729.080581	2234.241451	2368.579901	3491.015068	3191.775858	3037.186082	2890.609883	3053.435551	4009.788704
Ayo Command 6	5867.312442	7553.508054	7257.596701	7916.164591	4308.021265	4444.468453	5264.465989	7859.752917	7446.796163	6194.995067
Ayo Command 7	7679.776557	4219.398075	3631.436899	3475.111408	4772.716082	2679.305353	2477.69572	4278.723556	5467.439758	5500.094749
Ayo Command 8	2578.987606	3731.957421	4059.785757	3929.565371	4603.421256	5847.417768	4779.216584	3932.315398	3229.15948	2631.279836
Ayo Command 9	4916.657333	2094.493073	2204.799637	2531.721324	3634.16461	4509.534155	3684.63896	1598.393582	2101.394434	2848.228346
Ayo Command 10	4375.027201	2648.505526	2273.797176	2534.621688	3307.297262	2991.848535	2575.664607	2510.990129	2758.60786	3467.94768
Ayo Command 11	5861.079129	6021.942861	6124.479595	7097.799529	3935.749883	5429.882901	6286.664935	6322.135581	6162.705466	6052.230711
Ayo Command 12	6849.639061	2397.862709	2978.218825	3949.341395	3603.35887	4824.535103	4341.309585	2941.453742	3199.581905	4833.100367
Ayo Command 13	6516.716295	2591.648825	1587.13312	2526.510149	3549.383933	4009.333379	3202.198032	2215.077021	3130.232891	4056.805569
Ayo Command 14	5844.332502	2532.318266	2811.785293	1611.679055	4650.346246	4881.970924	3831.535715	2581.833605	2564.12078	3080.914463
Ayo Command 15	6338.478716	4370.680841	4126.083097	5924.013991	2293.21107	5044.524569	5894.832019	4731.546603	4100.038845	5570.424272
Ayo Command 16	8028.980795	7280.325609	7599.585716	5953.953429	8691.728165	6215.442741	4729.644393	7006.315773	8131.657755	6281.758622
Ayo Command 17	9479.296102	5469.702428	5060.324779	4913.828563	6007.314713	3479.785796	2799.788504	5541.617613	7457.681768	6905.501995
Ayo Command 18	5648.797448	2207.185868	2322.848959	2892.991264	3861.106437	4343.198773	3545.122798	1652.576891	2698.991862	3675.709676
Ayo Command 19	6134.885435	2178.053898	2135.703277	2548.820175	3540.105885	4259.221187	4029.930866	2375.156789	1720.508623	3208.662886
Ayo Command 20	4436.385892	3292.815507	3837.83834	4371.796654	3918.898092	4769.453467	4868.271495	3867.340708	2816.823253	2966.234772

Table 9.6

Timing analysis in MATLAB and CCS

Task	CCS for Single Frame (sec)	CCS for Complete Frame (sec)	MATLAB (sec)
Windowing	0.012982556	1.324220753	0.216439
FFT	0.03200512	3.264522281	0.503182
Power Spectrum	0.001080574	0.110218548	0.001157
Mel Filtering and LOG	2.612486945	266.4736684	0.000576
DCT	0.00420581	0.428992661	0.003605
Classification	0.055253163	5.635822606	0.002509
Total No. of Frames=102	2.718014169	277.2374453	0.727468

9.6 Conclusion

In this chapter we depicted the real time hardware implementation of speech recognition using DSP processor software development kit, DSK-TMS320C6713 with Code Composer Studio (CCS)[8]. MFCC algorithm calculates cepstral coefficients of Mel frequency scale. After feature extraction from recorded speech, each Euclidian Distance (ED) from all training vectors is calculated using Gaussian Mixture Model (GMM). The command/voice having minimum ED is applied as similarity criteria. The timing analysis is done for various individual blocks of algorithm. The time required for processing in DSP and PC processors are compared. Timing analysis in MATLAB is taking less time this is due to more Clock speed of CPU and more memory.

[8] Paper published on "Real Time Speech Recognition Using DSK TMS320C6713", *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)*, Volume 3, Issue 12, December 2013.

CHAPTER 10

CONCLUSIONS AND FUTURE SCOPES

The importance of speech based applications is increasing day by day not only in industries but also in every one's life. The applications like the digital voice communications, human-machine interfaces and automatic speech recognition systems have been so integral part of the human life. The major concern for speech based applications is a decrease in performance while using it in the noisy environment. In hands-free operation of cellular phones in car, the speech signal to be transmitted may be contaminated by vibration, engine and background noise. Among various classes of methods for speech enhancement, beamforming is promising technique for removing the noise from speech and keeping the useful information of speech intact. Our research work deals in the problem of speech enhancement using beamforming.

We have initiated research work with two-fold objectives: 1) To improve the speech recognition performance in multi-microphone environment and 2) We attempted to analyze the performance of speech recognition against the filter-bank parameters; filter length and number of sub bands. In next part of the research work, we have improved the performance of beamforming based speech recognition system using evolutionary computational algorithms (Genetic algorithm, GA). Additionally, the system is made to be working in real-time as time required for classifier has been reduced dramatically. This is particularly achieved by including the zeros at random places and in random amount in initial population chromosomes, which were generated randomly in the range of 0 to 1. This results in the reduction of feature elements in feature descriptor and have feature vector length. We have also analyzed the timing analysis of hardware implementation of speech recognition algorithm on DSP processor TMS320C6713–DSK kit. The results were compared with implementation using host PC in MATLAB.

As there is easy availability of the multi core processor like GPGPU now days, the speech recognition algorithm with complex computations can be implemented using multi-core processor. In future, speech enhancement and recognition techniques can benefit from the software tools and parallel hardware that are available at cheaper cost in the market.

CHAPTER 11

REFERENCES

CHAPTER 1

[1] L.R. Rabiner, R.W.Schafer, Digital Processing of Speech Signals, LPE, Pearson Education, Delhi, 2006.

[2] Douglas O'Shaughnessy, Speech Communications, 2nd Ed., University press (India) Ltd., Hydrabad, 2001.

[3] Thomas F. Quatieri, Discrete-time Speech Signal Processing, 1st Indian reprint, Pearson education signal processing series, Delhi, 2004.

[4] Udo Zolzer, 'Digital Audio Signal Processing' 2nd ed., Wiley publication 2002.

[5] Harry Urkowitz, "Some Properties and Effects of Reverberation in Acoustic Surveillance", IEEE Transactions on Aerospace and Electronic Systems, Vol. Aes-4, No. 1, January 1968.

[6] Ulrich Sauvagerd, "A Ten-Channel Equalizer for Digital Audio-Applications", IEEE Transactions on Circuits and Systems, Vol 36, No. 2, February 1989.

CHAPTER 2

[1] Paurav Goel, Anil Garg "Review of Spectral Subtraction Techniques for Speech Enhancement", International Journal of Electronics and Communication Technology (IJECT), Vol. 2, Issue 4, pp. 189-194, Oct.-Dec. 2011.

[2] S.V. Vaseghi, "Advance Digital Signal Processing and Noise Reduction", John Wiley & Sons Ltd, 3rd Edition, ISBN: 978-0-470-09495-2, February 2006.

[3] Ephraim, Y., "A minimum mean square error approach for speech enhancement", ICASSP-90, Vol. 2, pp. 829-832, 1990

[4] S.China Venkateswarlu, Dr. K.Satya Prasad, Dr. A.SubbaRami Reddy, "Improve Speech Enhancement Using Weiner Filtering", Global Journal of Computer Science and Technology", Vol. 11, Issue 7, ver.1.0, USA 2011.

[5] Marcel Gabrea, "Adaptive Kalman Filtering Based Speech Enhancement Algorithm", Proceeding of 7th International workshop on Acoustic Echo and Noise Control, Germany Sept. 2001.

[6] P.M. Halder and A. K. M. Fazlul Haque, "Improved Echo Cancellation in VOIP", International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 2, No. 11, pp. 122-125, 2011.

[7] Donoha D.L., "Denoising by soft thesholding," IEEE Transactions on Information Theory, vol.41, no.3, pp. 613- 627, 1995.

[8] M.A. Abd E-Fattah, M. I. Dessouky, S. M. Diab and F. E. Abd El-samie, "Speech Enhancement using Adaptive Wiener Filtering Approach," Progress in Electromagnetic Research M, Vol.4, pp.167-184, 2008.

CHAPTER 3

[1] Jae Lim, Oppenheim A, "All-pole modeling of degraded speech," IEEE Transactions on Acoustics, Speech and Signal Processing, vol.26, no.3, pp.197- 210, June 1978.

[2] McAulay R, Malpass, M, "Speech enhancement using a soft-decision noise suppression filter," IEEE Transactions on Acoustics, Speech and Signal Processing, vol.28, no.2, pp. 137-145, April 1980.

[3] Ephraim Y. Malah D., "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Transactions on Acoustics, Speech and Signal Processing, vol.32, no.6, pp.1109- 1121, December 1984.

[4] Mathews V. Dae Youn, Ahmed N., "A unified approach to nonparametric spectrum estimation algorithms," IEEE Transactions on Acoustics, Speech and Signal Processing, vol.35, no.3, pp. 338- 349, March 1987.

[5] Ahmed, M.S., "Comparison of noisy speech enhancement algorithms in terms of LPC perturbation," IEEE Transactions on Acoustics, Speech and Signal Processing, vol.37, no.1, pp.121-125, January 1989.

[6] Virag N., "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Transactions on Speech and Audio Processing, vol.7, no.2, pp.126-137, March 1999.

[7] Carnero B., Drygajlo A., "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms," IEEE Transactions on Signal Processing, vol.47, no.6, pp.1622-1635, June 1999.

[8] Jabloun Firas, Champagne Benoit, "A perceptual signal subspace approach for speech enhancement in colored noise," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol.1, no., pp.I-569-I-572, 13-17 May 2002.

[9] Martin Rainer, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol.1, no., pp.I-253-I-256, 13-17 May 2002.

[10] Lin L, Holmes W.H., Ambikairajah E., "Adaptive noise estimation algorithm for speech enhancement," Electronics Letters, vol.39, no.9, pp. 754- 755, 1 May 2003.

[11] Li Deng, Droppo J., Acero A., "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," IEEE Transactions on Speech and Audio Processing, vol.11, no.6, pp. 568- 580, Nov. 2003.

[12] O. L. Frost, III, “An algorithm for linearly constrained adaptive array processing,” Proc. IEEE, vol. 60, pp. 926–935, Jan. 1972.

[13] Griffiths, L., Jim C., "An alternative approach to linearly constrained adaptive beamforming," IEEE Transactions on Antennas and Propagation, vol.30, no.1, pp. 27-34, January 1982.

[14] Van Compernolle D., "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol.2, no.2, pp.833-836, 3-6 Apr 1990.

[15] Gannot S., Burshtein D., Weinstein E., "Signal enhancement using beamforming and nonstationarity with applications to speech," IEEE Transactions on Signal Processing, vol.49, no.8, pp.1614-1626, Aug 2001.

[16] Grbic, N., Nordholm, S., Cantoni, A., "Optimal FIR subband beamforming for speech enhancement in multipath environments," Signal Processing Letters, IEEE , vol.10, no.11, pp. 335- 338, Nov. 2003.

[17] Xianxian Zhang, Hansen, J.H.L., "CSA-BF: a constrained switched adaptive beamformer for speech enhancement and recognition in real car environments," IEEE Transactions on Speech and Audio Processing, vol.11, no.6, pp. 733- 745, Nov. 2003.

[18] Seltzer, M.L., Raj B., Stern, R.M, "Likelihood-maximizing beamforming for robust hands-free speech recognition," IEEE Transactions on Speech and Audio Processing, vol.12, no.5, pp. 489- 498, Sept. 2004.

[19] Yermeche Z., Grbic N., Claesson I., "Blind Subband Beamforming With Time-Delay Constraints for Moving Source Speech Enhancement," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.8, pp.2360-2372, Nov. 2007.

[20] Guangji Shi, Parham Aarabi Hui Jiang, "Phase-Based Dual-Microphone Speech Enhancement Using A Prior Speech Model," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.1, pp.109-118, Jan. 2007.

[21] Maganti H.K., Gatica-Perez D., McCowan I., "Speech Enhancement and Recognition in Meetings With an Audio–Visual Sensor Array," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.8, pp.2257-2269, Nov. 2007.

[22] Hai Huyen Dam, Hai Quang Dam, Nordholm S., "Noise Statistics Update Adaptive Beamformer With PSD Estimation for Speech Extraction in Noisy Environment," IEEE

Transactions on Audio, Speech, and Language Processing, vol.16, no.8, pp.1633-1641, Nov. 2008.

[23] Han S., Hong J., Jeong S., Hahn M., "Robust GSC-based speech enhancement for human machine interface," IEEE Transactions on Consumer Electronics, vol.56, no.2, pp.965-970, May 2010.

[24] Oppenheim A.V., Weinstein E., Zangi K.C., Feder M., Gauger D., "Single-sensor active noise cancellation," IEEE Transactions on Speech and Audio Processing, vol.2, no.2, pp.285-290, April 1994.

[25] Kuo S.M, Tahernezhadi M Li Ji, "Frequency-domain periodic active noise control and equalization," IEEE Transactions on Speech and Audio Processing, vol.5, no.4, pp.348-358, July 1997.

[26] Gan W.S., Kuo S.M, "An integrated audio and active noise control headset," IEEE Transactions on Consumer Electronics, vol.48, no.2, pp.242-247, May 2002

[27] Ying Song, Yu Gong, Kuo S.M., "A robust hybrid feedback active noise cancellation headset," IEEE Transactions on Speech and Audio Processing, vol.13, no.4, pp. 607- 617, July 2005.

[28] Hinamoto Y., Sakai H., "Analysis of the filtered-X LMS algorithm and a related new algorithm for active control of multitonal noise," IEEE Transactions on Audio, Speech, and Language Processing, vol.14, no.1, pp. 123- 130, Jan. 2006.

[29] Akhtar M. T., Abe M., Kawamata M, "On Active Noise Control Systems With Online Acoustic Feedback Path Modeling," Audio, Speech, and Language Processing, IEEE Transactions on , vol.15, no.2, pp.593-600, Feb. 2007.

[30] Ferrer M., Gonzalez A., de Diego, M. Pinero, G., "Fast Affine Projection Algorithms for Filtered-x Multichannel Active Noise Control," IEEE Transactions on Audio, Speech, and Language Processing, vol.16, no.8, pp.1396-1408, Nov. 2008.

[31] Sheng-Wei Zhang and T.J. Stonham. Universal Architectures for Logical Neural Nets. In: Second International Conference on Artificial Neural Networks, Conference Publication No. 349 IEE. Pages 262-266. 18-20 November 1991.

[32] Jinsoo Jiong, “A Kepstrum approach to Real Time Speech Enhancement”, Chapter 5-6, PhD Thesis, Massey University at Albany, 2007.

CHAPTER 4

[1] Andreas Zell, “Simulation Neuronaler Netze”, Addison-Wesley publication, 1994.

[2] Murray L. Barr & John A. Kiernan, “The Human Nervous System. An Anatomical Viewpoint”, Fifth Edition, Harper International, 1988.

[3] M. Minsky and S. Papert. Perceptrons, MIT Press, Cambridge, MA1969.

[4] Zadeh, L. A. Berkeley, “Fuzzy Sets”, Information and Control, Vol.8, pp.338-353, University of California, 1965.

[5] Jang, J.S.R., “ANFIS: Adaptive-Network-based Fuzzy Inference System”, Systems, Man and Cybernetics, Vol. 23, pp. 665-685 0018-9472, IEEE, June 1993.

[6] Wang, Hong Guang, “Fuzzy Control in Manufacturing Systems”, Eindhoven, Eindhoven University of Technology, 1997.

[7] Jantzen Jan., “Design of Fuzzy Controllers”, Department of Automation, Technical University of Denmark, 1998.

CHAPTER 5

[1] Barry D. Van Even and Kevin M. Buckley, “Beamforming: A versatile approach to spatial filtering”, IEEE Acoustics, Speech, and Signal Processing Magazine, pages 4-24, April 1988.

[2] H. Cox, R. Zeskind, and T. Kooij, “Practical super gain,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-34, pp. 393–397, June 1986.

[3] R. Taylor and G. Dailey, “The super-directional acoustic sensor,” in Proceedings of OCEANS ’92 -Mastering the Oceans through Technology, vol. 1, pp. 386–391, 1992.

CHAPTER 7

[1] O. L. Frost, III, “An algorithm for linearly constrained adaptive array processing,” Proc. IEEE, vol. 60, pp. 926–935, Jan. 1972.

[2] Griffiths L., Jim C., "An alternative approach to linearly constrained adaptive beamforming," IEEE Transactions on Antennas and Propagation, vol.30, no.1, pp. 27- 34, Jan 1982.

[3] Van Compernolle D., "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90, vol.2, pp.833-836, 3-6 Apr 1990.

[4] Gannot S., Burshtein D., Weinstein E., "Signal enhancement using beamforming and non stationarity with applications to speech," IEEE Transactions on Signal Processing, vol.49, no.8, pp.1614-1626, Aug. 2001.

[5] Grbic N., Nordholm S., Cantoni A., "Optimal FIR subband beamforming for speech enhancement in multipath environments," IEEE Signal Processing Letters, vol.10, no.11, pp. 335- 338, Nov. 2003.

[6] Xianxian Zhang, Hansen J.H.L., "CSA-BF: a constrained switched adaptive beamformer for speech enhancement and recognition in real car environments," IEEE Transactions on Speech and Audio Processing, vol.11, no.6, pp. 733- 745, Nov. 2003.

[7] Seltzer M.L., Raj B., Stern R.M., "Likelihood-maximizing beamforming for robust hands-free speech recognition," IEEE Transactions on Speech and Audio Processing, vol.12, no.5, pp. 489- 498, Sept. 2004.

[8] Yermeche Z., Grbic N., Claesson I., "Blind Subband Beamforming With Time-Delay Constraints for Moving Source Speech Enhancement," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.8, pp.2360-2372, Nov. 2007.

[9] Guangji Shi, Parham Aarabi, Hui Jiang, "Phase-Based Dual-Microphone Speech Enhancement Using A Prior Speech Model," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.1, pp.109-118, Jan. 2007.

[10] Maganti H.K., Gatica-Perez D., McCowan I., "Speech Enhancement and Recognition in Meetings With an Audio–Visual Sensor Array," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.8, pp.2257-2269, Nov. 2007.

[11] Hai Huyen Dam, Hai Quang Dam, Nordholm S., "Noise Statistics Update Adaptive Beamformer With PSD Estimation for Speech Extraction in Noisy Environment," IEEE Transactions on Audio, Speech, and Language Processing, vol.16, no.8, pp.1633-1641, Nov. 2008.

[12] Han S., Hong J., Jeong S., Hahn M., "Robust GSC-based speech enhancement for human machine interface," IEEE Transactions on Consumer Electronics, vol.56, no.2, pp.965-970, May 2010.

[13] John J. Shynk, “Frequency-domain and multirate adaptive filtering,” IEEE Signal Processing Magazine, vol. 9, pp. 14–37, 1992.

[14] Jan Mark de Haan, Nedelko Grbic, Ingvar Claesson, and Sven Erik Nordholm, “Filter bank design for subband adaptive microphone arrays,” IEEE Trans. Speech Audio Proc., vol. 11, no. 1, pp. 14–23, Jan. 2003.

[15] Kumatani K., McDonough J., Schachl S., Klakow D., Garner P.N., Weifeng Li, "Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming," IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, pp.1609-1612, March 31 2008-April 4 2008.

[16] P. P. Vaidyanathan, “Multirate Systems and Filter Banks”, Prentice Hall, Englewood Cliffs, 1993.

[17] Kenichi Kumatani, Tobias Gehrig, Uwe Mayer, Emilian Stoimenov, John McDonough, and MatthiasW¨olfel, “Adaptive beamforming with a minimum mutual information criterion,” IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 8, pp. 2527–2541, 2007.

[18] L. Rabiner and Biing-Hwang Juang, “Fundamentals of Speech Recognition”, Prentice Hall PTR, c1993.

[19] Joseph W. Picone, “Signal Modeling Techniques in Speech Recognition”, Proceedings of the IEEE, vol. 81, No. 9, pages 1215--1247, 1993.

[20] Steven B. Davis and Paul Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, No. 4, August 1980.

CHAPTER 8

[1] Chan K.Y., Low S.Y., Nordholm S., Yiu K.F.C., Ling S.H., "Speech Recognition Enhancement Using Beamforming and a Genetic Algorithm," Third International Conference on Network and System Security, NSS '09., pp.510-515, Oct. 2009.

[2] Chmulik M., Jarina R., "Bio-inspired optimization of acoustic features for generic sound recognition," 19th International Conference on Systems, Signals and Image Processing (IWSSIP), pp.629-632, April 2012.

[3] Harrag A., Saigaa D., Boukharouba K., Drif M., Bouchelaghem A., "GA-based feature subset selection: Application to Arabic speaker recognition system," 11th International Conference on Hybrid Intelligent Systems (HIS), pp.383-387, 5-8 Dec. 2011.

[4] Gao Wen-xi, Yu Feng-qin, "Feature dimension reduction based on genetic algorithm for mandarin digit recognition," 4th International Congress on Image and Signal Processing (CISP), vol.5, pp.2737-2740, 15-17 Oct. 2011.

[5] Selouani S., "Evolutionary discriminative speaker adaptation," IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp.164-168, 11-15 Dec. 2011.

[6] Yuan Yujin, Zhou Qun, Zhao Peihua, "Vector Quantization Codebook Design Method for Speech Recognition Based on Genetic Algorithm," 2nd International Conference on Information Engineering and Computer Science (ICIECS), pp.1-4, 25-26 Dec. 2010.

[7] Aggarwal R.K., Dave M., "Application of genetically optimized neural networks for Hindi speech recognition system," World Congress on Information and Communication Technologies (WICT), pp.512-517, 11-14 Dec. 2011.

[8] Shing-Tai Pan, Ching-Fa Chen, Jian-Hong Zeng, "Speech recognition via Hidden Markov Model and neural network trained by genetic algorithm," International Conference on Machine Learning and Cybernetics (ICMLC), vol.6, pp.2950-2955, 11-14 July 2010.

[9] Oudelha M., Ainon R.N., "HMM parameters estimation using hybrid Baum-Welch genetic algorithm," International Symposium in Information Technology (ITSim), vol.2, pp.542-545, 15-17 June 2010.

[10] Nemade M. U., Shah S.K., “Improvement in Speech Recognition Performance using Beamforming based Speech Enhancement”, International Journal of Electronics Communication and Computer Engineering (IJECCE) ISSN: 2249-071X (Online, http://ijecce.org) Volume 3 Issue 4, July 2012.

[11] Milind U. Nemade, Satish K. Shah, “Beamforming based Speech Recognition using Genetic Algorithm for Real Time Systems”, International Journal of Recent Technology and Engineering, ISSN: 2231-2307(online), Vol.2, Issue 2, pp. 96-104, 2013.

CHAPTER 9

[1] L. Rabiner, B.H. Juang and B. Yegnanarayana, “Fundamentals of Speech Recognition”, Pearson Education, first edition, ISBN 978-81-7758-560-5, 2009.

[2] H.S. Jayanna, S.R. Mahadeva, “Analysis, Feature Extraction, Modelling and Testing Techniques for Speaker recognition”, IETE Tech. Rev., 26:181-90, 2009.

[3] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani, Md. Saifur Rahman, “Speaker Identification Using Mel Frequency Cepstral Coefficients”, 3rd International Conference on Electrical and Computer Engineering ICECE, Dhaka, Bangladesh, 28-30 December 2004.

[4] D.A. Raynolds, “Speaker Identification and verification using Gaussian mixture speaker models”, Speech Communication, Vol.17, No. 1-2, pp. 91-108, 1995.

[5] TMS320C6713 datasheet available on www.ti.com.

[6] Spectrum Digital Incorporated, TMS320C6713 DSK Technical Reference, 2004.

[7] Texas Instruments, Code Composer Studio User’ Guide, 2005.

[8] S.M. Kuo, B.H. Lee, W. Tian, Real Time Digital Signal Processing: Implementations and Applications, 2nd Ed., John Wiley & Sons Ltd., West Susex, England, 2006.

[9] Molau S., Pitz M., Schluter R. and Ney H., "Computing Mel-frequency Coefficients on Power Spectrum", Proceeding of IEEE ICASSP-2001, Vol. 1, pp.73-76, 2001.

APPENDIX A

RESEARCH PROJECT

Details of the research project completed by the candidate as a part of PhD work is as follows

Project Title:	GUI Based Quantitative Performance Comparison of Single Channel Speech Enhancement Techniques for Personal Communication
Project No.:	318
Project Sanctioned Letter No.:	APD/237/318 of 2012
Academic Year:	2012-2013
Sponsored Agency:	University of Mumbai
Amount Sanctioned:	Rs.18500/-

APPENDIX B

PAPER PUBLICATIONS AND PRESENTATIONS

Table B.1 lists the papers presented/published in various national/international conferences/journals and indexed in databases; based on the work described in the thesis.

Sr. No.	Conference/Journal	Year	Paper Title
1	National Conference on Emerging Trends in Electronics and Telecommunication Engineering (ETETE-2011), Watumull Institite of Electronics Engineering and Computer Technology, Worli, Mumbai, 16th-17th Sept. 2011	2011	Exploring of Real Time Speech Processing Strategies: A Review of Applications
2	National Conference on Emerging Technologies and Applications in Engineering and Science (NCETAES). Journal Published by International Society of Science and Technology, Mumbai, (ISSN 0974-0678)	2011	Digital Signal Processing based Implementation of Auditory System Parameters
3	Fourth International Conference on Electronics, Computer Technology (ICECT), Published in International Journal of Future Computer and Communication (IJFCC), ISSN: 2010-3751, DOI: 10.7763/IJFCC, Volume 3, No.3.	2012	Speech Enhancement Techniques: Quality vs. Intelligibility
4	International Conference on New Development and Challenges in Engineering, Technology and Management	2012	Performance Evaluation in VOIP Network with Acoustic Echo Cancellation and Adaptive Wiener Filter
5	International Journal of Electronics Communication and Computer Engineering (IJECCE), (ISSN:2249-071X), Paper ID: 730, Vol.3, Issue 4, Pages 745-751	2012	Improvement in Speech Recognition Performance using Beamforming based Speech Enhancement
6	International Journal of Innovative Research In Computer and Communication Engineering (IJIRCCE), Paper ID: V1I10C043, ISSN (online):2320-9801, ISSN (print): 2320-9798, Vol.1, Issue 1.	2013	Performance Comparison of Single Channel Speech Enhancement Techniques For Personal Communication

7.	International Journal of Advance Research in Computer and Communication Engineering (IJRCCE), ISSN: 2278-1021, Paper ID: V25105, Vol.2, Issue 5, Pages 2039-2043.	2013	Survey of Soft Computing based Speech Recognition Techniques for Speech Enhancement in Multimedia Applications
8.	International Journal of Recent Technology and Engineering, Paper ID: B0618052213, ISSN: 2231-2307(online), Vol.2, Issue 2, Pages 96-104.	2013	Beamforming based Speech Recognition using Genetic Algorithm for Real Time Systems
9	International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 3. Issue 12, December 2013	2013	Real Time Speech Recognition Using DSK TMS320C6713

Table B.1 List of papers published/presented

APPENDIX C

SHORT TERM TRAINING PROGRAM ATTENDED

Table C.1 lists the short term training programs attended during the Ph.D. work.

Sr. No	Duration	Topic Name	Learning Objectives	Venue
1	04th -08th Jan. 2010 (One Week)	Neural Network and Fuzzy System	Study of soft computing techniques	K. J. Somaiya Institute of Engineering and Information Technology, Sion(E), Mumbai
2	10th March 2011 (One Day)	MATLAB and Simulink for Engineering Education	MATLAB Simulink for Model Preperation	MathWorks India Pvt. Ltd, Pune
3	17th -21st Dec. 2012 (One Week)	Research Methodology and Related Open Source Tools	LATEX language for research report writing	K. J. Somaiya Institute of Engineering and Information Technology, Sion(E), Mumbai
4	15th-20th April 2013 (One Week)	Recent Trends in Optimization and its Application in Engineering	Study of Soft Computing Techniques	Shri. Sant Gadgebaba College of Engineering and Technology, Bhusawal

Table C.1 List of short term training programs attended

APPENDIX D

ADDITIONAL RESOURCES FOR RESEARCH WORK

Table D.1 shows list of the persons contacted for additional resources for research work

Sr. No.	Person Name and Institute	Guidance / Discussion
1	Dr. Pandey, IIT Mumbai	Guidance on various issue related to speech processing
2	Dr. B.K. Mohan, IIT Mumbai	Attended lectures delivered on Neural Network and Fuzzy Logic at KJSIEIT, Sion, Mumbai
3	Dr. Milind Shah, Head of Electronics and Telecommunication Department, Fr. C. Rodrigues Institute of Technology, Vashi, Navi Mumbai	Guidance and discussion about real time speech processing for multimedia applications
4	Dr. R. D. Kanfade, Principal, Dhole Patil College of Engineering, Pune	Topic selection and guidelines for writing research paper and proposal
5	Dr. Hemant Patil, DAIICT, Gandhinagar	Attended seminar on “Wavelet Transform” at KJSIEIT, Sion, Mumbai
6	Dr. Mandar Sahastrbudhe, Visiting Faculty, Charusat, Changa	Worked on project “Audio Deblurring” under his guidance

D1.List of the persons contacted for additional resources for research work

APPENDIX E

SUMMARY OF MATLAB CODE DEVELOPED FOR RESEARCH WORK

Summary of MATLAB Code developed for different techniques

1. Dataset used for research work:

SpeechData	Jim (Speaker1)	Recorded wave files 10 for numbers and 10 for commands for speech length of 2 seconds (5_1.wav to 5_20.wav)
	Sameh(Speaker2)	
SpeechData1	Amir (Speaker3)	
	Ayo (Speaker4)	

2. Single channel Speech Enhancement:

MATLAB GUI developed for following process

Speech_enhancement.fig	Figure file for MATLAB GUI
Speech_enhancement.m	Code developed for comparison of different Speech enhancement techniques
ss.m	Code developed for spectral subtraction
kalman.m	Code developed for Kalman filtering process
wiener.m	Code developed for wiener filtering process
awf.m	Code developed for adaptive wiener filtering process

3. Code used for Real Time implementation of Speech Recognition

enframe.m	Split signal into overlapping frames
mel2freq.m	Compute frequency from mel value
rdct.m	Compute discrete cosine transform of real data
freq2mel.m	Compute mel value from frequency
melbankm.m	Determine matrix for mel spaced filter bank
rfft.m	Return FFT value of real data
melcepst.m	Calculate mel cepstrum of a signal
Speech_reco_03	Speech recognition code

4. Speech Enhancement using Beamforming Process:

BeamFormProcess.m	Code developed for beamforming process
MultiMicVoiceGen.m	Code developed for multi mic voice generation for Beamform process
Beamform_filter_design.m	Code developed for Beamform Filter design process
Subband_filter_design_analysis.m	Code developed for analysis of subband filter design

5. Speech Recognition using Genetic Algorithm

Speech_reco_GA.m	Code developed for speech recognition and optimization of parameters using Genetic algorithm

6. CCS codes:

DAT file	5_1.dat to 5_20.dat for four spakers	Converted wave file to DAT file for CCS
melcepst	melcepst.c	Split signal into overlapping frames
	melbank.h	Header file for mel filter bank
	melcepst.pjt	Project file for melcepst operation
feature_average	feature_average.c	Calculate feature average for speech recognition
	feature_average.h	Header file for feature average calculation
	feature_average.pjt	Project file for feature average operation
speech_recognition	speech_recognition.c	Speech recognition task code in C
	speech_recognition.pjt	Project file for speech recognition