SMARTPHONE-BASED SINGLE AND DUAL MICROPHONE SPECH ENHANCEMENT ALGORITHMS FOR HEARING STUDY

by

Gautam Shreedhar Bhat

APPROVED BY SUPERVISORY COMMITTEE:

_____

Dr. Issa M. S. Panahi

_____

Dr. Carlos Busso

_____

Dr. Lakshman S. Tamil

To God, to my parents and

to everyone who trusted my abilities

SMARTPHONE-BASED SINGLE AND DUAL MICROPHONE SPECH ENHANCEMENT ALGORITHMS FOR HEARING STUDY

by

GAUTAM SHREEDHAR BHAT, BE

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN

ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

April 2018

ACKNOWLEDGMENTS

This thesis represents not only my work at the computer but this work is also the outcome of two years of hard work and brain storming at UT Dallas and especially in the Statistical Signal Processing Research Lab. The journey towards my Master's degree is a memorable one and there are several people who helped to make it remarkable. It is with great pleasure that I take this opportunity to express my gratitude towards these people who have inspired me and who have become the backbone of my aspirations and accomplishments. I am grateful to each person who has impacted my career in a positive way.

Foremost, I am thankful to my mentor, Dr. Issa M. S. Panahi for providing me an opportunity to work in the Statistical Signal Processing Research Laboratory (SSPRL) and to pursue thesis under him. I would like to express sincere gratitude for his continuous support of my study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Carlos Busso and Dr. Lakshman Tamil for agreeing to be on my supervisory committee.

I also thank all my colleagues who were and are at SSPRL: Chandan K. A. Reddy, Anshuman Ganguly, Yiya Hao, Nikhil Shankar, Ram Charan Chandrashekhar, Abduallah Kucuck, Parth Mishra, Serkan T, Ziyan Zou and Holden Hernandez for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last two years. Also, I would like to recognize Chandan K. A. Reddy for his collaborative efforts and major contributions in Algorithm development and Nikhl Shankar for all his help in Smartphone implementation.

I would like to thank my friends Adarsh, Varun, Akshay, Amith, and Nitin, for all their support during my tough times and for their involment in my growth and success. I would like to thank my professor in India Dr. D. Ganesh Rao for his able guidance to shape my career.

Last but not the least, I would like to thank my parents, my siblings and my relatives for supporting me spiritually throughout my life and being with me during my toughest times. This thesis is an outcome of 2 years of combined hard work and determination. Thank you all for being an inspiration and trusting me.

This work was supported by the National Institute of the Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health (NIH) under the grant number 5R01DC015430-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors are with the Statistical Signal Processing Research Laboratory (SSPRL), Department of Electrical and Computer Engineering, The University of Texas at Dallas.

April 2018

SMARTPHONE-BASED SINGLE AND DUAL MICROPHONE SPECH ENHANCEMENT ALGORITHMS FOR HEARING STUDY

Gautam Shreedhar Bhat, MSEE
The University of Texas at Dallas, 2018

Supervising Professor: Dr. Issa M. S. Panahi

Speech Enhancement (SE) is elemental in many real world applications. In the last two decades, extensive studies have been carried out on single and multi-channel SE techniques. In this thesis, three novel SE algorithms have been proposed that can be used for Hearing Aid Devices using a smartphone as their assistive device. The first SE method exploits the information of formant locations to improve the speech quality and intelligibility of the Super-Gaussian Joint Maximum aposterori (SGJMAP) SE method. The second method is the extension of this work on the Log Spectral Minimum Mean Square Error Amplitude Estimator (Log-MMSE) which is a well-known SE algorithm. The third method is a real time Blind Source Separation (BSS) method based on Independent Vector Analysis (IVA) for convolutive mixtures. Objective and subjective evaluation of the developed techniques show substantial improvements in speech quality and intelligibility.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Since time immemorial, speech is one of the most important communication forms of humanity. While in former times conversations were possible only face-to-face, speech communication environments have changed drastically over the past 2 decades with the advances in speech processing technologies and ubiquity in telecommunications. A new generation of speech acquisition applications are developed such as Hands-free audio communication, Mobile telephony, Hearing Aids, Automatic Information systems i.e., Voice Controlled Systems, Video Conferencing Systems and many of the multimedia applications. In all these applications, the received speech signal of interest should be of high perceptual quality and intelligibility. This places high demands on the robustness of these devices to operate well in acoustically challenging conditions. The performance of these devices deteriorates considerably in the presence of background noise and depending on the amount of speech contaminated by the noise, the effects can be terrible. The speech perception is greatly affected due to the addition of some frequency components, masking of desired speech spectral components and smearing of speech spectra near the occurrences of phonemes.

Speech enhancement (SE) is used to process the noisy speech signal, reduce the impact of disturbances and improve the quality and intelligibility of the degraded speech signal at the receiving end. SE or noise reduction is a key feature in many applications like hearing aid devices (HADs), wearable technology, and virtual reality (VR) & gaming industries. Though, over the decades extensive research has been conducted in the area of SE, the robustness of the SE

algorithm to the rapidly changing acoustical environmental noise has been a limiting factor. Thus an SE method that is robust to different environmental noisy conditions becomes critical.

1.1 Problem Statement

According to World Health Organization (WHO), 360 million people across the globe have disabling hearing loss. Statistics obtained by National Institute on Deafness and Other Communication Disorders (NIDCD) show that approximately 15% of American adults (37.5 million) aged 18 and over report some concern in hearing. About 2 to 3 out of every 1,000 children in the United States are born with a detectable level of hearing loss. However, more than 90% of individuals with hearing impairment can be helped with HADs, cochlear implants and other hearing instruments [1, 2].

SE is a vital block in HAD signal processing pipeline. In the real world, the environment is surrounded by different types of acoustic noise which appears in different shapes and forms. In many conditions, to understand the speech in a noisy environment becomes an extremely difficult task for the normal hearings and the condition gets worse for hearing impaired people. As a result, obtaining the enhanced speech signal with an improved quality and minimal speech distortion is the ultimate goal. Figure 1.1 shows the block diagram of the signal processing pipeline used in the HADs which typically has a Feedback suppression block, Direction of Arrival (DOA) Estimation

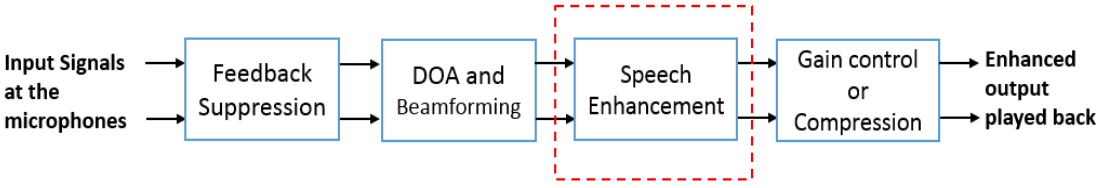


Figure 1.1. Block Diagram of HAD signal processing pipeline

block to find the directionality of the source, an SE system to reduce the background noise and the output gain control and Dynamic Audio compression block.

1.2 Solution and Approaches

The existing HADs have their own limitations like size, processor, and power consumption. Therefore, these limited computing powers make it highly impractical to implement complex yet useful signal processing algorithms on HADs to improve their performance [3]. One practical solution is to use a smartphone as an assistive tool for HADs as they have the superior processing power and large population anyways possess smartphones.

1.3 Single and Dual Microphone Speech Enhancement

In the HADs, as shown in figure 1.1, SE is the main block which concentrates on improving the speech quality and intelligibility and boost the Signal to Noise Ratio (SNR). The main objective of any SE technique must be concerned with improving the speech intelligibility and perceptual quality which has been corrupted due to different types of background noises. Among all the methods used for SE, the single microphone SE is the most challenging task due to the lack of the spatial information. This makes it highly challenging to reduce the background noise without inducing speech distortion in single channel SE. Single channel SE can be generally divided into five main classes: spectral subtraction algorithms, Wiener filtering, subspace algorithms, binary mask algorithms and statistical-model based algorithms [4]. Among these SE techniques, the statistical model based methods are widely used and well known to reduce background noise. In this thesis, the statistical model based methods are used to improve the overall speech quality. In these methods, a statistical model is assumed for the speech spectral coefficients. Therefore, to the

estimate the true speech spectra in the noisy observation, Bayesian rule is usually used to compute the conditional mean of the estimator. The minimum mean square error (MMSE) estimator [5] which is one of the most popular statistical model-based methods for suppressing the background noise (especially which are non-stationary) and also improving the perceptual speech quality for single channel SE [6-7] was proposed by Ephraim and Malah. The estimate of a priori SNR which is the key parameter in statistical model-based methods for computing frequency spectral weights was also proposed by Ephraim and Malah known as the decision-directed approach (DDA) [5]. To improve the performance of the SE algorithms especially when the SNR is too low, there are numerous noise estimation algorithms that have been introduced. There are many computationally efficient alternatives for the MMSE method, in this new method, speech is estimated by applying the joint maximum a posteriori (JMAP) estimation rule [8]. In [9], super-Gaussian extension of the JMAP (SGJMAP) is proposed which is shown to outperform algorithms proposed in [5, 7]. Super-Gaussian statistical model of the clean speech and noise spectral components (especially Babble) attains a lower mean squared error compared to Gaussian model. Recent developments include SE based on deep neural networks (DNN) [10, 11], which requires rigorous training data. Although these methods yield supreme noise suppression, the preservation of Spectro-temporal characteristics of speech, the quality and natural attributes remains as a prime challenge. To enhance the speech using multiple microphones, adaptive algorithms like Least Mean Square (LMS), Normalized Least Mean Square (NLMS) have been used in the past [12]. As the spatial information of the speech and the noise can be exploited using multiple microphones, with a good estimate of the DOA, a null can be steered towards the noise source and a beam towards the signal source using beamforming techniques. Researchers have found interest in BSS techniques for

enhancing the speech [13]. BSS techniques use the information of the mixed signals observed at the input microphones. Independent Component Analysis (ICA) is a widely used BSS technique [14].

1.4 Thesis Objectives and Outline

Improving the quality and the intelligibility of enhanced speech is the major concentration of this thesis. The conventional single channel SE algorithms have their own disadvantages based on particular noise environment, i.e. some of them create annoying musical tones in the highly non-stationary noise environment. In certain conditions, some of the algorithms induce speech distortion in important frequency regions. This work will put the main efforts on the modification and development of alternative statistical-model based SE methods and their real-time implementation on smartphones. Out of the three SE methods are proposed in this thesis, the first two methods are based on single microphone:

I. Formant frequency based Single Channel Super Gaussian Speech Enhancement

The proposed SE method is based on new super Gaussian joint maximum *a Posteriori* (SGJMAP) estimator [15]. We use the formant trajectory information to improve the overall quality and intelligibility of the speech. '*Tradeoff'* parameters are introduced to modify the gains over multiple frequency bands. The algorithm is implemented on smartphone for online playback applications and this arrangement can be used as an assistive device for hearing aids.

II. Improved Log spectral amplitude estimator

The developed SE method is an extension of first method to Log-MMSE SE technique i.e. an MMSE estimator based on the formant frequency is proposed. A scaled value of the MMSE gain function is applied over different bands which is estimated based on formant locations.

The proposed method does not induce any residual or musical noise so there is no requirement of any post filter after the enhancement.

III. Real time Independent Vector Analysis

Independent Vector Analysis (IVA) is a frequency domain BSS technique used to find an optimal demixing matrix for convolutive mixtures of source signals. In our approach, we use two microphones to separate speech and noise signals. The demixing matrix is calculated batch wise and updated based on the DOA information. Improved speech quality is observed especially for low SNR conditions (SNRs equal to and lesser than 0dB).

Performance evaluation of all the proposed methods are presented using objective and subjective test results.

The outline of the thesis is as follows: Chapter 2 provides a brief review of single channel and multi-channel SE algorithms. It also illustrates the pros and cons of existing standard SE algorithms. Chapter 3 introduces Smartphone based super Gaussian single microphone SE using formant information. Its real-time implementation is discussed. Performance evaluation is made by measuring the objective measures and subjective tests. Chapter 4 discusses about formant based Real time Log-Spectral Amplitude Estimator SE technique. Chapter 5 discusses IVA based real-time source separation, analysis, and results. Chapter 6 draws the important conclusion on the proposed SE techniques.

CHAPTER 2

LITERATURE REVIEW ON SPEECH ENHANCEMENT

2.1 Overview of Single Channel SE Techniques

In single channel SE, only one microphone is used to estimate the desired speech. Typically, the estimation of clean speech is obtained by passing the noisy speech signal through a filter. The prime challenge is to find the optimal filter which can suppress the noise as well as maintain the speech integrity. Due to high non-stationary nature of speech signal, most of the single channel SE techniques are operated on short time frames to obtain the optimal filter, as speech is stationary in short duration in the order of 10-40 milliseconds [5]. Spectral-Subtraction algorithm [16] is one of the first algorithms proposed for noise reduction. The principle of this algorithm is that an estimate of the clean signal spectrum can be accomplished by subtracting an estimate of the noise spectrum from the noisy speech spectrum. The challenge with the Spectral-Subtraction algorithm is that the estimation of the noise spectral components from the noisy speech spectra needs to be pretty accurate. Some of the speech components will be removed if excess noise is subtracted. The noise components remain if too little is subtracted. It is not facile to get a good estimate of the noise spectral components when SNR of the noisy speech is low. Sometimes, subtraction of the noise spectrum does create the annoying effect called 'musical noise'.

Wiener filtering approach derives the enhanced signal by optimizing the tractable error criterion, mean-square error [17]. The optimal filter is computed by minimizing the estimation error between the desired signal and the output signal. The optimum filter is derived based on the well-known orthogonality principle with the assumption that the noise and speech Discrete Fourier Transform

(DFT) coefficients are independent Gaussian random variables. The Wiener filter approach yields a linear estimate of the complex spectrum of the signal. Though Wiener filters are much less subject to signal corruption but were never really able to fully remove the background noise.

Statistical model based methods mainly focus on the nonlinear estimators of the modulus of the DFT coefficients i.e. magnitude of the signal rather than the complex spectrum of the signal as done by the Wiener filter. Various methods exist for deriving nonlinear estimators, firstly, Maximum-Likelihood (ML) estimators assume that the parameter to be estimated is deterministic but unknown [18]. Bayesian estimators assume that the parameter to be estimated is a random variable and works on the fact that we have available *a priori* knowledge of the parameter to be estimated. Hence it performs better than the ML estimator as it makes use of distribution (probability density function) of the speech signal. A statistical model that utilizes the asymptotic statistical properties of the Fourier transform coefficients has been proposed by Ephraim and Malah [5] and the optimal estimator is found using this model which minimizes the mean-squared error between the estimated and the true spectral magnitudes. The estimator takes probability density function (pdf) of speech and noise DFT coefficients, into account and is combined with soft decision gain modifications that take speech presence probability into account. A metric based on the squared error of the log-magnitude spectra is more suitable for speech processing [6, 19] as the metric based on just the squared error of the magnitude spectra may not be subjectively meaningful. In [6], the researchers concluded that the statistical-model based methods performed the best across in terms of quality enhancement across all conditions. These conclusions were made based on the subjective evaluation of speech quality which was performed for various SE algorithms to track the amount of speech distortion, noise reduction and overall speech quality of

the enhanced signal. However, in [20], the authors studied the intelligibility of enhanced speech across the numerous SE algorithms and concluded the intelligibility of the noisy speech was equal to that of the enhanced speech in most of the algorithms, with the exception of certain signal noise condition. Also, the algorithms that were found to perform the best in terms of overall quality of the enhanced signal and that performed the best in terms of speech intelligibility were different.

In the subspace filtering approach, the noisy speech signal space is divided into a signal space containing both speech and noise and a noise space containing only noise [21]. An estimator which projects the noisy speech onto the signal subspace linearly by minimizing the speech signal distortion can be estimated. Numerous variations of this algorithm have been proposed for various noise types [22, 23].

2.2 Overview of Multi-Channel SE Techniques

In the multi-channel SE techniques, the main advantage that it has at least 2 microphones and this will assist to exploit the spatial information of the speech source and the noise source to enhance the speech. Practically, the location of these sources in the space is different. Therefore, with a good estimate of the DOA of the sources, using beamforming approaches, a null can be steered towards the noise source and a beam towards the signal source [24]. Adaptive filters can be used to enhance the speech by using the input signals at multiple microphones and then subtract the estimated noise from the noisy speech. This can be achieved by well-known adaptive algorithms like LMS, NLMS [25] methods. Though these methods do not provide substantial elimination of noise, it acts as an SNR booster. Literature provides many instances where beamforming and adaptive filter methods are used as pre-filter for single channel SE techniques [12, 26]. The advantage of this technique is, these pre-filters increase the SNR without inducing any speech

distortion. The remaining residual noise is eliminated by single channel SE methods. Multichannel noise reduction can be performed in frequency domain by applying a complex weight to the output of each microphone, at each frequency bin [27, 28]. Since speech signals are correlated at successive frames, beamforming can be improved by taking interframe correlation into account. In [29], researchers have proposed a dual-microphone SE technique, which is based on the magnitude of coherence between input signals to suppress the coherent noise, emanating from a single interfering source. In the recent times, researchers have found interest in enhancing the speech quality using BSS methods [13]. The mixing information observed at the input microphones is used in BSS techniques. ICA is one of the BSS technique [14] which is commonly used for linear mixtures. IVA is mostly used for signals having convolutive mixture model.

CHAPTER 3

SMARTPHONE BASED SUPER GAUSSIAN SINGLE MICROPHONE SPEECH ENHANCEMENT USING FORMANT INFORMATION

3.1 Introduction

In the United States about 2 percent of adults aged 45 to 54 have disabling hearing loss. The rate increases to 8.5% for adults aged 55 to 64. Nearly 25% of those aged 65 to 74 and 50% of those who are 75 and older have disabling hearing loss. These Statistics obtained by National Institute on Deafness and Other Communication Disorders (NIDCD) show the importance of the personal hearing devices like HADs for the hearing impaired. Though SE is an elemental block in HAD signal processing pipeline due to the drawbacks in HADs it is impractical to implement important signal processing algorithms on them to further improve their performance. One viable solution is to use smartphone as an assistive tool for HADs as they have sophisticated processors and large population anyways possess smartphones. The microphone on the smartphone captures the noisy speech. The SE algorithm running on the processor of the smartphone reduces the background noise and the enhanced speech is wirelessly transmitted to the HADs. Recently, extensively used smartphones such as Apple iPhone have come up with new HA features such as Live Listen [30] to enhance the overall quality and intelligibility of the speech perceived by hearing impaired.

Literature offers extensive studies where SE algorithms are developed to improve the performance of HADs in the presence of background noises for the purpose of gaining better hearing capability by users of these devices. However, the prime challenge in single microphone SE is to suppress the background noise without inducing any sort of speech distortion. Traditional SE methods like spectral subtraction [16] and statistical model based methods proposed by Ephraim and Malah [5,

7] can be implemented on a smartphone in real-time. But, these algorithms induce musical noise and do not substantially improve speech intelligibility. There are some computationally efficient alternatives for [4-5] which are proposed in [8-9]. Recent developments include SE based on deep neural networks (DNN) [10], which is not suitable for real-time applications, as it requires rigorous training data and extensive training period also these methods fail to retain the natural attributes of speech. But, none of these methods provide a control to suppress the amount of noise reduction in real-time. Recent development in SE algorithms [15] gives provision to control the amount of noise reduction in real time. Studies on speech intelligibility [15] show that maintaining speech intelligibility and also reducing the background noise is inadequate in many widely used algorithms. Studies on speech intelligibility [31] show that speech intelligibility improvement is inadequate in many widely used algorithms. Researchers have shown that ideal binary mask in SE could improve intelligibility [32], but accurate estimation of the binary mask is challenging, especially in lower SNR conditions. In [33], Loizou explained two types of distortions that play a significant role in speech intelligibility. One of the major acoustical cues to identify the vowels [34], Diphthongs [36], vowel-consonant transitions [37, 38] and also nasal consonants [35] are the formant frequency trajectories which represent typical characteristics of a speech signal. Several studies have been carried out to increase the intelligibility in speech coders [39] using formant frequencies. Recently in [40], formant shaping method is used to improve speech intelligibility. In this chapter, an SE technique is presented to improve the quality and intelligibility of speech perceived by Hearing Aid users using smartphone as an assistive device. The formant frequency information is used to improve the overall quality and intelligibility of the speech. The proposed SE method is based on new super Gaussian joint maximum *a Posteriori* (SGJMAP) estimator [15]

in which there is a "*tradeoff"* factor introduced in the optimization of SGJMAP cost function to estimate the clean speech magnitude spectrum. Therefore, the derived gain function is a function of a priori, a posteriori and the tradeoff factor. However, the gain function used in this method is designed such that the tradeoff parameter is applied over the entire frequency range. To this method we use the priori information of formant frequency locations, in order to divide the entire frequency range of the signal to multiple bands i.e. the formant frequency bands and the non-formant frequency bands. Therefore, the derived gain function has two parameters based on the acoustically important bands. This allows to suppress more noise in acoustically unimportant bands without inducing distortion in clean speech and residual noise. The formant frequency information helps the hearing aid user to control the gains over the non-formant frequency band and the formant frequency band, allowing the HA users to attain more noise suppression while maintaining the speech intelligibility using a smartphone application. The resulting enhanced signal is of good perceptual quality.

3.2 Brief Overview of Formant Frequencies

A formant is a concentration of acoustic energy around a specific frequency in the speech signal. Formants are vocal tract resonances, the frequency depends on length of the vocal tract which makes it speaker dependent. There are several formants, each at a different frequency, roughly one in each 1000Hz band. Or, to put it differently, formants occur at roughly 1000Hz intervals. Each formant corresponds to a resonance in the vocal tract. The first two formants are particularly important in speech recognition. Frequencies of formants change only within 15% between female and male speakers [41]. Formant frequencies are widely used in speaker recognition, speaker ID, and speech coders etc. Formant frequencies represent particular speech characteristics and it is

widely used to identify phonemes including vowels, nasal consonants, and consonants in consonants - vowel transitions. Sound-induced hearing impairment can cause cochlear hair cell damage, leading to the degradation of the auditory nerve response to formant frequencies [42], [43]. It is likely that these degradations contribute to decreased speech intelligibility for people suffering from sound-induced hearing loss. Hence it becomes very important to determine the formant locations and make sure that there is minimal speech distortion in formant specific regions. Formants can be tracked for clean speech by estimating the resonance frequencies of the speaker's vocal tract using linear prediction method at regular intervals of the signals and the formants are identified by picking the peaks. However, formant tracking becomes challenging with an intruding background noise and most of the formant tracking algorithms are neither robust to real world noise or nor suitable for real time implementations. In our method, we use a well-known technique for formant tracking [38] which obtains formant frequency estimates from voiced segments of continuous speech by using a linear predictive method to track individual formant frequencies. The formant tracker incorporates an adaptive voicing detector and a gender detector for formant extraction from continuous speech, for both male and female speakers.

3.3 Conventional SGJMAP Method

In the SGJMAP [9] method, a super Gaussian speech model is used by considering non-Gaussianity property in spectral domain noise reduction framework [44, 45] and by knowing that Speech spectral coefficients have a super-Gaussian distribution. Spectral amplitude estimator using super Gaussian speech model allows the probability density function (PDF) of the speech spectral amplitude to be approximated by the function of two parameters μ and v. These two parameters can be adjusted to fit the underlying PDF to the real distribution of the speech

magnitude spectrum. Considering the additive mixture model for noisy speech $y(n)$, with clean speech $s(n)$ and noise $w(n)$,

$$y(n) = s(n) + w(n) \tag{3.1}$$

The noisy k^{th} Discrete Fourier Transform (DFT) coefficient of $y(n)$ for frame λ is given by,

$$Y_k(\lambda) = S_k(\lambda) + W_k(\lambda) \tag{3.2}$$

where S and W are the clean speech and noise DFT coefficients respectively. In polar coordinates, (2) can be written as,

$$R_k(\lambda)e^{j\theta_{Y_k}(\lambda)} = A_k(\lambda)e^{j\theta_{S_k}(\lambda)} + B_k(\lambda)e^{j\theta_{W_k}(\lambda)} \tag{3.3}$$

where $R_k(\lambda)$, $A_k(\lambda)$, $B_k(\lambda)$ are magnitude spectrums of noisy speech, clean speech and noise respectively. $\theta_{Y_k}(\lambda)$, $\theta_{S_k}(\lambda)$, $\theta_{W_k}(\lambda)$ are the phase spectrums of noisy speech, clean speech and noise respectively. The goal of any SE technique is to estimate clean speech magnitude spectrum $A_k(\lambda)$ and its phase spectrum $\theta_{S_k}(\lambda)$. We drop λ in further discussion for brevity. The JMAP estimator of the magnitude and phase jointly maximize the probability of magnitude and phase spectrum conditioned on the observed complex coefficient given by,

$$\hat{A}_k = \arg\max_{A_k} \frac{p(Y_k|A_k,\theta_{S_k})p\left(A_k,\theta_{S_k}\right)}{p(Y_k)} \tag{3.4}$$

$$\hat{\theta}_{S_k} = \arg\max_{\theta_{S_k}} \frac{p(Y_k|A_k,\theta_{S_k})p\left(A_k,\theta_{S_k}\right)}{p(Y_k)} \tag{3.5}$$

Assuming uniform distribution for phase, the joint PDF

$$p\left(A_k,\theta_{S_k}\right) = \frac{1}{2\pi}p(A_k) \tag{3.6}$$

The super-Gaussian PDF [44] of the amplitude spectral coefficient with variance σ_{S_k} is given by,

$$p(A_k) = \frac{\mu^{v+1}}{\Gamma(v+1)} \frac{A_k^v}{\sigma_{S_k}^{v+1}} \exp\left\{ -\frac{\mu A_k}{\sigma_{S_k}} \right\} \tag{3.7}$$

Assuming the Gaussian distribution for noise and super-Gaussian distribution (3.7) for speech, (3.4) is given by [9],

$$\hat{A}_k = \left(u + \sqrt{u^2 + \frac{v}{2\hat{\gamma}_k}} \right) R_k, \; u = \frac{1}{2} - \frac{\mu}{4\sqrt{\hat{\gamma}_k \hat{\xi}_k}} \tag{3.8}$$

where $\hat{\xi}_k = \frac{\hat{\sigma}_{S_k}^2}{\hat{\sigma}_{W_k}^2}$ is the *a priori* SNR and $\hat{\gamma}_k = \frac{R_k^2}{\hat{\sigma}_{W_k}^2}$ is the *a posteriori* SNR. $\hat{\sigma}_{W_k}^2$ is estimated using a voice activity detector (VAD). $\hat{\sigma}_{S_k}$ is the estimated instantaneous clean speech power spectral density. In [9], $v = 0.126$ and $\mu = 1.74$ is shown to give better results. The optimal phase spectrum is the noisy phase itself $\hat{\theta}_{S_k} = \theta_{Y_k}$.

3.4 Proposed Formant Based SE Method.

The model explained in [9] is less robust and inaccurate as it depends on parameters like μ , v and accuracy of the voice activity detector (VAD), it is very difficult to estimate the magnitude spectrum of the clean speech in the real world rapidly fluctuating acoustical environment. In [15], these inaccuracies were compensated by introducing a tradeoff factor β into the cost function for estimating the optimal clean speech magnitude spectrum. Taking natural logarithm of (3.4), and differentiating with respect to A_k gives,

$$\frac{d}{dA_k} \log(p(Y_k|\beta A_k, \theta_{S_k}) p(\beta A_k, \theta_{S_k})) =$$

$$\frac{-(Y_k^* - A_k \beta e^{-j\theta_{S_k}})\left(-jA_k \beta e^{j\theta_{S_k}}\right) + (Y_k - A_k \beta e^{j\theta_{S_k}})(jA_k \beta e^{-j\theta_{S_k}})}{\hat{\sigma}_{W_k}^2} \tag{3.9}$$

Setting (9) to zero and substituting $Y_k = R_k e^{j\theta_{Y_k}}$ simplifies to

$$\frac{2R_k}{\hat{\sigma}_{W_k}^2} - \frac{2A_k\beta}{\hat{\sigma}_{W_k}^2} + \frac{v}{A_k\beta} - \frac{\mu\beta}{\hat{\sigma}_{S_k}} = 0 \tag{3.10}$$

On simplifying (10), the following quadratic equation is obtained,

$$A_k^2 + \frac{A_k}{2\beta\hat{\sigma}_{S_k}}\left(\hat{\sigma}_{W_k}^2\mu\beta - 2R_k\hat{\sigma}_{S_k}\right) - \frac{v\hat{\sigma}_{W_k}^2}{2\beta^2} = 0 \tag{3.11}$$

Solving the above quadratic equation and writing in terms of $\hat{\xi}_k$ and $\hat{\gamma}_k$ yields

$$A_k = \left[\left(\frac{1}{2\beta} - \frac{\mu}{\sqrt[4]{\hat{\gamma}_k\hat{\xi}_k}}\right) + \sqrt{\left(\frac{\mu}{\sqrt[4]{\hat{\gamma}_k\hat{\xi}_k}} - \frac{1}{2\beta}\right)^2 + \frac{v}{2\hat{\gamma}_k\beta^2}}\right] R_k \tag{3.12}$$

The speech magnitude spectrum estimate is

$$\hat{A}_k = G_k\, R_k \tag{3.13}$$

Where,
$$G_k = \left[\left(\frac{1}{2\beta} - \frac{\mu}{\sqrt[4]{\hat{\gamma}_k\hat{\xi}_k}}\right) + \sqrt{\left(\frac{\mu}{\sqrt[4]{\hat{\gamma}_k\hat{\xi}_k}} - \frac{1}{2\beta}\right)^2 + \frac{v}{2\hat{\gamma}_k\beta^2}}\right] \tag{3.14}$$

3.4.1 Formant frequency Band Estimation

As an extension to the proposed method in [15] and in order to improve the performance, we make use of formant information. The formant frequency bands are approximated by calculating the exact formants to improve the intelligibility of speech. The pitch and the formant frequency trajectories $(f_0 - f_3)$ of the clean speech or speech degraded with noise at high SNR can be calculated by the method explained in [38]. We require a frequency range to approximate the presence of speech and apply considerably less noise suppression over that band and more noise suppression on the other bands. Therefore, the mean of pitch and the formants $f_0 - f_3$ are

calculated for large data sets and mean absolute error for each formant is determined over the data sets to find the frequency band of probable formant location. The frequency band is given by (3.15)

$$F_S = \left[\left(f_S - \frac{f_a}{2}\right), \left(f_S + \frac{f_a}{2}\right)\right] \tag{3.15}$$

where, F_S is the frequency band for a particular formant. f_S represents mean formant frequency computed over entire database for S = 0, 1, 2 and 3. f_a is the mean absolute error determined for each formant. f_S can be estimated in the real time for the noisy speech and f_a which is calculated over the large datasets can be used to find frequency band F_S in real time. Thus, we estimate four frequency bands ($F_0\ to\ F_3$) from the respective mean formant locations. The FFT bins corresponding to the four frequency bands are thus calculated.

3.4.2 Gain function customization based on Frequency Bands

The block diagram of developed SE method is shown in Fig. 3.1. In traditional SE methods, the gain function shown in (3.7) is applied over the entire frequency range inducing speech distortion due to inaccuracies in gain function estimation. The proposed method allows more noise suppression on acoustically unimportant bands and far lowering noise suppression on significant formant frequency bands to retain clean speech intelligibility. Thus, we obtain two different gain functions based on the frequency bands. The gain function for the formant frequency bands is given by,

$$G_{k_F} = \left[\left(\frac{1}{2\beta_F} - \frac{\mu}{\sqrt[4]{\hat{\gamma}_k \hat{\xi}_k}}\right) + \sqrt{\left(\frac{\mu}{\sqrt[4]{\hat{\gamma}_k \hat{\xi}_k}} - \frac{1}{2\beta_F}\right)^2 + \frac{v}{2\hat{\gamma}_k {\beta_F}^2}}\right] \tag{3.16}$$

$$\hat{G}_k = \begin{cases} G_{k_F}, & if\ k \in F_S\ for\ S = 0,\ 1, 2, 3 \\ G_k, & otherwise \end{cases} \tag{3.17}$$

Where k represents the k^{th} frequency bin, F_S represents the bins associated with formant frequency bands. β and β_F allow the hearing-impaired smartphone user to obtain more noise suppression without speech distortion. The β_F can be kept constant or can be adjusted along with β by HAD user in real time based on his/her listening preference under continuously varying acoustical environment. $\beta_F < \beta$ and β can be varied from 0.5 to 5.

We reconstruct the signal by considering the phase of the noisy speech signal. The final clean speech estimate is,

$$\hat{S}_k = \hat{G}_k \, Y_k \tag{3.18}$$

The time domain reconstruction signal $\hat{s}(n)$ is obtained by taking Inverse Fast Fourier Transform (IFFT) of $\hat{S}_k$. As we consider band approximations, the inaccuracy in calculating the exact formant frequency does not affect the proposed method to a great extent. Near estimation of the formant frequency bands can improve the speech intelligibility substantially. Another advantage of the proposed method is it does not induce any musical noise, eliminating the need of any post filter after the enhancement. This reduces computational complexity and latency in real time.

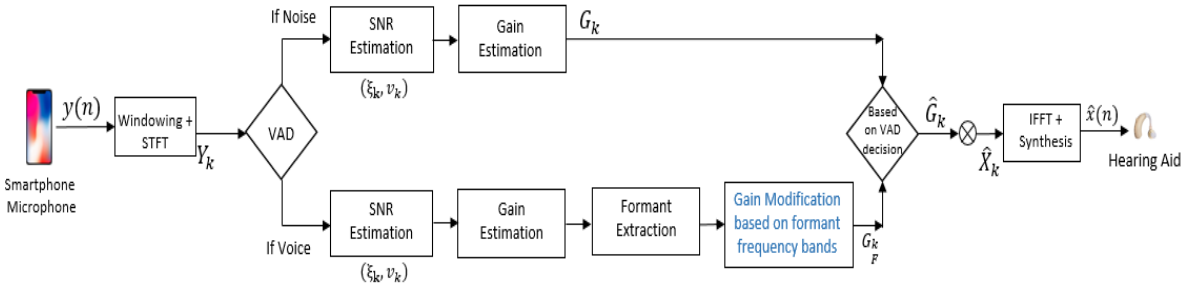


Figure 3.1. Block Diagram of Proposed SE Method

3.5 Smartphone Implementation to Function as an Assistive Device to HA

Present work considers iPhone 7 as an assistive device to HA. The data is captured using the default mic on the smartphone at a sampling rate of 48 kHz. After acquiring the input, the data is converted to float, and a frame size of 256 is used for the input buffer. Fig. 3.2 shows a snapshot of the configuration screen of the algorithm implemented on iPhone 7. Switching the 'ON' button enables SE module to process the incoming audio stream by applying the proposed SE algorithm on the magnitude spectrum of noisy speech. The enhanced signal is then played back through the HAD. Once the noise suppression is on, we have provided parameters, which allow more noise suppression without inducing speech distortion and musical noise. Varying β will change the gains over non-formant regions and varying β_F will change the gains over formant regions. From our experiments, we have seen that by setting β high and β_F low, can achieve more noise suppression and maintain the speech integrity, as we do not distort the significant bands. The user can control these parameters by adjusting its values on the touch screen panel of the smartphone to attain more noise suppression based on their level of hearing comfort. For user's ease, we have also provided a button where he/she can keep β_F set automatically based on the value of β. However, β has to be adjusted manually as keeping the β constant is not feasible in real world noisy acoustic conditions. The processing time for the frame size of 10 ms (480 samples) is 1.4ms. The smartphone application consumes very less power because of the computational efficiency of the developed algorithm. Through our experiments, we found that a fully charged smartphone can run the application for 6.2 hours on iPhone 7 which has 1960 mAh battery. We use Starkey live listen [46] to stream the data from iPhone to the HAD. The audio streaming is encoded for Bluetooth Low Energy consumption.

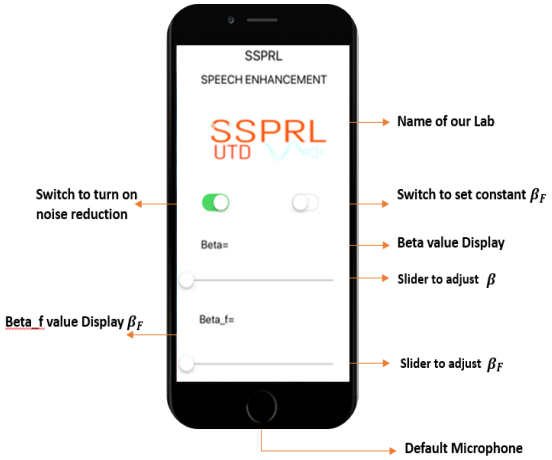


Figure 3.2. Snapshot of the developed SE application

3.6 Experimental Results and Discussion

3.6.1 Objective Evaluations

To objectively measure the performance of the proposed SE method, we use Perceptual Evaluation of Speech Quality (PESQ) [47] as it has high correlation with subjective tests than any other objective measures. The amount of noise reduction and the residual noise reduction is generally measures using segmental SNR (SegSNR), so we also use this measure to evaluate the performance of our proposed method [48 ch. 11.1]. The proposed method is compared to Log-MMSE [7] and traditional SGJMAP [9] method to evaluate the performance. The formant frequency bands were calculated by determining the mean of the formants and mean absolute error for over 300 clean speech files from TIMIT database. The experimental evaluations are performed for 3 different noise types: machinery, multi-talker babble, and traffic noise. The reported results are the average over 20 sentences from TIMIT database. For objective evaluation, noisy speech files sampled at 16 kHz, and 20ms frames with 50% overlap were considered. The β and β_F were

adjusted empirically to give the best values for both PESQ and SegSNR and for each noise type. PESQ and SegSNR values show significant improvements over Log-MMSE and SGJMAP methods for all three noise types considered. Objective measures shown in Fig. 3.3 and 3.4 reemphasize the fact that the proposed method achieves comparatively more noise suppression without distorting speech by using formants and varying user adjustable parameters in real time.

3.6.2 Subjective Evaluations

Objective tests provide useful evaluation during the development phase of the proposed method. However, the practical usability of the application can be assessed by subjective tests. We performed mean opinion scores (MOS) [49] tests on 10 normal hearing adults including both male and female subjects. The subjects were presented with the noisy speech and enhanced speech using the Log-MMSE, SGJMAP and the proposed SE methods at the SNR levels of +5 dB, 0 dB and -5 dB for 3 different noise types. Subjects were asked to rate between 1 to 5 for each speech file based on how pleasant it was and how many words they could recognize. They were also allowed to go back and change the scores after listening to other speech files. Before starting the test the subjects were instructed regarding the parameters β and β_F and were asked to set these parameters according to their listening preference. Different subjects chose to vary β and β_F for different types of noise. This test supported our claim that proposed SE method and its application is user adaptive and noise dependent. We also did field testing of our application where the acoustic environment changed dynamically. Subjective evaluation in Fig. 3.5 illustrates the usefulness of the proposed method in reducing the background noise and without inducing unpleasant musical noise or speech distortion.

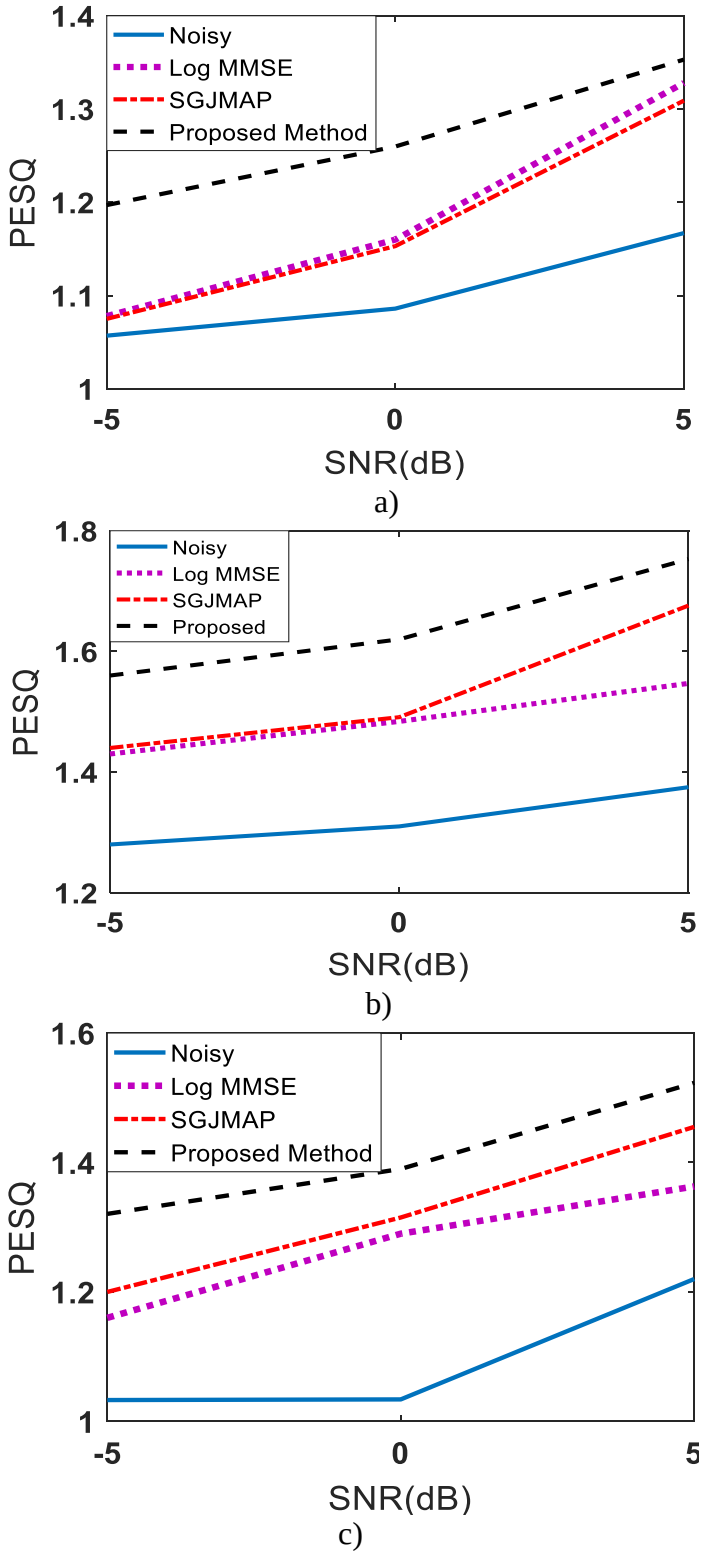
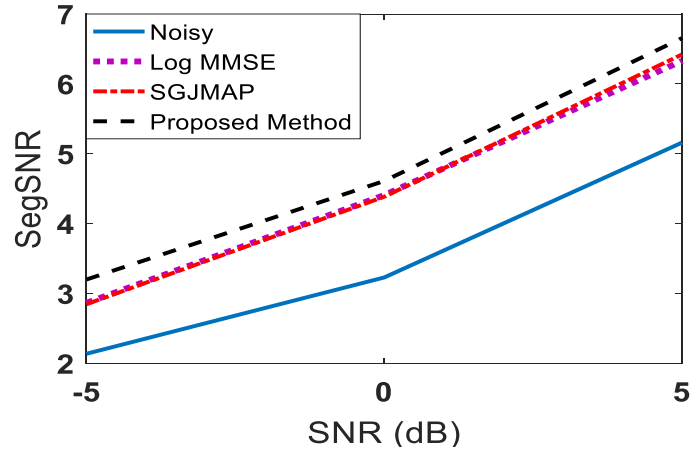
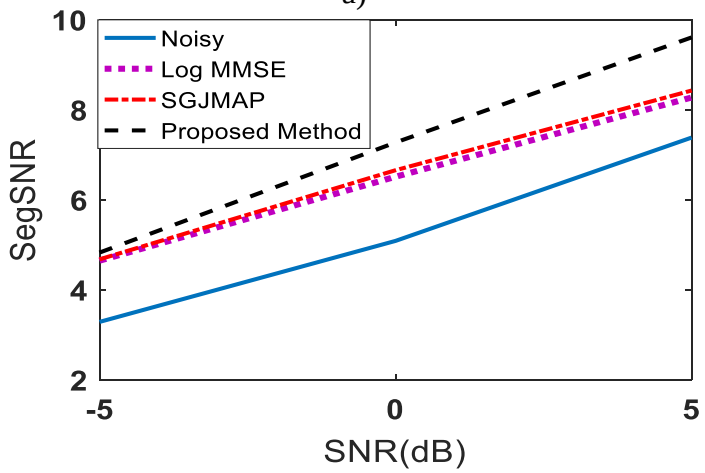


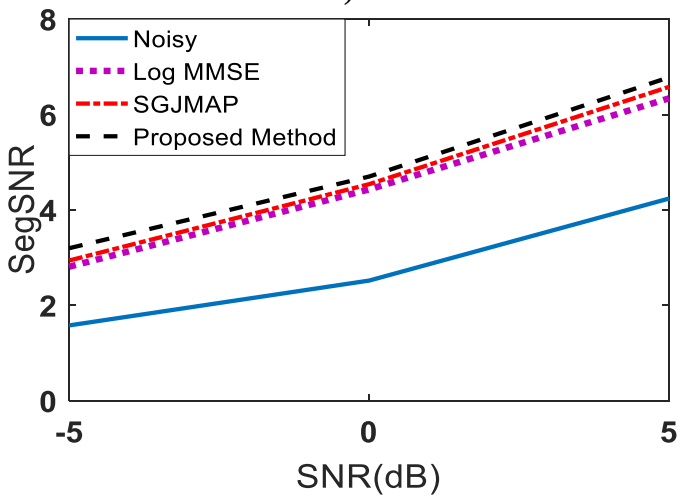
Figure 3.3. Comparison of PESQ scores a) Babble Noise b) machinery Noise c) Traffic Noise



a)



b)



c)

Figure 3.4. Comparison of Segmental SNR scores a) Babble Noise b) machinery Noise c) Traffic Noise

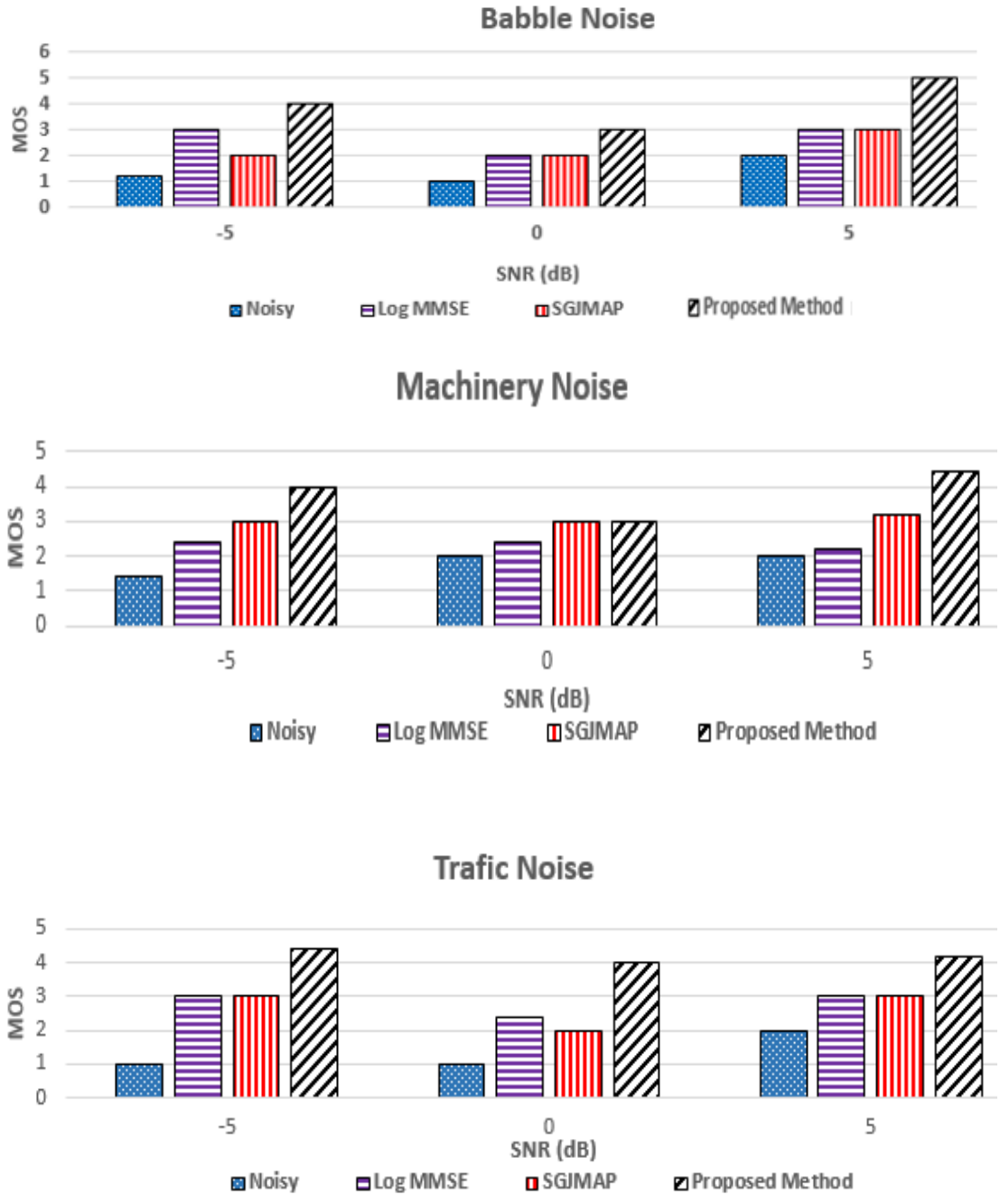


Figure 3.5. Comparison of Subjective Test scores

3.6.3 Analysis and Discussion

Spectrogram: A spectrogram is a visual representation of the spectrum of frequencies in a sound or other signals which vary with time or some other variable. Spectrograms are also called as voice prints or voice` grams and extensively used in the field of speech processing. Spectrograms are used to identify spoken words phonetically. The sample outputs on the right show select block of frequencies going up the vertical axis and time horizontal axis. The third dimension indicates the amplitude of a particular frequency at a particular time which is represented by the intensity or the color of each point. Clearly, we can see that the proposed method effectively eliminates the background noise and residual musical noise. The results are shown graphically in Fig. 3.6 using spectrograms. Actual recorded machinery noise is mixed synthetically with clean speech obtained from TIMIT database sampled at 16 KHz at an SNR of 0 dB. Periodic components of machinery noise are suppressed well in the enhanced speech. This is due to the improved noise tracking capability of the developed method.

Surface Plot Analysis with Different values of β and β_F

In this part, we see the detailed analysis of the proposed method which considers considerably less noise suppression on Formant frequency bands and more on Non-formant regions. In this evaluation, we vary the β and β_F such that, $\beta_F < \beta$. We consider 3 different Noise types (Machinery, Babble and Traffic) which is mixed with the 10 clean speech file obtained from the TIMIT database at 0 dB SNR. We calculate the PESQ values by varying β from 0.1 to 3.5 in terms of 0.05 and varying β_F from 0.1 to β. For all the three noise type, we have seen that by setting β high and β_F low, can achieve maximum PESQ values. For the Machinery noise file, β = 1.25 and β_F = 0.85, for the babble noise file β = 2.11 and β_F = 1.21 and for the traffic noise file β =

1.41 and β_F = 0.995 yielded maximum PESQ value. Similar results were obtained for all 10 files, i.e. different values of the *tradeoff* factor on different bands attained higher PESQ scores than keeping it the same over the entire frequency range. These tests and results supported our claim that proposed SE method and its application is user adaptive and can achieve more noise. Figure 3.7 shows the surface plots for different noises for attaining optimal PESQ value. The elliptical curves on the graphs show the regions which showed highest PESQ values.

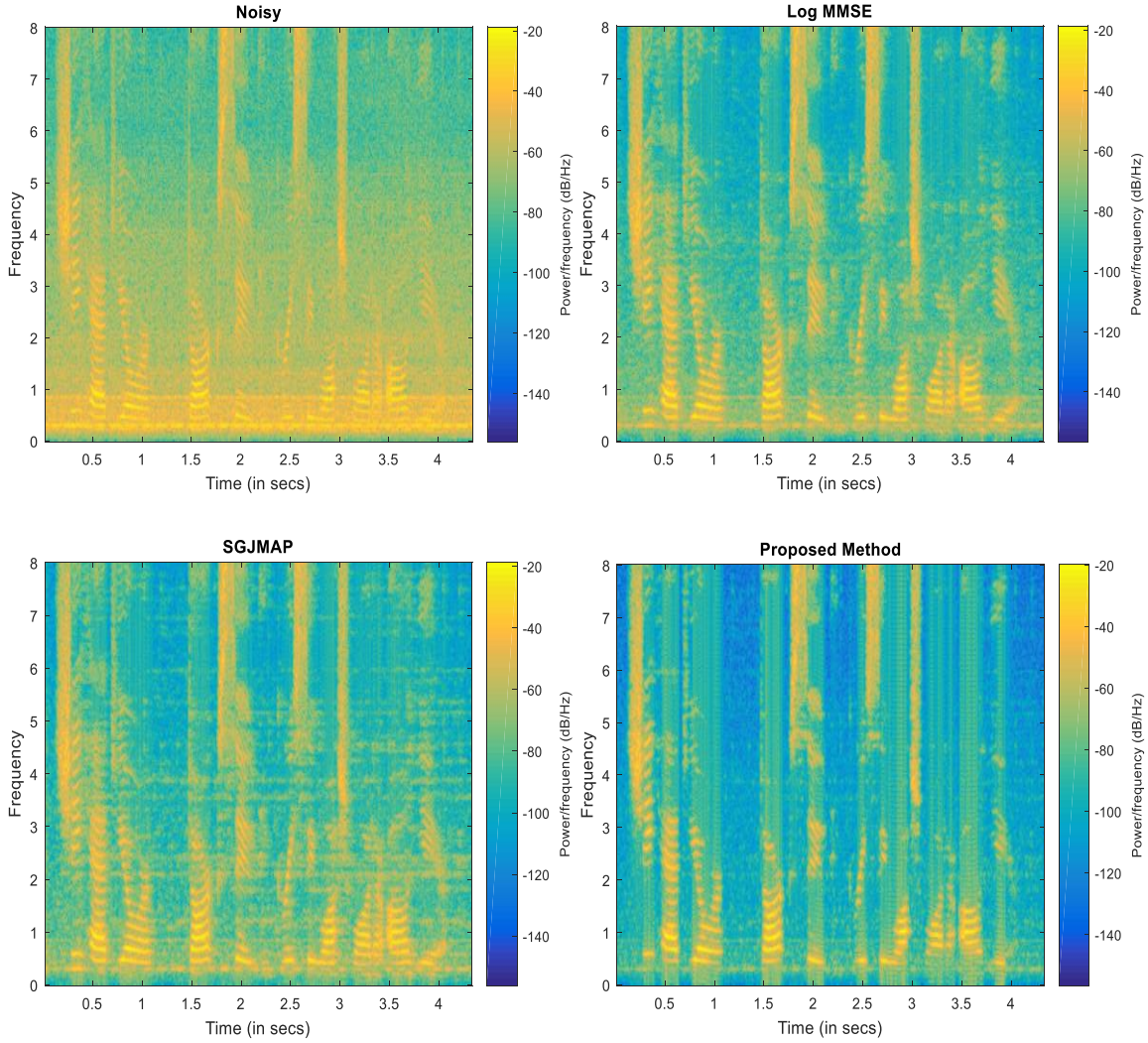
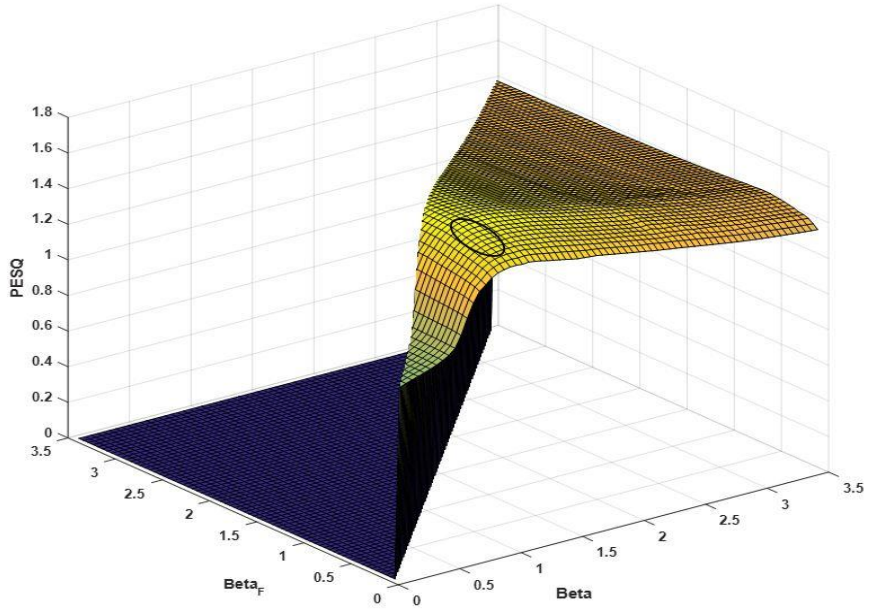
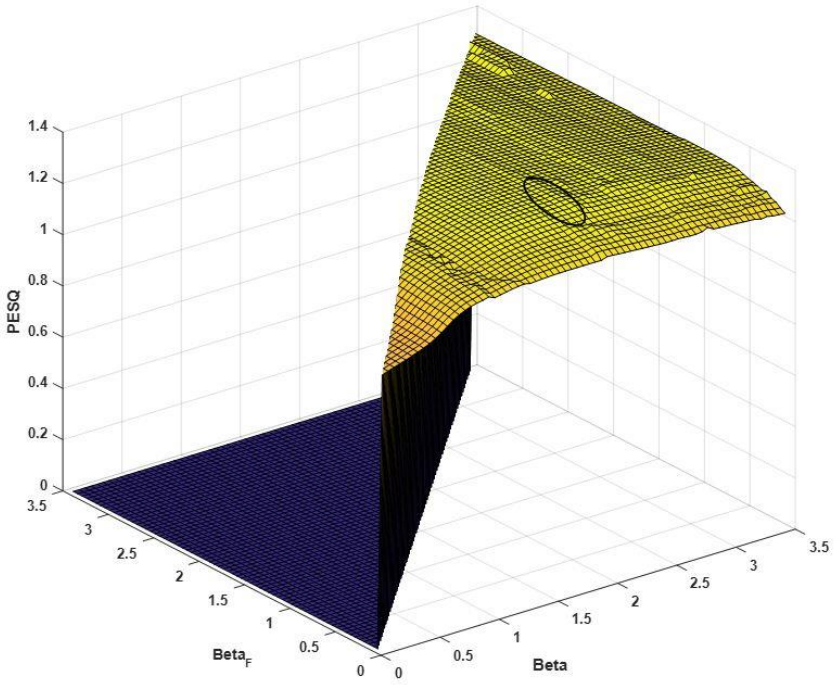


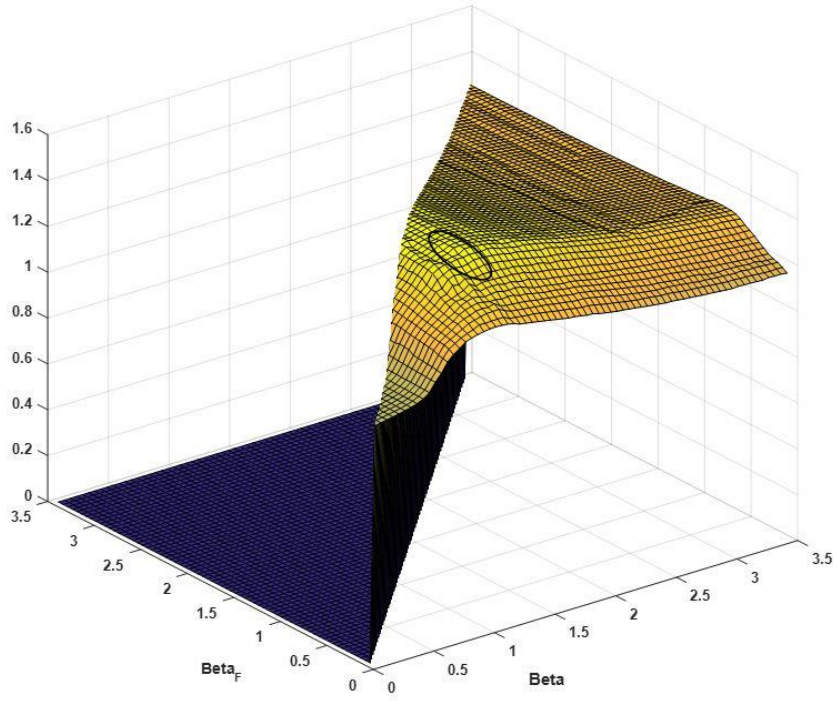
Figure 3.6. Spectrogram plot of Noisy Speech (SNR = 0 dB) and Enhanced Speech using the Log-MMSE, SGJMAP and Proposed method



a) Machinery Noise



b) Babble Noise



c) Traffic Noise

Figure 3.7. Choice of tradeoff Factors for attaining optimal value of PESQ a) Machinery Noise b) Babble Noise c) Traffic Noise

3.7 Chapter Outcomes

In this chapter, The SE method makes use of formant information in the given speech to define acoustically significant frequency bands and the gain function of the new super Gaussian Joint Maximum *a Posteriori* (SGJMAP) based SE method explained in [15] is applied to suppress more noise in acoustically unimportant bands without inducing distortion in clean speech and residual noise. The '*tradeoff*' factors that are introduced in the proposed method allows us to control the gains over formant and non-formant locations in real time allowing the hearing-impaired smartphone user to attain more noise suppression without speech distortion and thereby

maintaining the speech intelligibility. Detailed analysis of choosing the two tradeoff factors to attain the optimal intelligibility score was illustrated. The proposed method was implemented on a smartphone, which works as an assistive device for HA. The proposed method is inexpensive and computationally efficient. Objective evaluations show good improvements in quality and intelligibility showing the effectiveness of the developed method. Subjective evaluations show the usefulness of the developed smartphone application in real-world noisy conditions.

CHAPTER 4

LOG SPECTRAL AMPLITUDE ESTIMATOR BASED SINGLE MICROPHONE SPEECH ENHANCEMENT USING FORMANT INFORMATION[1]

4.1 Introduction

In this chapter, we discuss about the developed SE method based on Log spectral amplitude estimator which is an extension of the first chapter. The formant frequency analysis is applied to Log-MMSE speech enhancement technique i.e. an MMSE estimator based on the formant frequency is proposed. We apply scaled value of the MMSE gain function over different bands which is estimated based on formant locations. The formant frequency based SE method presented in this work makes use of the *priori* information of formants to apply scaled value of the Minimum mean square error Log spectral amplitude estimator (Log-MMSE) gain function on the acoustically unimportant bands which in turn suppresses the background noise without inducing speech distortion and any residual musical noise so there is no requirement of any post filter after the enhancement. The 'scaling' factor for the gain in the non-formant locations can be varied in real time allowing the hearing impaired smartphone user to control the amount of noise suppression and speech distortion. The proposed SE method is computationally efficient, and inexpensive. Objective and subjective evaluations of the proposed method show good improvement in quality and intelligibility reveals the overall usability of the developed algorithm.

[1] © [2017] IEEE. Reprinted, with permission, from [G. S. Bhat, N. Shankar, C. K. A. Reddy and I. M. S. Panahi, "Formant frequency-based speech enhancement technique to improve intelligibility for hearing aid users with smartphone as an assistive device," 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)]

4.2 Conventional Log-Spectral Amplitude Estimator

In the Log-MMSE method, speech and noise models are considered to be statistically independent Gaussian Random Variables. The goal is to minimize the mean squared error of log magnitude spectra between estimated and true speech signals. Considering the additive mixture model for noisy speech $y(n)$, with clean speech $s(n)$ and noise $w(n)$, as

$$y(n) = s(n) + w(n) \tag{4.1}$$

The noisy k^{th} Discrete Fourier Transform (DFT) coefficient of $y(n)$ for frame λ is given by,

$$Y_k(\lambda) = S_k(\lambda) + W_k(\lambda) \tag{4.2}$$

Where S and W are the clean speech, and noise DFT coefficients respectively. In polar coordinates, (2) can be written as,

$$R_k(\lambda)e^{j\theta_{Y_k}(\lambda)} = A_k(\lambda)e^{j\theta_{S_k}(\lambda)} + B_k(\lambda)e^{j\theta_{W_k}(\lambda)} \tag{4.3}$$

Where $R_k(\lambda)$, $A_k(\lambda)$, $B_k(\lambda)$ are magnitude spectra of noisy speech, clean speech, and noise respectively. $\theta_{Y_k}(\lambda)$, $\theta_{S_k}(\lambda)$, $\theta_{W_k}(\lambda)$ are the phase spectra of noisy speech, clean speech and noise respectively. Looking at the estimator $\hat{A}_k$, which minimizes the distortion measure as explained in [6], the mean-square error of the log-magnitude spectra is given by,

$$E\left\{\left(log\ A_k\ -\ log\ \hat{A}_k\right)^2\right\} \tag{4.4}$$

Where, A_k is the k^{th} bin of magnitude spectrum, and $\hat{A}_k$ is the k^{th} bin of estimated clean speech magnitude spectrum. The optimal log-MMSE estimator can be obtained by evaluating the conditional mean of the $log\ A_k$, that is,

$$log\ \hat{A}_k = E\{log\ A_k\ |Y_k(\lambda)\} \tag{4.5}$$

Hence, the estimate of the speech magnitude is given by,

$$\hat{A}_k = exp(E\{log\, A_k |\, Y_k(\lambda)\}) \tag{4.6}$$

Solving the above expectation, the final estimate of speech magnitude spectrum according to [7] is given by,

$$\hat{A}_k = \frac{\xi_k}{\xi_k+1} \exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}\mathrm{dt}\right\} R_k \tag{4.7}$$

$$\triangleq \boldsymbol{G_{LSA}(\xi_k, v_k) R_k}$$

Where $v_k = \frac{\xi_k}{1+\xi}\gamma_k$ here $\xi_k = \frac{\sigma_{S_k}^2}{\sigma_{W_k}^2}$ is the *a priori* SNR and $\gamma_k = \frac{R_k^2}{\sigma_{W_k}^2}$ is the *a posteriori* SNR. $\sigma_{W_k}^2$ is estimated using a voice activity detector (VAD) [50]. σ_{S_k} is the estimated instantaneous clean speech power spectral density. The optimal phase spectrum is the noisy phase itself $\theta_{S_k} = \theta_{Y_k}$.

4.3 Proposed SE Method

In this section we describe the developed speech processing method implemented on Smartphone. Fig. 4.1 shows the block diagram of the proposed method that runs on the smartphone in real time (Timing is discussed in section 4.4).

4.3.1 Formant Frequency Band Estimation

In this method, we approximate the formant frequency bands by using the method shown in chapter 1 i.e. by calculating the exact formant locations and the first four formant frequency trajectories $(f_0 - f_3)$ of the clean speech or speech degraded with noise at high SNR can be calculated by the method explained in [38] which employs adaptive voice detector, and gender detector for formant extraction from the voice segments of continuous speech. We require a frequency range to approximate the presence of speech and apply considerably less noise

suppression over that band. Therefore, the mean of formants $f_0 - f_3$ are calculated for large data sets and mean absolute error for each formant is determined over the data sets to find the frequency band of probable formant location. The frequency band is given by (4.8)

$$F_S = \left[\left(f_S - \frac{f_a}{2}\right), \left(f_S + \frac{f_a}{2}\right)\right] \tag{4.8}$$

Where F_S is the frequency band for a particular formant. f_S represents mean formant frequency computed over entire database for S = 0, 1, 2 and 3. f_a is the mean absolute error determined for each formant. Thus, we estimate four frequency bands (F_0 to F_3) from the respective mean formant locations. The FFT bins corresponding to the four frequency bands are thus calculated.

4.3.2 Gain function Customization based on Frequency Bands

In conventional methods, the gain function shown in (4.7) is usually applied over the entire frequency range. This induces speech distortion if the estimate of gain function is inaccurate. The proposed method allows suppressing more noise on acoustically unimportant bands and far lowering noise suppression on formant frequency bands to retain the integrity of clean speech. Thus, we obtain two different gain functions based on the frequency bands.

$$\hat{G}_k = \begin{cases} G_{LSA}(\xi_k, v_k), & if\ k \in F_S\ for\ S = 0,\ 1, 2, 3 \\ \delta\ G_{LSA}(\xi_k, v_k), & otherwise \end{cases} \tag{4.9}$$

Where k represents the k^{th} frequency bin, F_S represents the bins associated with formant frequency bands. δ represents the scaling factor which allows the smartphone user to obtain more noise suppression without speech distortion. The δ ranges from 0 to 1 which the HAD user can adjust in real time based on his/her listening preference under continuously varying acoustical environment. We know from the literature that the phase is perceptually unimportant. Therefore,

we consider the phase of the noisy speech signal for reconstruction. The final clean speech estimate is,

$$\hat{S}_k = \hat{G}_k \, Y_k \tag{4.10}$$

The time domain reconstruction signal $\hat{s}(n)$ is obtained by taking Inverse Fast Fourier Transform (IFFT) of $\hat{S}_k$. At lower values of δ there is more noise suppression. When $\delta = 1$ the algorithm acts as basic Log-MMSE method. The inaccuracies in calculating the exact formants does not affect the proposed method to a greater extent as it considers the band approximation. Near approximation of the formant frequency bands can improve the speech intelligibility substantially. Another advantage of the proposed is it does not induce any residual or musical noise so there is no requirement of any post filter after the enhancement. This reduces computational complexity and latency in real time.

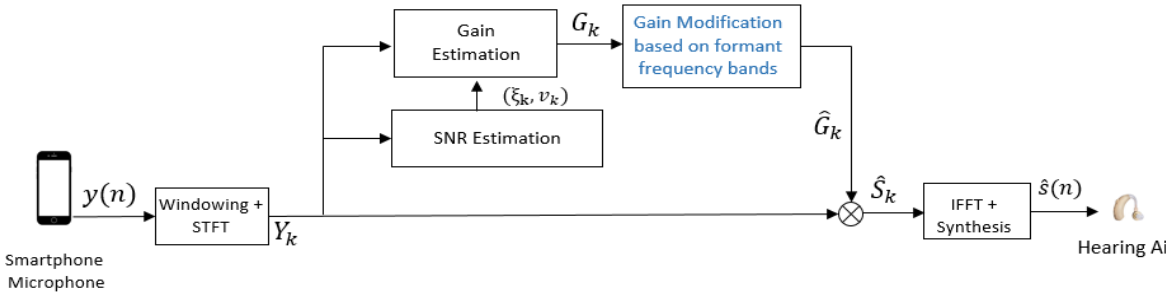


Figure 4.1. Block Diagram of the Proposed SE Method

4.4 Real Time Implementation on Smartphone to Function as an Assistive Device to HA

As considered in Chapter 1, in this work, iPhone 7 running iOS 10.3 operating system is considered as a HA assistive device. The default microphone (Figure 4.2) on iPhone 7 is used to capture the

audio signal, process the data and wirelessly transmit the enhanced signal to the HAD. Xcode [51] is used for coding and debugging of the SE algorithm. The data is acquired at a sampling rate of 48 kHz. Core Audio [52], an open source library from Apple Inc. was used to carry out input/output handling. After input callback, the short data is converted to float, and a frame size of 256 is used for the input buffer. Fig. 4.2 shows a snapshot of the configuration screen of the algorithm implemented on iPhone 7. When the switch button shown is in 'OFF' mode, the application merely plays back the audio through the smartphone without processing it. Switching 'ON' the button enables SE module to process the incoming audio stream by applying the proposed noise suppression algorithm on the magnitude spectrum of noisy speech. The enhanced signal is then played back through the HAD. Once the noise suppression is on, we have provided a parameter, which allows more noise suppression without inducing any speech distortion and musical noise. Unlike in this chapter 1, here we have only one parameter to control the gains over the non-formant regions of the frequency range. In (4.10), the gain function depends on δ which needs to be determined empirically. Through our experiments, it is known that the optimal value of δ depend on the type of noisy signal, background noise and acoustic environment. Hence, it is not advisable to fix the values of δ irrespective of changing conditions. In our smartphone application, the user can control this parameter by adjusting its value on the touch screen panel of smartphone to attain more noise suppression based on their level of hearing comfort. The processing time for a frame of 10ms (480 samples) is 1.4ms. Computation efficiency of the developed algorithm allows the smartphone application to consume very less power. Through our experiments, we found that a fully charged smartphone can run the application seamlessly for 6.3 hours on iPhone 7 with 1960

mAh battery. We use Starkey live listen [46] to stream the data from iPhone to the HAD. The audio streaming is encoded for Bluetooth Low Energy consumption.

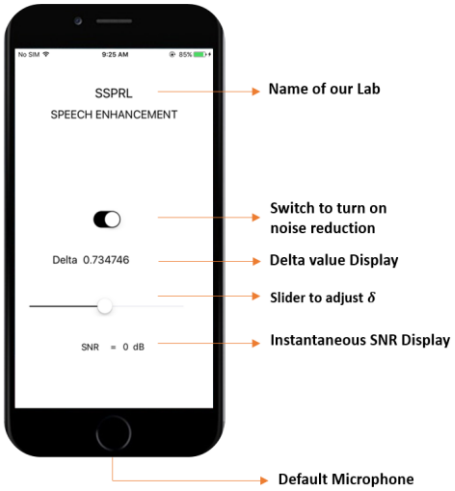


Figure 4.2. Snapshot of developed SE method

4.5 Experimental Results

4.5.1 Objective Evaluation

To the best of our knowledge, there are no SE algorithms published that use formant frequencies and a user adjustable parameter to attain more noise suppression in particular bands and thereby personalizing and retaining the speech quality and intelligibility simultaneously in real time varying noisy conditions. We therefore fix the values of few parameters and evaluate the performance of the proposed method by comparing with Log-MMSE [7] method which has been known to show promising results. We note that the proposed SE method is an improved extension of the Log-MMSE method. In our experiment, the formant frequency bands were calculated by determining the mean of the formant locations and mean absolute error for over 300 clean speech

files from TIMIT database. The experimental evaluations are performed for 3 different noise types: machinery, multi-talker babble and traffic noise. The reported results are the average over 30 sentences from TIMIT database. For objective evaluation, all the files are sampled at 16 kHz, and 20ms frames with 50% overlap are considered. As objective evaluation criteria, we choose the perceptual evaluation of speech quality (PESQ) [47] as it had better correlation with subjective tests than the other objective measure. Another objective measure is Segmental SNR (SegSNR) as the amount of noise reduction, residual noise and speech distortion is generally measured by SegSNR [48]. Fig. 4.3 and 4.4 show the plots of SegSNR and PESQ versus SNR for the 3 noise types. The δ was adjusted empirically to give the best values for both PESQ and SegSNR and for each noise type. PESQ and SegSNR values show statistically significant improvements over Log-MMSE method for all three noise types considered. Objective measures reemphasize the fact that the proposed method achieves comparatively more noise suppression without distorting speech.

4.5.2 Subjective Test setup and Evaluation

The objective measures are extremely useful evaluation results during the development phase of our method, they give very little information about the practical usability of our application. We performed Mean Opinion Score (MOS) tests [49] on 11 normal hearing subjects including male and female adults. Subjects were presented with noisy speech and enhanced speech using the proposed, and Log-MMSE methods at SNR levels of -5 dB, 0 dB and 5 dB. Subjects were asked to rate between 1 and 5 for each audio file based on how pleasant it is and how many words they can identify. They were also given the flexibility to go back change the score after listening to all the other audio files. This test provided a good comparison between proposed method and Log-

MMSE method. The key contribution of this paper is in providing the user the ability to customize the parameter to suppress more noise without compromising much on intelligibility. Before

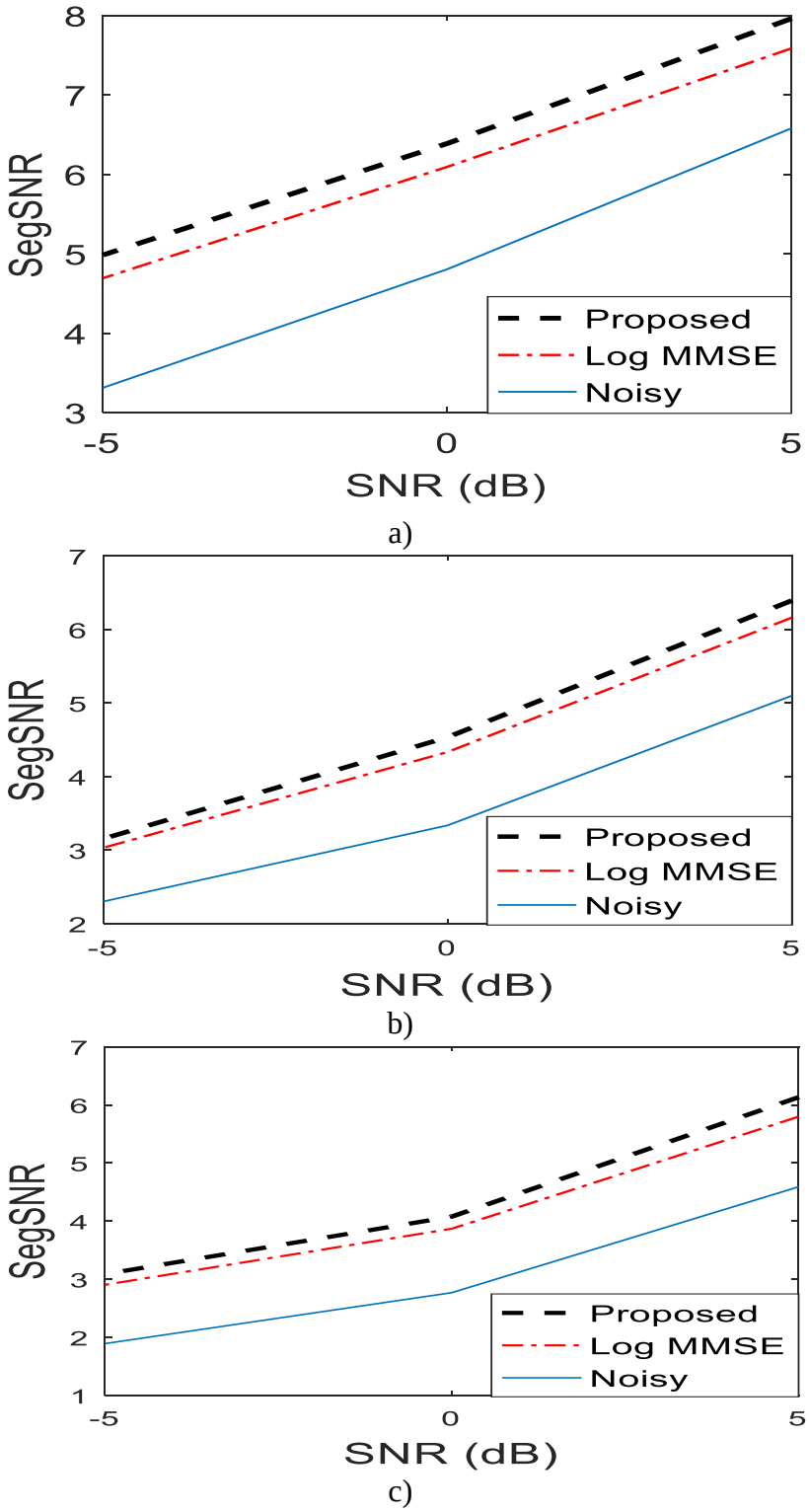


Figure 4.3. Comparison of Segmental SNR scores for (a) Machinery noise, (b) Babble noise and (c) Traffic noise

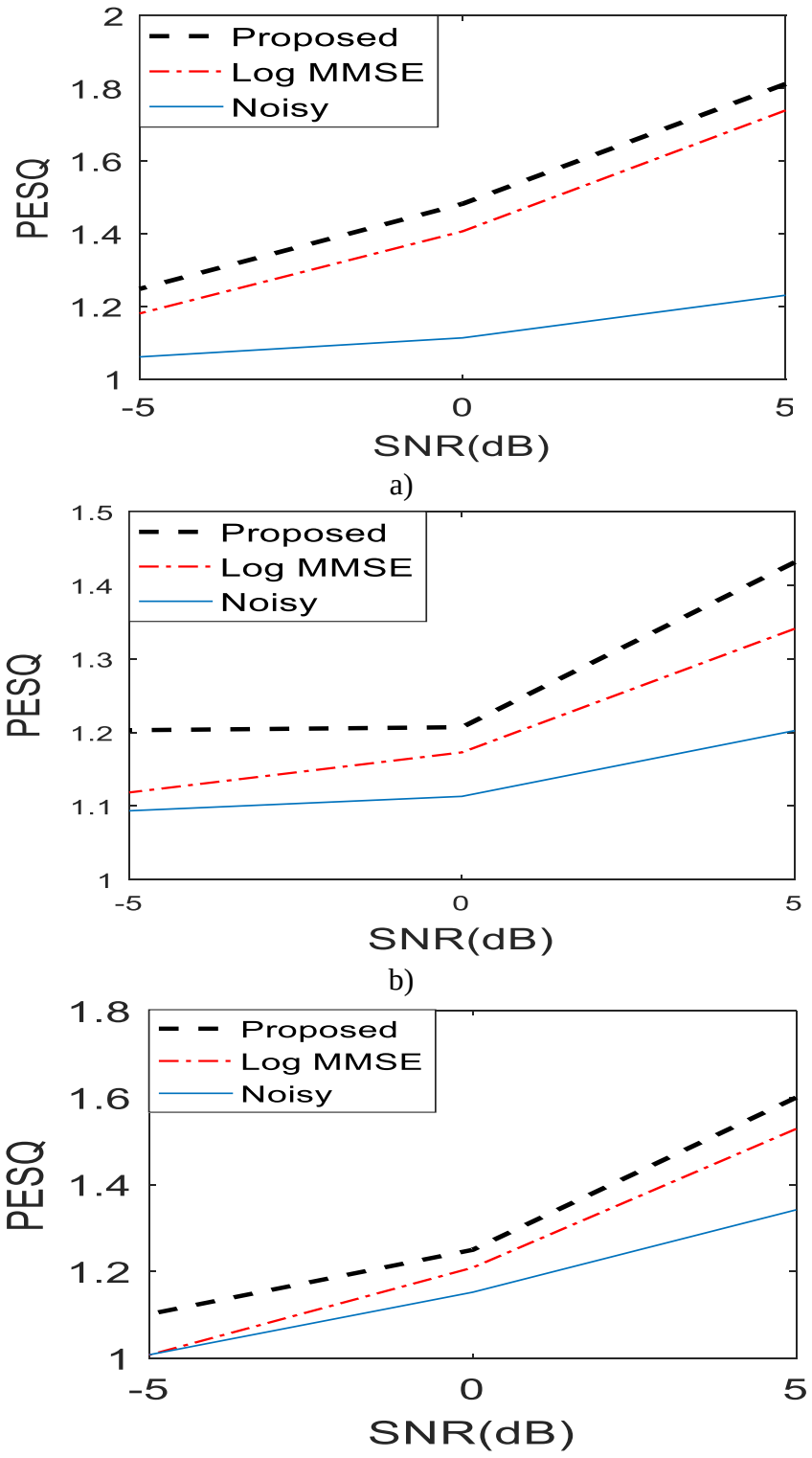


Figure 4.4. Comparison of PESQ scores for (a) Machinery noise, (b) Babble noise and (c) Traffic noise

starting the actual tests, the subjects were instructed to set δ for each noise type as per their preference. One key observation was, the preferred value of δ varied across subjects. This supports our claim that the proposed SE method and its developed application is user adaptive. We also conducted field test of our application in real world noisy conditions, which change dynamically. Subjective test results in Fig. 4.5 illustrate the effectiveness of the proposed method in reducing the background noise and musical noise, simultaneously preserving the quality and intelligibility of the speech.

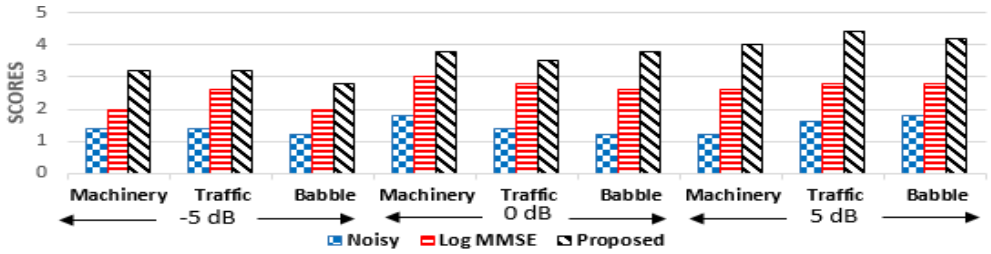


Figure 4.5. Comparison of Subjective results

4.6 Chapter Outcomes

We developed a formant frequency based single microphone SE technique by introducing two gain functions depending on acoustical importance. The resulting gain allows smartphone user to suppress more noise on non-formant bands by retaining speech intelligibility and strike a balance between the amount of noise suppression and speech distortion. The proposed algorithm was implemented on a smartphone device, which works as an assistive device for HA. The objective and subjective results demonstrate the usability of the method in real world noisy conditions.

CHAPTER 5

COMPUTATIONALLY EFFICIENT TWO MICROPHONE SPEECH ENHANCEMENT FOR CONVOLUTIVE MIXTURES

5.1 Introduction

In the previous chapters, the single microphone SE techniques were discussed. In this chapter we discuss about real-time dual channel BSS technique. Recently, microphone array-based SE techniques have been widely accepted as a viable solution for noise suppression. Generally, by increasing the number of microphones in a speech enhancement system, further noise reduction is expected [29]. However, the design of a microphone array for hearing aids faces serious difficulties in terms of size, power consumption and computational efficiency. Therefore, dual microphone speech enhancement systems can be considered as a tradeoff. Blind Source separation has gained a lot of limelight in the past decade and researchers are using BSS techniques for enhancing speech. BSS techniques use the information of the mixed signals at the input and they have shown promising results in source separation by using dual microphones [13]. ICA is one of the most commonly used BSS technique [14]. Generally, in the real-world signals, often the mixing procedure is convolutive in nature. ICA fails to separate the sources when tested with actual recorded data because ICA is designed to find an optimal demixing matrix for instantaneous mixtures of source signals in time domain and not for acoustical signals having convolutive mixtures. Hence a BSS method that can separate the two convolutedly mixed sources is essential. Frequency Domain Blind Source Separation (FDBSS) approach [53-55] can be used to extend ICA to convolutive mixtures. The time domain convolutive mixture can be transformed into frequency domain instantaneous mixtures at individual frequency bins by using Short Time Fourier

Transform (STFT). Now ICA can be used for source separation at individual frequency bin [56, 57]. But calculation of demixing matrices may permute the sources at each frequency bin causing permutation problem. In order to solve the permutation problem, an extension to ICA [58], Independent Vector Analysis (IVA) was developed. The sources in the IVA model are not scalars but are considered as vectors. The optimization of IVA not only considers the independency between sources, but also takes the inter dependency in each source vector into account. Hence the permutation problem is inherently solved and there is no post processing stage required to align the components compared to ICA. The conventional IVA uses Kullback-Leibler (KL) divergence as an objective function and the demixing matrices at each frequency bin is optimized using natural gradient-based updating rule.

5.2 IVA Formulation

5.2.1 Frequency domain Source Separation model

The real-world noisy acoustic environment mixing process include delays, losses and reverberations, i.e., signals are time delayed and convolved. Consider there are P sources and Q sensors,

$$y_q(t) = \sum_{p=1}^{P} \sum_{\tau}^{T-1} h_{qp}(t) * x_p(t-\tau) \tag{5.1}$$

where (Q>=P), (*) is the convolution. $h_{qp}(t)$ is the time domain finite impulse response mixing filter from source p to sensor q which has T length in time. $x_p(t)$ is the pth source signal at time t. In our application, the number of sensors and the number of sources are equal and limited to two. The time domain equation will be as follows,

$$y_1(t) = h_{11}(t) * x_1(t) + h_{12}(t) * x_2(t) \tag{5.2}$$

$$y_2(t) = h_{21}(t) * x_1(t) + h_{22}(t) * x_2(t) \tag{5.3}$$

Continuing with the derivation and applying STFT, the time domain expression in (5.1) can be converted to frequency domain given by,

$$y_q{}^{[f]}(n) = \sum_{p=1}^{P} h_{qp}{}^{[f]}\, x_p{}^{[f]}(n) \tag{5.4}$$

$$\boldsymbol{y}^{[f]}(n) = \boldsymbol{H}^{[f]}\boldsymbol{x}^{[f]}(n) \tag{5.5}$$

Here, $x_p{}^{[f]}(n)$, $y_q{}^{[f]}(n)$ and $h_{qp}{}^{[f]}$ are frequency domain representation of $x_p(t)$, $y_q(t)$ and $h_{qp}(t)$ respectively. $\boldsymbol{y}^{[f]}(n) = [y^{[f]}(n), \ldots\ldots, y_q^{[f]}(n)]$, $\boldsymbol{x}^{[f]}(n) = [x_1{}^{[f]}(n), \ldots\ldots\ldots, x_P{}^{[f]}(n)]$ and $\boldsymbol{H}^{[f]}$ is the mixing matrix for each frequency bin index *f*, with $h_{qp}{}^{[f]}$ as its entries for each frame *n*.

The goal of IVA or any BSS method is to find a demixing matrix $\boldsymbol{G}^{[f]}$ at each individual frequency bin *f* such that,

$$\boldsymbol{s}^{[f]}(n) = \boldsymbol{G}^{[f]}\boldsymbol{y}^{[f]}(n) \tag{5.4}$$

where $\boldsymbol{s}^{[f]}(n)$ is the estimate of $\boldsymbol{y}^{[f]}(n)$.

In (5.4), if algorithms like ICA are used, the permutation problem should be solved, otherwise, the source separation fails miserably. On the other hand, to solve the permutation problem, IVA makes use of inter-frequency bin information. The difference between ICA and IVA is that, signals are considered as vectors and not scalars, also the optimization of IVA uses multivariate variables instead of univariate scalars.

5.2.2 Objective function

Mutual information is used as an objective function in most of the BSS algorithms. Kullback-Leibler (KL) divergence method is used to calculate the mutual information and is given by,

$$I(\boldsymbol{s}) = KL(p_{\boldsymbol{s}}||\prod_{p} p_{s_p}) = \int p_{\boldsymbol{s}}(z)\log\frac{p_{\boldsymbol{s}}(z)}{\prod_s p_{y_s}(z_s)}dz \tag{5.5}$$

where $p_{\boldsymbol{s}}$ is the probability density function (PDF) of a random vector $\boldsymbol{s}$. p_{s_p} denotes the pth marginal PDF of $\boldsymbol{s}$ and z is a dummy variable used for the integral. The objective function of IVA is that, each $\boldsymbol{y}_p$ is a vector instead of a scalar. The objective function of BSS using IVA is given by,

$$J_{IVA} = KL(p_{\boldsymbol{s}}||\prod_{p} p_{\boldsymbol{s}_p})$$

$$= \sum_{p=1}^{P} \mathrm{H}(\boldsymbol{s}_p) - \mathrm{H}(\boldsymbol{s}_1; \dots\dots\dots; \boldsymbol{s}_P)$$

$$= \sum_{p=1}^{P} \mathrm{H}(\boldsymbol{s}_p) - \mathrm{H}(\boldsymbol{s}^{[1]}; \dots\dots\dots; \boldsymbol{s}^{[F]})$$

$$= \sum_{p=1}^{P} \mathrm{H}(\boldsymbol{s}_p) - \mathrm{H}(\boldsymbol{G}^{[1]}\boldsymbol{y}^{[1]}; \dots\dots\dots; \boldsymbol{G}^{[F]}\boldsymbol{y}^{[F]}) \tag{5.6}$$

$$= \sum_{p=1}^{P} \mathrm{H}(\boldsymbol{s}_p) - \mathrm{H}(\boldsymbol{G}\boldsymbol{y}) \tag{5.7}$$

$$= \sum_{p=1}^{P} \mathrm{H}(\boldsymbol{s}_p) - \sum_{f=1}^{F} \log\left|\det(\boldsymbol{G}^{[f]})\right| - C \tag{5.8}$$

In (5.8), for a linear invertible transformation $\boldsymbol{G}$, $\mathrm{H}(\boldsymbol{Gy}) = \log|\det(\boldsymbol{G})| + \mathrm{H}(\boldsymbol{y})$ and the determinant of diagonal matrix is given by, $\det(\boldsymbol{G}) = \prod_{f=1}^{F} \det(\boldsymbol{G}^{[f]})$. Here we have to note that H represents the entropy of the signal and not the mixing matrix. Since the observed signals will not change in the optimization procedure [59], the term $C = \mathrm{H}(\boldsymbol{y})$ is a constant. Assuming that the observed signals are zero mean and whitened at individual frequency bin, the term $\sum_{f=1}^{F} \log\left|\det\left(\boldsymbol{G}^{[f]}\right)\right|$ becomes zero. Also by noting that, $\mathrm{H}(\boldsymbol{s_p}) = \sum_{\boldsymbol{f=1}}^{\boldsymbol{F}} \mathrm{H}\left(s_p{}^{[f]}\right) - I(\boldsymbol{s}_p)$, the objective function in (5.8) becomes,

$$J_{IVA} = \sum_{p=1}^{P} \left(\sum_{f=1}^{F} \mathrm{H}\left(s_p{}^{[f]}\right) - I(\boldsymbol{s}_p)\right) \tag{5.9}$$

Minimizing the $\mathrm{H}\left(s_p{}^{[f]}\right)$ and maximizing the $I\left(\boldsymbol{s}_p\right)$ term balances the minimization of (5.9). Here, non-Gaussianity property is used to measure the independency, and in order to increase the non-Gaussianity which is responsible for separating the sources at each frequency bin, $\mathrm{H}\left(s_p{}^{[f]}\right)$ is minimized. On the other hand, to increase the inter dependency of variables in $\boldsymbol{s}_p$, $I(\boldsymbol{s}_p)$ is maximized and this in turn solves the permutation problem by itself.

The objective function shown in (5.8) is minimized by considering the estimate of the entropy functions of the sources. The actual PDF of each $\boldsymbol{s}_p$ is not available and a prior target PDF $\hat{p}(\boldsymbol{s_p})$ is used. Therefore the objective function becomes,

$$J_{IVA} = -\sum_{p} E(\log \hat{p}(\boldsymbol{s_p})) \tag{5.10}$$

5.2.3 Optimization of Objective Function

Learning Algorithm

Using the natural gradient based approach [59, 60] we minimize the cost function by differentiating the objective function with respect to the demixing matrices,

$$\Delta \boldsymbol{G}^{[f]} = (\boldsymbol{I}\text{-E}\,[\Phi^{[f]}(\boldsymbol{s}^{[f]})(\boldsymbol{s}^{[f]})^{\mathrm{H}}])\,\boldsymbol{G}^{[f]} \tag{5.12}$$

The coefficients of separating matrices can be updated with the batch update rule. The batch update rule is given by [60],

$$\boldsymbol{G}^{[f]} = \boldsymbol{G}^{[f]} + \eta\,\Delta \boldsymbol{G}^{[f]} \tag{5.11}$$

The final update equation is given by,

$$\boldsymbol{G}^{[f]} = \boldsymbol{G}^{[f]} + \eta\,(\boldsymbol{I}\text{-E}\,[\Phi^{[f]}(\boldsymbol{s}^{[f]})(\boldsymbol{s}^{[f]})^{\mathrm{H}}])\,\boldsymbol{G}^{[f]} \tag{5.13}$$

In (5.13), the learning rate is η, and the multivariate nonlinear function for frequency bin f is $\Phi^{[f]}(.)$. This can be termed as Multivariate score function and the nonlinear function is related to the chosen source prior PDF:

$$\Phi^{[f]}(\boldsymbol{s_p}) = -\frac{\partial \log \hat{p}\left(s_p^{[1]}, \ldots\ldots, s_p^{[F]}\right)}{\partial s_p^{[f]}} \tag{5.14}$$

Optimizing the Multivariate Score Function

Among a number of functional forms for $\Phi^{[f]}(\boldsymbol{s_p})$, one of the simplest but effective one is given as follows:

$$\Phi^{[f]}(\boldsymbol{s_p}) = \frac{\partial \sqrt{\sum_{f=1}^{F}\left|s_p^{[f]}\right|^2}}{\partial s_p^{[f]}} = \frac{\boldsymbol{s_p^{[f]}}}{\sqrt{\sum_{f=1}^{F}\left|\boldsymbol{s_p^{[f]}}\right|^2}} \tag{5.15}$$

Though we use the fixed form of the score function, the multivariate score function might vary with different types of dependencies [60]. The source pdf is defined as a dependent multivariate super-Gaussian distribution, which can be written as

$$p(\boldsymbol{s_p}) = \alpha \exp(-\sqrt{(\boldsymbol{s_p} - \mu_p)^T \Sigma_p^{-1} (\boldsymbol{s_p} - \mu_p)} \quad (5.17)$$

where μ_p and Σ_p^{-1} are the mean vector and inverse of covariance matrix of the $p - th$ source signal, respectively. By assuming zero mean and identity covariance matrix, we obtain (5.17).

5.3 Drawbacks and Implementation Challenges of Conventional IVA

Conventional IVA methods consider an entire length of a signal in order to find the demixing matrix at each frequency bin but in order to use IVA in real time, the algorithm should be operated frame wise as we receive the input signals at the microphone in frames. Also to use IVA in our smartphone-HA setup, which are also real-time systems, the computational complexity of the algorithm should be very low. A new BSS method for convolutive mixtures with fast convergence and low computational complexity is explained in [61] based on Auxiliary-function (AuxIVA). Several variations on IVA are proposed for real time operations [62, 63]. However, these methods cannot effectively counteract the processing delay of each frame. In the case of IVA, a delay of at least one frame length is necessary for frame analysis [64]. This delay depends on the incoming signal, computational complexity of the algorithm, and the processing power of the platform on which the algorithm is running which in our case is a smartphone. For example, if a 50ms frame is processed, the processing time of that frame should be less than 50ms so that by the time a new audio frame comes into the microphone, the previous frame should be processed and be an output

frame. IVA involves iterative approach to calculate the de-mixing matrix [62]. The number of iterations required to converge and achieve good source separation depends on the characteristics of signal. Hence, there is high possibility of the processing time exceeding the frame size. Such a large delay causes various problems to HA users such as discomfort due to the loss of lip synchronization and difficulty in speaking due to the delayed feedback. This delay also induces skipping of audio frames, causing distortion in the separated speech. This decreases the quality and speech intelligibility.

5.4 Proposed Real-time IVA

This framework is based on the idea that the impulse response between the microphone and the speech source do not change significantly in short durations of time. In the proposed method we use the formulation discussed in section 5.2 to calculate the demixing matrix for first 100 frames of the signal, we use this demixing matrix for the rest of the signal until we detect the change in the impulse response. The change in the impulse response is detected by tracking the change in the location of the source, which is achieved using the Neural Network based DOA estimation proposed in [65]. Whenever the change in the DOA is detected, the demixing matrix for successive 100 frames of the signal is calculated and used for the separating the two sources. Convolutedly mixed noisy speech from the two microphones $x_1(t)$ and $x_2(t)$ are processed to estimate the DOA using the proposed Neural Network approach only in the speech portions. A reliable VAD based on Spectral flux [66] is used to identify the voice only parts. Once the DOA is estimated, the algorithm checks for a change in the direction of arrival of the source, we use this as a criterion to update the demixing matrix. If the criterion is satisfied, the demixing matrix is updated using the new data that reflects the change in the angle of the source. If the criterion is not satisfied, the old

demixing matrix is used to separate the sources. $y_1(t)$ and $y_2(t)$ are the separated sources. This approach saves lot of computations. The separated sources can also be input to the VAD, this improves the performance of the VAD. Whenever there is change in the source direction, though the source separation is inaccurate, the SNR of the output signal will be much better than the SNR of noisy signal alone. Figure 5.1 shows the flowchart of the proposed Real-Time implementation framework of IVA.

5.5 Experimental Results and Analysis

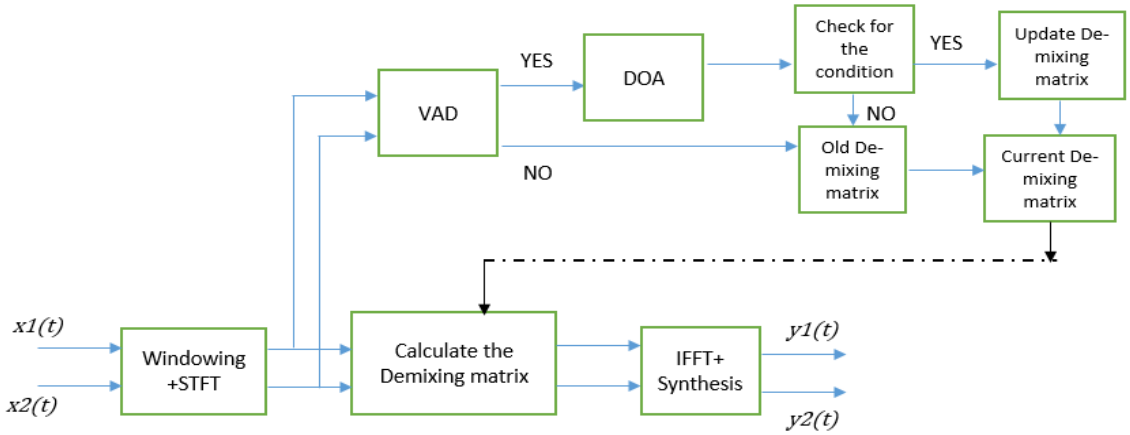


Figure 5.1. Block Diagram of the proposed method

5.5.1 Objective results

The data for assessing the performance of the proposed method using objective measures was generated using the ISM toolbox [67]. The clean speech was used from the TIMIT and HINT database. The clean speech source was placed at different angles to the microphone pair. The angles considered for the speech source direction were $[0^0, 45^0, 90^0, 135^0$ and $179^0]$. The noise is assumed to be diffused. The speech and the noise files were sampled at 16 kHz and the noisy

speech files were generated using ISM toolbox. The distance between the two microphones is 13 cm. The distance between the speech source and the microphones is 2.5 m. The noisy speech files with a frame size of 64ms is used for processing the IVA with 75% overlap.

The proposed method is evaluated using 4 different objective measures. For evaluating the source separation performance, Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and Signal to Artifact Ratio (SAR) [67] are used. PESQ is also used to measure the quality of the speech as it has high correlation with the subjective measures. Figure 5.2 shows the performance evaluation plots using the above-mentioned objective measures for speech mixed in Babble noise at SNR levels of -5 dB, 0 dB and 5 dB. The proposed method is compared with Noisy speech, Log MMSE, and BSS using the traditional IVA which is in non-real time. Single microphone Log-MMSE is included in only PESQ measures to give a perspective about the advantages of using 2 mics over single microphone. The BSS using the proposed setup outperforms Noisy and Log-MMSE in terms of all measures. As expected, the performance of the proposed method is on par with the traditional IVA in terms of PESQ. However, in terms of SIR, SDR and SAR, the traditional IVA gives better results than the proposed approach at the cost of high computational complexity, due to which it is impractical for real-time implementation. In Figures 5.3 and 5.4, the results for both Machinery and Traffic noise types are shown and they follow similar trends as Babble Noise. Objective measures reemphasize the fact that the IVA implemented using the proposed method achieves better speech quality and intelligibility without any speech distortion. However, there is a tradeoff in computational time and the accuracy of the proposed method in comparison to the traditional IVA approach.

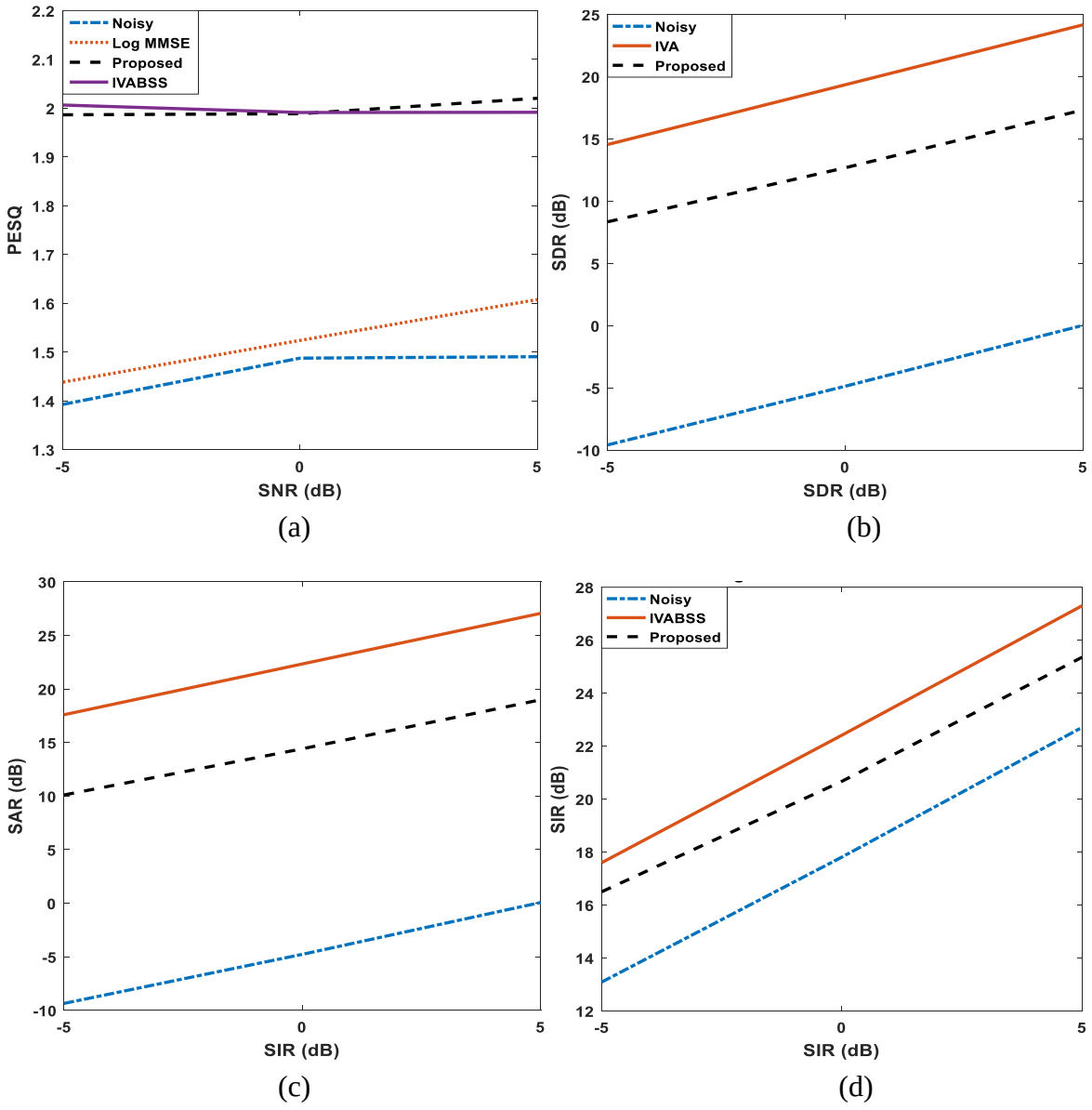


Figure 5.2. Performance evaluation for speech mixed with Machinery Noise using (a) PESQ, (b) SDR (c) SAR and (d) SIR

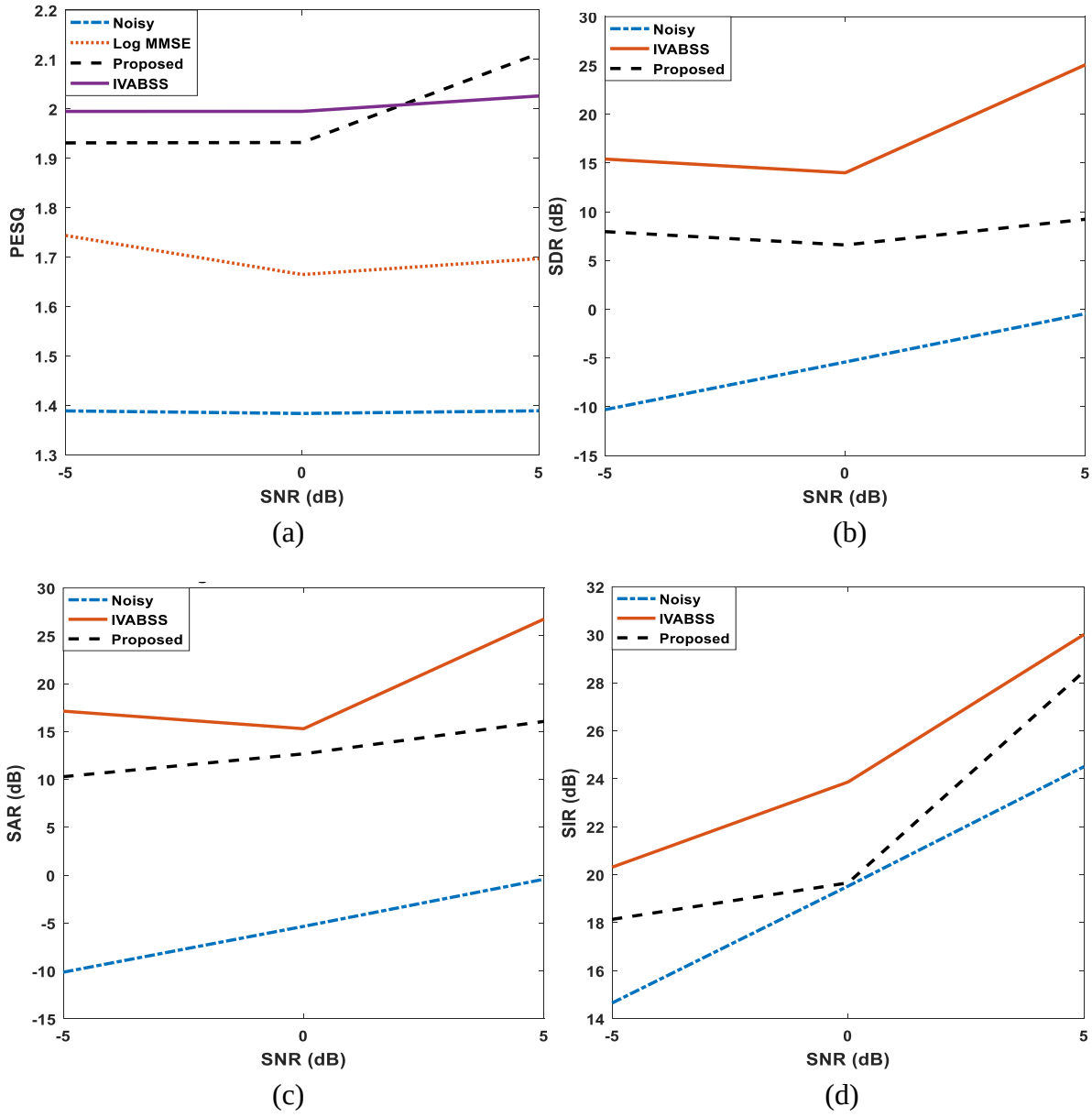


Figure 5.3. Performance evaluation for speech mixed with Babble Noise using (a) PESQ, (b) SDR (c) SAR and (d) SIR

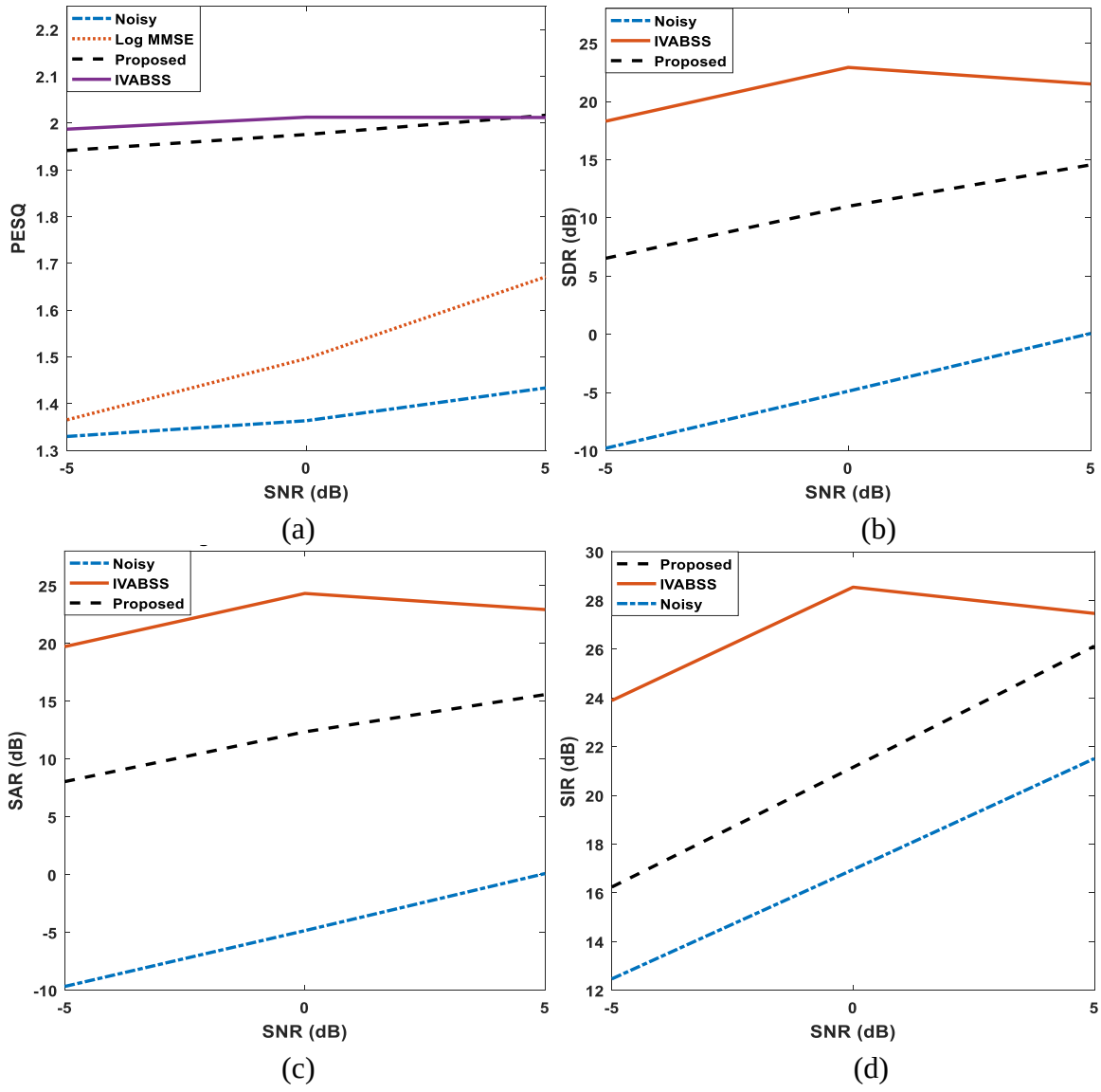


Figure 5.4. Performance evaluation for speech mixed with Traffic Noise using (a) PESQ, (b) SDR (c) SAR and (d) SIR

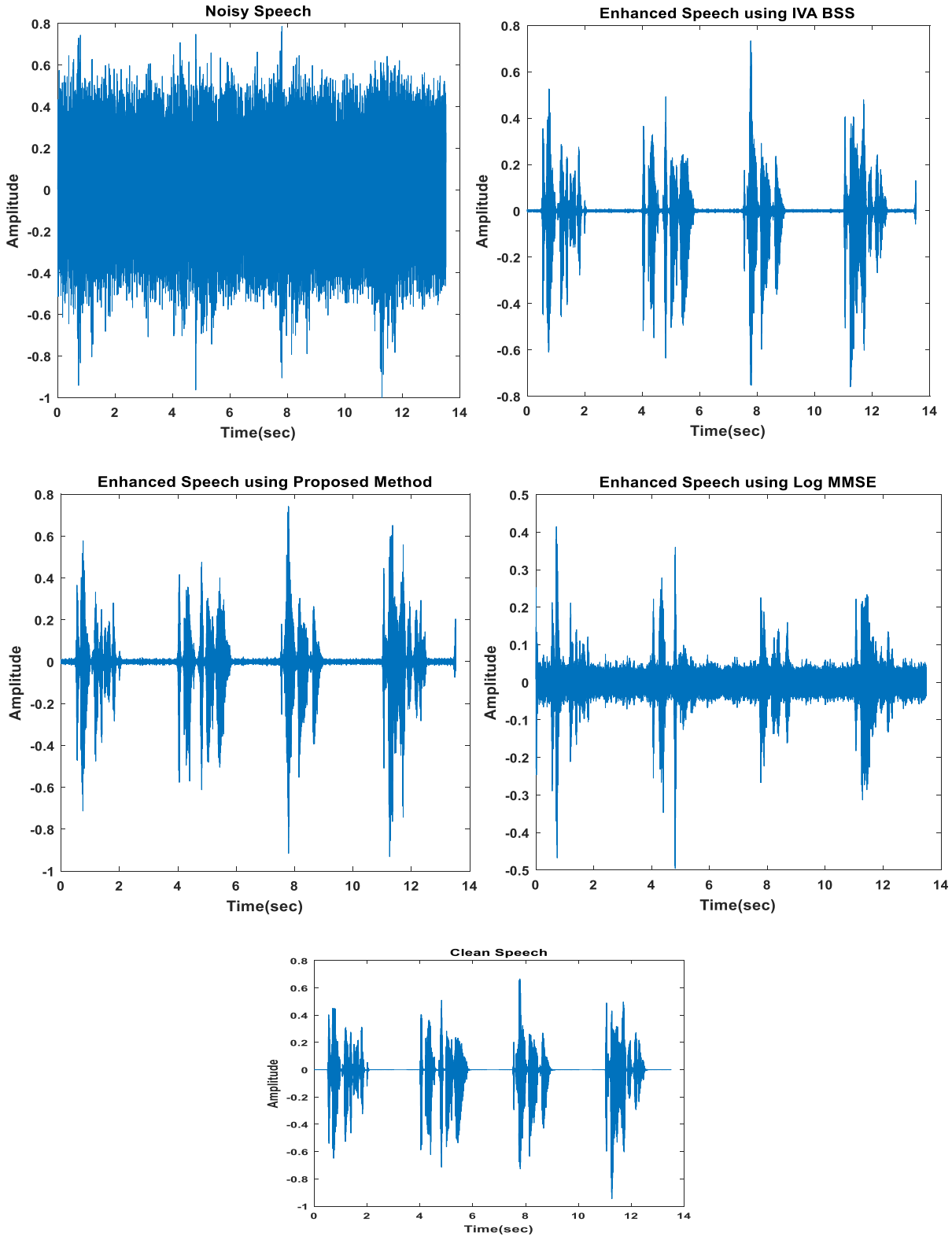


Figure 5.5. Comparison of the results using Time domain plots.

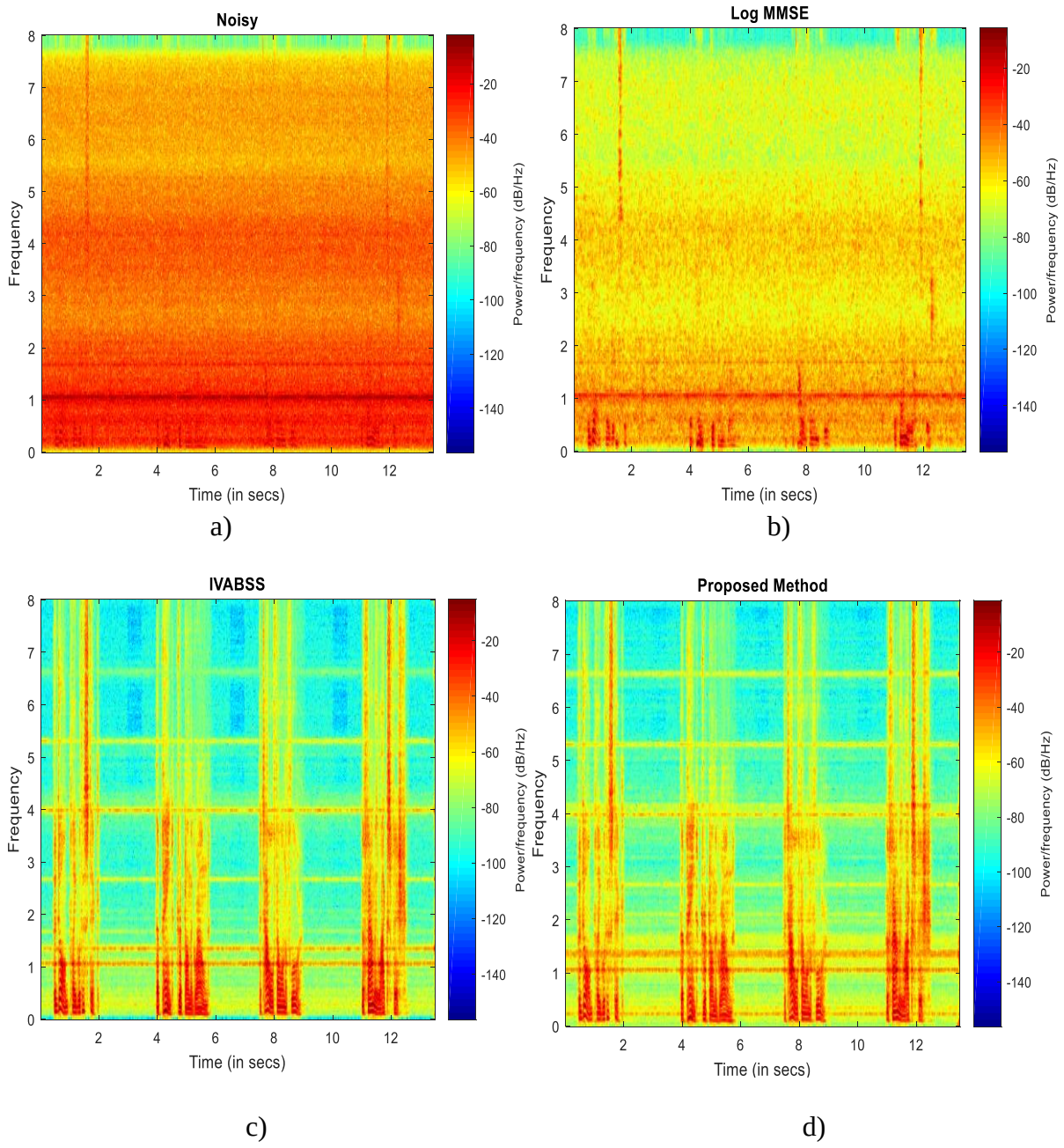


Figure 5.6. Comparison of the results using spectrograms. (a) Machinery noise mixed with clean speech at SNR=-5 dB, (b) Output of the Log MMSE, (c) Output of Non Real time IVA d) Output of the proposed IVA

5.5.2 Subjective Evaluation

Although we had 4 objective measures which provided useful evaluation results during the development phase, we carried out subjective tests to check the preference of the end user. We performed MOS tests [38] on 10 expert normal hearing subjects who were presented with noisy speech and enhanced speech using the proposed IVA method at SNR levels of -5 dB, 0 dB and 5 dB. For each audio file the subjects were instructed to score in the range of 1 to 5 with 5 being excellent speech quality and 1 being bad speech quality. They were given the flexibility to go back and change the score as well. The detailed description of scoring procedure is in [38]. Subjective test results in Figure 5.7 illustrate the effectiveness of the proposed method in reducing the separating the two sources, simultaneously preserving the quality and intelligibility of speech.

5.5.3 Computational Complexity of the proposed IVA

The computational complexity of the proposed method can be analyzed firstly by comparing it with the non-real time IVA. For instance, let us consider a noisy speech file of 30secs with a sampling rate of 16 kHz with a frame length of 65ms, so there will be approximately 462 frames if there is no overlap. The traditional IVA computes demixing matrix by waiting for 30secs to get the complete data and then compute the demixing matrix for the entire signal. On the other hand, the proposed method calculates the demixing matrix for first 100 frames and use this demixing matrix for the rest of the signal until the DOA of the source is changed. So, in the proposed method we have to wait for 6.5secs until the source separation is started. The traditional IVA takes an average of 30.65secs to separate the two sources when tested on a PC with Intel i7 octa core

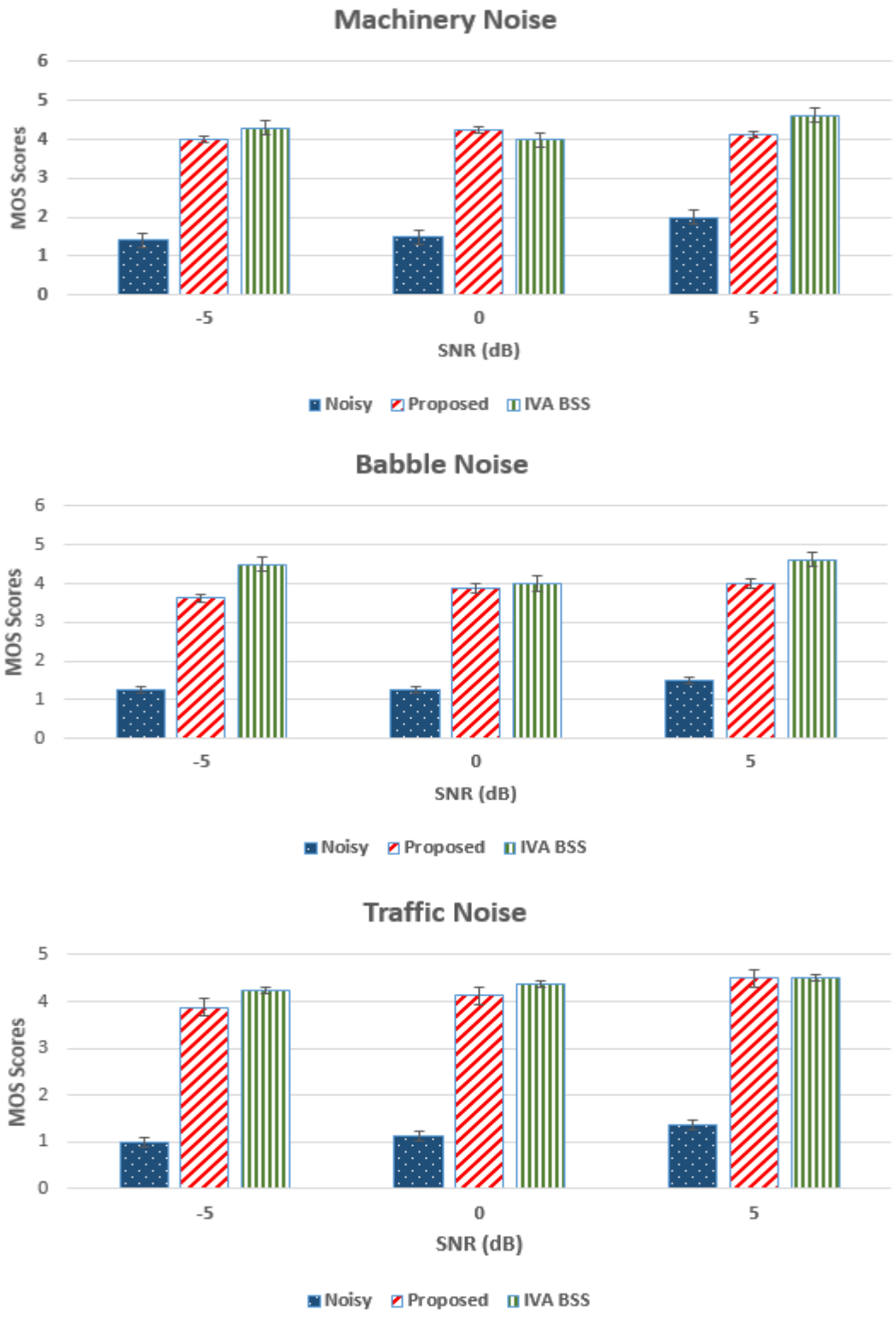


Figure 5.7. Subjective Test Results of IVA

processor with 3.6 GHz clock speed. So, if we have to use this algorithm in real time, we have to wait for 30secs to get the input data and an additional 30.65secs to separate the two sources and this is highly impractical. Whereas, the proposed method takes about 3.5secs to calculate the demixing matrix. The source separation process is then operated in frames. Each frame of 65ms takes about 22ms to process. As discussed in section 5.2, the processing time here does not exceed the frame size and thus can be implemented on the smartphone without inducing any delay or skipping of frames.

Secondly, the computational complexity of the proposed method can be analyzed by checking the number of times the demixing matrix or the separating matrix is updated for a particular length of noisy speech. This is compared to traditional online IVA. For example, let us consider a noisy speech file of 15secs with a sampling rate of 16 kHz in which the direction of the source changes every 3secs. Therefore, according to the proposed method in Figure 5.1, the changes in the source direction should be tracked 4 times. Even if a larger frames (say 100ms) is processed without any overlap, a total of 150 frames should be processed. The traditional online-based IVA computes demixing matrix 150 times. And this complexity escalates as the number of frames increase. On the other hand, the proposed method tracks the change in the DOA of the source and updates the demixing matrix only 4 times in the entire 15secs of the data. Hence, for this scenario, the proposed method is 37 times computationally efficient than the traditional online approach.

5.6 Chapter Outcomes

In this Chapter, a computationally efficient real-time dual channel IVA technique is developed which operates frame wise and uses few frames to separate the two sources. A neural network based DOA method [65] is used track a change in the source direction and update the demixing matrix, this provides stability to the proposed IVA and can be operated even when there is change in the impulse response between the microphone and speech source. With a little compromise in the accuracy of the results, the computational complexity of the traditional method can be greatly reduced and the proposed method can be implemented in real-time.

CHAPTER 6

CONCLUSION

In this thesis, single and dual microphone SE techniques are developed that are designed to run on a smartphone and this setup works as an assistive device to Hearing Aids.

In Chapter 3, a single microphone SE was proposed that makes use of formant frequency information to attain more noise suppression on certain acoustical bands without inducing any speech distortion or residual noise. Two parameters are introduced to control the amount of noise reduction over different bands. The proposed SE method makes use of the gain function of the new super Gaussian Joint Maximum *a Posteriori* (SGJMAP). The objective and subjective results show substantial improvement over traditional methods.

In Chapter 4, an MMSE extension of the formant frequency based SGJMAP SE is proposed. In this approach, a single parameter called scaling factor is provided on the smartphone Graphical User Interface (GUI) that can be controlled by user in real-time, which controls the gains over non formant regions. The objective and subjective results show significant improvements and proves the usability of the developed application in real-world noisy conditions.

In Chapter 5, a computationally efficient Blind Source Separation framework is introduced and analyzed for convolutive mixtures. A DOA based criterion is used to update the separating matrix and reduce the overall computational complexity of the BSS. The developed method is computationally fast and operates in real-time. The experimental results showcase superior performance in terms of source separation capability and speech quality and intelligibility.

REFERENCES

[1] D. L. Blackwell, J. W. Lucas, T. C. Clarke, “Summary health statistics for U.S. adults: National Health Interview Survey, 2012”, National Center for Health Statistics. Vital Health Stat. vol. 10, no. 260, 2014.

[2] Quick Statistics (n. d.) retrieved from http://www.nidcd.nih.gov/health/ statistics/pages/-quick.aspx.

[3] C. K. A. Reddy, Y. Hao and I. Panahi, "Two microphones spectral-coherence based speech enhancement for hearing aids using smartphone as an assistive device," 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, 2016, pp. 3670-3673.

[4] P. Loizou, “Comparison of Speech Enhancement Algorithms” in Speech Enhancement: Theory and Practice. 2nd Edition. Boca Raton, FL, USA: CRC, 2013, ch 12, pp. 598-599

[5] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean- square error shorttime spectral amplitude estimator,” IEEE Trans., Acoust., Speech and Signal Process., vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[6] Y. Hu and P. Loizou, “Subjective comparison of speech enhancement algorithms,” *in Proceedings of IEEE Inter. Conf. on Acoust., Speech and Signal Process., ICASSP 2006*, vol. 1, no. 6, pp. 153-156, April. 2006.

[7] Y. Ephraim and D.Malah, “Speech enhancement using a minimum meansquare error log-spectral amplitude estimator,” *IEEE Trans. Acoustics,Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[8] P. J. Wolfe and S. J. Godsill, “Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement*,” EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 10, pp. 1043–1051, 2003, special issue: Digital Audio for Multimedia CommunicationsT.

[9] P. V. Lotter , “Speech Enhancement by MAP Spectral Amplitude Estimation using a super-gaussian speech model,” EURASIP Journal on Applied Sig. Process, pp. 1110-1126, 2005.

[10] Y. Xu, J. Du, L-R. Dai, C-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Proc. Letters*, pp. 65-68, Nov 2013.

[11] F. Weninger, J. R. Hershey, J. L. Roux,B. Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” *IEEE Global Conf. on Signal and Inf Processing*, Dec 2014.

[12] Y. Rao, Y. Hao, I. M. S. Panahi and N. Kehtarnavaz, "Smartphone-based real-time speech enhancement for improving hearing aids speech perception," *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, 2016, pp. 5885-5888

[13] S. Makino, T-W. Lee, H. Sawada, “Blind Speech Separation,” *Springer Signals and Communication technology*, 2007.

[14] Karadagur Ananda Reddy, C., & The University of Texas at Dallas. Graduate Program in Electrical Engineering, degree granting institution. (2015). *Single and dual microphone speech enhancement techniques for hearing devices*.

[15] C. Karadagur Ananda Reddy, N. Shankar, G. Shreedhar Bhat, R. Charan and I. Panahi, "An Individualized Super-Gaussian Single Microphone Speech Enhancement for Hearing Aid Users With Smartphone as an Assistive Device," in *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1601-1605, Nov. 2017.

[16] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustic, Speech and Signal Process*, vol. 27, pp. 113-120, Apr 1979.

[17] J. Chen, J. Benesty, Y. Huang, and S. Doclo, “New insights into the noise reduction wiener filter,” *IEEE Trans. Speech Audio Process.*, vol. 14, Pp. 1218-1234, July 2006.

[18] S. Kay, “Fundamentals of Statistical Signal Processing: Estimation Theory,” *Upper Saddle River, NJ:Prentice Hall.*

[19] R. Gray, A. Buzo, A. Gray, and Matsuyama, Y., “Distortion measures for speech processing,” *IEEE Trans. Acoust. Speech Signal Process*

[20] Y. Hu, and P. Loizou, “A Comparative Intelligibility Study of Speech Enhancement Algorithms,” *IEEE Int. Conf. Acoustic, Speech, Signal Process. (ICASSP)*, Honolulu, HI, vol.4, Pp. 561-564, Apr 2007.

[21] Y. Ephraim and H. L. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Speech Audio Process.*, 3(4); 251-266, July 1995.

[22] Y. Hu and P. C. Loizou, “A generalized subspace approach for enhancing speech corrupted by colored noise,” *IEEE Trans. Speech Audio Process.*, 11(4):334-341, July 2003.

[23] S. So, K. K. Wojcicki and K. K. Paliwal, “Single-channel speech enhancement using Kalman filtering in the modulation domain,” in *11th Annual Conf. of the Int. Speech Communication Association,* 2010.

[24] S. S. Priyanka, "A review on adaptive beamforming techniques for speech enhancement," *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Vellore, 2017, pp. 1-6.

[25] Siddappaji and K. L. Sudha, "Performance analysis of New Time Varying LMS (NTVLMS) adaptive filtering algorithm in noise cancellation system for speech enhancement," *2014 4th World Congress on Information and Communication Technologies (WICT 2014)*, Bandar Hilir, 2014, pp. 224-228.

[26] D. Fischer and T. Gerkmann, "Single-microphone speech enhancement using MVDR filtering and Wiener post-filtering," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 201-205.

[27] J. P Dmochowski, J. Benesty, “Microphone arrays: fundamental concepts,” *Speech Processing in Modern Communication- Challenges and Perspectives*, ed. by I. Cohen, J. Benesty, S. Gannot (Springer, Berlin, 2010), Chapter 8: pp. 199-223

[28] G. W. Elko, J. Meyer, “Microphone arrays,” *Springer Handbook of Speech Processing,* ed. by J. Benesty, M. M. Sondhi, Y. Huang, (Springer, Berlin, 2008) Chapter 48: pp. 1021-1041.

[29] N. Yousefian and P. C. Loizou, "A Dual-Microphone Speech Enhancement Algorithm Based on the Coherence Function," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 599-609, Feb. 2012. doi: 10.1109/TASL.2011.2162406

[30] Apple. Mar. 2017. [Online]. Available: https://support.apple.com/en-us/HT203990

[31] J. Lim, “Evaluation of a correlation subtraction method for enhancing speech degraded by additive noise,” IEEE Trans. Acoust., Speech.Signal Process., vol. ASSP-37, pp. 471-472, Dec. 1987.

[32] L Ning, P. C. Loizou, “Factors influencing intelligibility of ideal binary- masked speech: implications for noise reduction,” J. Acoust. Soc. Amer., vol. 123(3), pp. 1673-1682, 2008.

[33] P. C. Loizou, “Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions,” IEEE Trans. Speech Audio Process., Vol.19, pp. 47-56, 2011.

[34] “Modeling the perception of concurrent vowels: Role of formant transitions,” J. Acoust. Soc. Amer., pt. 1, vol. 100, no. 2, pp. 1141–1152, Aug. 1996.

[35] R. N. Ohde, “The development of the perception of cues to the [m] [n]distinction in CV syllables,” J. Acoust. Soc. Amer., pt. 1, vol. 96, no. 2, pp. 675–686, Aug. 1994.

[36] A. K. Nábelek, Z. Czyzewski, and H. Crowley, ˇ “Cues for perception of the diphthong /ai/ in either noise or reverberation. Part I. Duration of the transition,” J. Acoust. Soc. Amer., pt. 1, vol. 95, no. 5, pp. 2681–2693, May 1994.

[37] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, “The role of consonant-vowel transitions in the perception of the stop and nasal consonants,” Psychol. Monographs, vol. 68, pp. 1–13, 1954

[38] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 435-444, March 2006.

[39] McLoughlin, Ian Vince, and R. J. Chance. "LSP-based speech modification for intelligibility enhancement." *Digital Signal Processing Proceedings, 1997. DSP 97., 1997 13th International Conference on*. Vol. 2. IEEE, 1997.

[40] Zorila, Tudor-Catalin, Varvara Kandia, and Yannis Stylianou. "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression." *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[41] Indiana.University.Mar.2018.Available: http://www.physics.indiana.edu/~courses/p109/P109fa08/11.pdf

[42] R. L. Miller, J. R. Schilling, K. R. Franck, and E. D. Young, “Effects of acoustic trauma on the representation of the vowel ="=in cat auditory nerve fibers,” J. Acoust. Soc. Amer., vol. 101, no. 6, pp. 3602–3616, Jun. 1997.

[43] M. B. Sachs, I. C. Bruce, R. L. Miller, and E. D. Young, “Biological basis of hearing-aid design,” Ann. Biomed. Eng., vol. 30, no. 2, pp. 157–168, Feb. 2002.

[44] R. Martin,"Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, vol. 1, pp. 253–256, Orlando, Fla, USA, May 2002.

[45] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC'03)*, pp. 87–90, Kyoto, Japan, September 2003.

[46] Apple. Mar. 2017. [Online]. Available: http://www.starkey.com/blog/2014/04/7-halo-features-that-will-enhanseevery-listening-experience

[47] A. W. Rix, J. G. Beerends, M. P Hollier, A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 2, pp. 749-752., May 2001.

[48] P. C. Loizou , 'Speech Enhancement Theory and Practice'

[49] ITU-T Rec. P.830, "Subjective performance assessment of telephoneband and wideband digital codecs," 1996.

[50] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*., vol. 6, no. 1, pp. 1–3, 1999.

[51] Apple. March. 2017. [Online]. Available: https://developer.apple.com/xcode/

[52] Apple.Mar.2017.[Online].Avaliable: https://developer.apple.com/library/content/documentation/MusicAudio/Conceptual/CoreAudioOverview/WhatisCoreAudio/WhatisCoreAudio.html

[53] S Makino, H Sawada, R Mukai, S Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," IEICE Trans. Fund. Electron. E88-A(7), pp. 1640-1655, 2005.

[54] MS Pederson, J Larsen, U Kjems, LC Parra, "A survey of convolutive blind source separation methods," Springer Handbook on Speech Processing and Speech Communication, Springer, Newyork, 2007.

[55] P Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing 22(1-3), pp. 21-34, 1998.

[56] V Calhoun, T Adali, “Complex infomax: convergence and approximation of infomax with complex nonlinearities,” J. VLSI Signal Process. 44, pp. 173-190, 2006

[57] E Bingham, A Hyvarinen, “A fast fixed-point algorithm for independent component analysis of complex valued signals,” Int. J. Neural Syst. 10(2), pp.1-8, 2000.

[58] A Hiroe, “Solution of permutation problem in frequency domain ICA, using multivariate probability density functions,” Int. Conf. on Independent Component Analysis and Blind Source Separation, vol. 3889, pp. 601-608, Charleston, SC, USA, 2006.

[59] Na et al, “Independent vector analysis using subband and subspace nonlinearity,” EURASIP Journal on Advances in Signal Processing 2013 2013:74.

[60] T. Kim, HT Attias, S-Y Lee, T-W Lee, “Blind source separation exploiting higher-order frequency dependencies,” IEEE Trans. Audio Speech, 15(1), pp. 70-79, 2007.

[61] Y. Liang, J. Harris, Gaojie Chen, S. M. Naqvi, C. Jutten and J. Chambers, "Auxiliary function based iva using a source prior exploiting fourth order relationships," *21st European Signal Processing Conference (EUSIPCO 2013)*, Marrakech, 2013, pp. 1-5.

[62] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 7, pp. 1431-1438, 2010.

[63] A. H. Khan, M. Taseska, and E. A. P. Habets, *A Geometrically Constrained Independent Vector Analysis Algorithm for Online Source Extraction.* Cham: Springer International Publishing, 2015, pp. 396-403

[64] F. Nesta and Z. Koldovský, "Supervised independent vector analysis through pilot dependent components," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 536-540.

[65] C. K. A Reddy User Customizable Real-Time Single and Dual microphone Speech Enhancement and Blind Source Separation for Smartphone Hearing Aid Applications

[66] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux," in *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197-200, March 2013.

[67] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462-1469, July 2006.

BIOGRAPHICAL SKETCH

Gautam Shreedhar Bhat is from Bangalore, India. He completed his Bachelor of Engineering in Electronics and Communication at Nitte Meenakshi Institute of Technology Bangalore May 2016. He worked as a Research intern at Indian Space Research Organization (ISRO), Bangalore, from Dec 2015 to May2016 where he worked on image processing. He is pursuing his Master's thesis program in Electrical Engineering at UT Dallas in from August 2016. He started to work in Statistical Signal Processing Research Laboratory (SSPRL) from September 2016. His research interests include, developing single and dual microphone Speech Enhancement techniques for hearing study, Blind Source Separation and Beamforming algorithms for Smartphone- Hearing Aid applications. He is also well versed in developing Machine Learning based algorithms for audio and speech applications.

CURRICULUM VITAE

Gautam Shredhar Bhat

Email: gautam.bhat4136@gmail.com

EDUCATION

Master of Science (Thesis), Electrical Engineering CGPA: 3.62
The University of Texas at Dallas (UTD), TX, USA May 2018
Bachelor of Engineering, Electronics & Communication CGPA: 8.92
Nitte Meenakshi Institute of Technology, Bangalore, KA, India May 2016

SKILLS

Programming Languages: MATLAB, Simulink, and C/C++. Familiar with Python
Tools: Audacity, Visual Studio, Eclipse IDE. Familiar with Android Studio and Core Audio
Courses: Random Processes, Digital Signal Processing I, Digital Signal Processing II, Digital Communication Systems, Special topics in biomedical applications of EE, Speech Perception Laboratory, Detection and Estimation Theory, **Machine Learning** by coursera, **Deep Learning** A-Z by Udemy Academy.

WORK EXPERIENCE

Student Research Worker at **Statistical Signal Processing Research Lab [September 2016-Present]**
Developed, implemented and tested algorithms for single channel and multichannel Speech Enhancement, Independent Vector analysis, Beamforming and Audio Compression. Coded in Matlab and C and implemented on PC and Smartphones.
Project Intern at **Indian Space Research Organization (ISRO) [Jan 2016-May2016]**
Implementation of image Compression using Lifting based 2-D Discrete Wavelet Transform algorithm.
Student Volunteer at DSP lab, Nitte Meenakshi Institute of Technology [Jun 2015-Dec 2015]
Developed audio signal processing pipeline using multi microphone array involving Speech recognition and Speech enhancement.

PROJECTS

A. Research Projects

1. **'Formant Based Single Microphone Speech Enhancement to improve speech Intelligibility' [June 2017- Present]**
   - Developed a Speech Enhancement algorithm to improve the speech intelligibility for Hearing aid users.
   - Analyzed and incorporated Pitch and Formants information to improve performance of Statistical Model based Speech enhancement techniques.
   - Implemented on Smartphone (iOS and Android)

2. **'Dual Channel Speech Enhancement using Blind Source Separation and Beamforming' [Dec 2017- present]**
    - Developed dual microphone speech enhancement algorithms using Independent vector analysis and Beamforming
    - Implemented on Smartphone and PC
3. ***'Neural network based Direction of Arrival to update Source Separation'*** **[Feb 2018-present]**
    - Implemented a GCC based DOA using simple Feed Forward Neural network.
    - Extracted Cross correlation features, trained the model, tested it using real data.
4. **'Super Gaussian single microphone Speech Enhancement for hearing aid users' [Jan 2017- May 2017]**
    - Developed Speech Enhancement algorithm using Super Gaussian Speech Modeling**.**
    - Implemented it on Smartphone (iOS) which acts as assistive device for hearing aids.
5. **Smartphone based Multi-channel Dynamic Range Audio Compression for hearing Aid Users [Sep 2016-Dec 2016]**
    - Developed Multi channel compression algorithm for Hearing impaired Smartphone users
    - Coded in Matlab and C. Implemented on PC and Smartphone.

B. Academic Projects

1. **System Identification and a comparative study using Least Means Square(LMS), Normalized Least Means Square(NLMS) and Wiener Optimum Filter for Real Data**
    - Implemented LMS and NLMS adaptive filtering algorithms to identify the unknown system.
    - Compared these methods with Winer Optimum method.
2. **Simple Voice Activity detection using Deep Neural Networks.**
    - Implemented a Voice Activity Detector using Feed Forward Neural network.
    - Extracted Spectral Features, trained the model, tested it using real data.
3. **ECG Signal Classification using SVM Classifier and Perceptron Method and Real Time Implementation on Smartphone.**
    - Implemented an SVM Classifier on PC and Smartphone to classify Abnormal and Normal Heartbeats.
    - Extracted features, trained the classifier and coded in both C and Matlab

PUBLICATIONS

1. **G. Bhat**, N. Shankar, C. K. A. Reddy, I. Panahi, "Formant Frequency-based Speech Enhancement technique to improve intelligibility for hearing aid users with smartphone as an assistive device," *IEEE-NIH 2017 Special Topics Conference on Healthcare Innovations and Point-of-Care Technologies,* Bethesda, MD, Nov 2017.
2. **G. S. Bhat**, C.K.A. Reddy, N. Shankar and I. M. S. Panahi, "Smartphone based real time Super Gaussian Speech enhancement to improve intelligibility for hearing aid users using formant information" – *40th International conference of the IEEE Engineering in medical and Biology* (under review)

3. Yiya Hao, R.Charan, **G.S.Bhat** and I. Panahi, “Robust Real-time Sound Pressure Level Stabilizer for Multi-Channel Hearing Aids Compression for Dynamically Changing Acoustic Environment” – *Signals, Systems and Computers, 2017 51st Asilomar Conference* (Accepted)
4. N. Shankar, **G. Shreedhar Bhat**, C. Karadagur Ananda Reddy, and I. Panahi, “ Noise dependent Super Gaussian-Coherence based dual microphone Speech Enhancement for hearing aid application using smartphone” - *175th Meeting of the Acoustical Society of America* (Accepted)
5. Karadagur Ananda Reddy, N. Shankar, **G. Shreedhar Bhat**, R. Charan and I. Panahi, "An Individualized Super-Gaussian Single Microphone Speech Enhancement for Hearing Aid Users with Smartphone as an Assistive Device," in *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1601-1605, Nov. 2017.

6. N. Shankar, Abduallah Kucuk, C. Karadagur Ananda Reddy, **G. Shreedhar Bhat** and I. Panahi, “Influence of MVDR Beamformer on speech enhancement based Smartphone applications” – *40th International conference of the IEEE Engineering in medical and Biology* (under review)

VISA STATUS: F1/ Eligible to work in U.S.