DUAL MICROPHONE SPEECH ENHANCEMENT ALGORITHMS ON ANDROID-BASED DEVICES FOR HEARING STUDY

by

Nikhil Shankar

APPROVED BY SUPERVISORY COMMITTEE:

Dr. Issa M. S. Panahi, Chair

Dr. Carlos Busso

Dr. Mehrdad Nourani

Dr. P. K. Rajasekaran

To my parents, mentors and friends,

who supported me and believed in me

DUAL MICROPHONE SPEECH ENHANCEMENT ALGORITHMS ON ANDROID-BASED DEVICES FOR HEARING STUDY

by

NIKHIL SHANKAR, 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

August 2018

ACKNOWLEDGMENTS

It was always a dream to do my Master's in the United States of America and I would like to take this opportunity to thank all the people who made this dream come true. It is my pleasure to show appreciation towards the people who made this journey at UT Dallas, especially those in the Statistical Signal Processing Research Lab (SSPRL), a memorable one.

Firstly, I would like to express my gratitude to my advisor, Dr. Issa M. S. Panahi for believing in me and providing me with the opportunity to work in the SSPRL. This Thesis would not have been possible without his support for my study and his knowledge. His innovation, motivation and guidance helped me at all times during my research.

I thank the rest of my thesis committee: Dr. P. K. Rajasekaran, Dr. Mehrdad Nourani and Dr. Carlos Busso for your input towards the betterment of my research.

I would want to thank Dr. Chandan K A Reddy and Dr. Anshuman Ganguly for clarifying my doubts and helping me grow at SSPRL. I also thank all my friends who were and are at SSPRL: Yiya Hao, Gautam S Bhat, Ram Charan, Abduallah Kucuck, Parth Mishra, Ziyan Zou, Serkan Tokgoz, and Holden Hernandez for providing me with helpful solutions and for working as a team. Thank you for maintaining a happy work environment for the past two years. Also, I would like to recognize Dr. Chandan K A Reddy and Gautam S Bhat for the collaborative efforts and contributions in the development of Speech Enhancement Algorithms.

I would like to thank Inchara Raveendra for her support during the tough times and keeping me motivated and happy throughout my journey.

Finally, I would like to thank my parents, and my relatives for their unconditional love and being there for me throughout my life during both my happiest and toughest times. I am what I am because of my parents.

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 5R01DC015430-02. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH. The author is with the Statistical Signal Processing Research Laboratory (SSPRL), Department of Electrical and Computer Engineering, The University of Texas at Dallas.

June 2018

DUAL MICROPHONE SPEECH ENHANCEMENT ALGORITHMS ON ANDROID-BASED DEVICES FOR HEARING STUDY

Nikhil Shankar, MS
The University of Texas at Dallas, 2018

Supervising Professor: Dr. Issa M. S. Panahi

Speech Enhancement (SE) is a key module in the Hearing Aid (HA) signal processing pipeline and improves the listening comfort. Over the last few decades, researchers have developed many single and dual-microphone SE techniques. In this thesis, two novel dual-channel SE techniques have been proposed and are implemented on Android-based smartphones as an assistive device for HA. In the first algorithm, the coherence between speech and noise signals is used to obtain an SE gain function, in combination with a Super-Gaussian Joint Maximum a Posteriori (SGJMAP) single microphone SE gain function. The second technique uses the Minimum Variance Distortionless Response (MVDR) as a Signal to Noise Ratio (SNR) booster for the SE method. The considered SE gain is based on the Log Spectral Minimum Mean Square Error Amplitude Estimator (Log-MMSE) to improve the speech quality in the presence of different background noise. Objective evaluation and subjective results of the developed methods show significant improvements in speech quality and intelligibility in comparison with existing SE methods.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Speech is eventually a communication tool; it gives us the possibility to interact with the world. Speech communication in the presence of background noise degrades the quality and intelligibility of speech. Quality is a subjective measure which reflects on listener's preference and Intelligibility is an objective measure which predicts the number of words correctly identified by the listeners [1]. Audio processing can improve communication by means of enhancing speech. Speech Enhancement (SE) plays a vital role in hands-free communication devices such as Hearing Aid Devices (HADs), cellular phones, teleconferences and Automatic information systems. In all these applications, the goal of SE varies, such as to increase the quality and intelligibility of the speech signal, and to improve the performance of speech communication. We can process the speech with several microphones (two or more) or only by one. With the progress in technology and the increasing demand for permanent reachability and connectivity, the exchange of information via speech is feasible nowadays from anywhere at any time. The use of smartphones has become an inherent part of daily life.

Digital signal processing plays a very important role in ensuring a high transmission quality on smartphones. The increasing computational performance of the technical devices allows the realization of more and more sophisticated and complex algorithms in mobile phones. Depending on the Signal-to-Noise-Ratio (SNR), interferences make a conversation uncomfortable for the user. In order to manage with such acoustic environments, SE algorithms are implemented in many speech communication systems. SE is used to attenuate the intensity of noise while preserving the

quality and intelligibility of the speech. Thus, SE or noise reduction is an essential feature in HADs and other wearable technology.

1.1 Motivation

Records by National Institute on Deafness and Other Communication Disorders (NIDCD) indicate that nearly 15% of adults (37million) aged 18 and over report some kind of hearing loss in the United States. Amongst the entire world population, 360 million people suffer from hearing loss. Researchers have developed numerous solutions for hearing impaired in the form of HADs and other hearing assistive devices. Performance of HADs and Cochlear Implants (CI) degrade in the presence of background noise, thus reducing the quality and intelligibility of speech [2]. SE plays a vital role in suppressing the noise in various stationary and non-stationary environments while preserving the speech features.

Figure 1.1 shows the simplified block diagram of the speech processing pipeline used in the HADs which typically has two paths: A Critical path and an Auxiliary path. The algorithms on the Critical path ensure continuous operation i.e. real time, while the algorithms present in the Auxiliary path

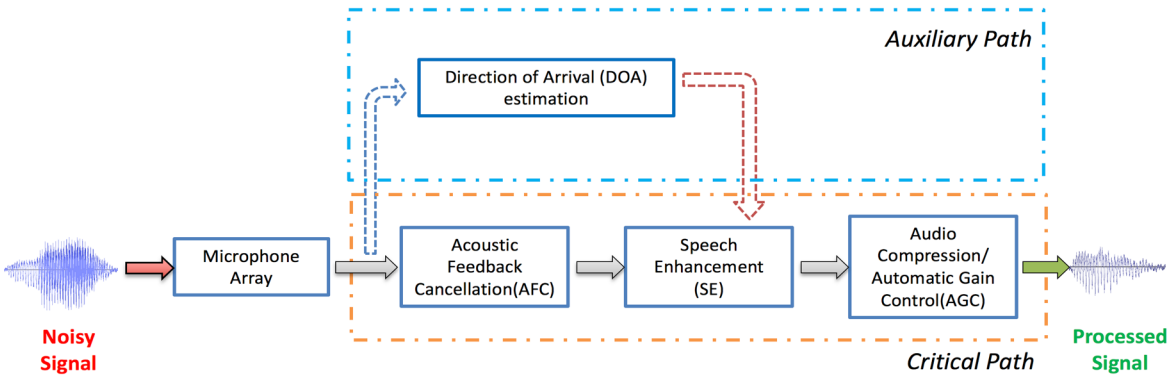


Figure 1.1 Block Diagram of HAD signal processing pipeline

improve/ support the performance of the Critical path. The noisy speech is captured by the microphone array and then can be used as an input to both the paths. The critical path has Acoustic Feedback Cancellation block, SE block to reduce the background noise, and an Audio Compression and Automatic Gain Control block. Direction of Arrival (DOA) Estimation block in the Auxiliary path finds the directionality of the source. The final output will be the processed signal, transmitted to the HADs.

1.2 Solution

The existing HADs lacks the computational power, due to the design constraints and to handle obligatory signal processing algorithms. Lately, HADs manufacturers are using a pen or a necklace as an external microphone to capture speech and transmit the signal and data by wire or wirelessly to HADs [3, 4]. An alternative solution is the use of smartphone which can capture the noisy speech data using the two microphones, perform complex computations using the SE algorithm and transmit the enhanced speech to the HADs through a wire or wirelessly.

1.3 Single and Dual Channel Speech Enhancement

About 10-20 percent of the population suffers from hearing impairment, that is caused by damage to inner ear hair cells in the process of aging or exposure to loud noise. The exposure to loud noise is mainly in the environments such as traffic from vehicles, machines in an industry, by listening to loud music, and engines. When ears are exposed to these types of environments, may lead to temporary or permanent hearing loss. HADs amplifies the received speech signal without considering the SNR level. In a noisy environment, noise is also amplified along with speech signal as hearing-impaired people are incapable of distinguishing the noise and speech signals.

As shown in figure 1.1, SE is the main block which concentrates on improving the Speech quality and intelligibility and reduces the background noise. SE is being used for more than 30 years. It depends on the type of noise, number of microphones and the source speech signal. Widely used methods in Single channel SE are spectral subtraction, Wiener filtering and other statistical-model based algorithms [5]. In this thesis, we operate in the frequency domain and Short-Time Fourier Transform (STFT) is used to transform the time domain signal to the frequency domain. In the statistical-model based methods, a statistical model for the speech Fourier transform coefficients are considered. Finally, the spectral magnitude is multiplied with a nonlinear gain function. The obtained gain function is used to enhance the speech by optimizing a cost function. Ephraim and Malah proposed a new SE method known as the minimum mean square error (MMSE) estimator [6] which suppresses non-stationary background noise and improves the perceptual quality of the speech signal [7-8]. In this statistical-model based SE techniques, the gain function is a function of a priori SNR and a Posteriori SNR. Suppression rules that are derived under a Gaussian model are interpreted as spectral estimators in a Bayesian statistical framework to obtain a computationally efficient alternative for the MMSE method. In this new method, speech estimation is done by applying the joint maximum a posteriori (JMAP) estimation rule [9]. In [10], a super-Gaussian extension of the JMAP (SGJMAP) is proposed which is shown to outperform algorithms proposed in [6, 8]. Spectral amplitude estimators with a super-Gaussian speech model are considered. SE based on deep neural networks (DNN) [11] requires rigorous training data, but yield supreme noise suppression. The preservation of Spectro-temporal characteristics of speech, the quality, and natural attributes remains a prime challenge using such methods.

1.4 Thesis Objectives and Outline

The conventional single microphone SE algorithms have limitations when compared to dual channel enhancement techniques, i.e. no information about spatial mixing process, highly dependent on the noise power, dependent on the accuracy of Voice Activity Detector (VAD), and some of the methods introduces annoying musical noise in a highly non-stationary noise environment. In certain conditions, there are no measures taken to counteract the effects of reverberation and room impulse response. Thus, dual microphone methodology was considered which provide a favorable solution for the above-mentioned disadvantages. Two dual-microphone SE methods are proposed in this thesis:

I. Super Gaussian-Coherence based Dual Microphone Speech Enhancement

The proposed SE method is based on super-Gaussian joint maximum a Posteriori (SGJMAP) estimator [10, 12]. We use the coherence between speech and noise signals to obtain an SE gain in combination with the above SGJMAP estimator. We introduce a parameter called 'weighting' factor in the proposed SE gain function that can be varied by the user to control the weighted combination, which in turn controls the amount of noise suppression and speech distortion. The algorithm is implemented on a smartphone that works as an assistive device to hearing aids.

II. Influence of MVDR beamformer on Log spectral amplitude estimator

The developed algorithm is a combination of MVDR beamformer and Minimum mean square error Log spectral amplitude estimator [8] SE gain function, for suppressing noise and extracting the desired speech. The beamformer is used as an SNR booster and the proposed method works as a real-time application on Android-based smartphones.

The outline of the thesis is as follows: Chapter 2 provides a literature review of single-channel and multi-channel SE algorithms. Chapter 3 introduces Super Gaussian-Coherence based dual microphone SE method. Its real-time implementation is also discussed along with objective measures and subjective tests. Chapter 4 discusses the influence of MVDR beamformer on Real-time Log-Spectral Amplitude Estimator SE technique. Implementation of smartphone and the performance evaluations are discussed. Chapter 5 gives a conclusion on the proposed SE techniques.

CHAPTER 2

LITERATURE REVIEW ON SPEECH ENHANCEMENT

2.1 Overview of Single Channel SE Techniques

The SE algorithms are developed based on an additive noise signal model and by taking STFT of that signal. Figure 2.1 shows the overview of conventional single microphone SE technique. The enhanced speech spectrum is given by,

$$\hat{S}_k = G_k \, Y_k \tag{2.1}$$

The gain function G_k dictates the performance of the SE algorithm.

Boll [13] presented a single channel SE (Spectral Subtraction) to enhance the desired speech degraded by white noise. Spectral Subtraction algorithm measures the signal present during non-speech activity and uses it as a noise estimate. Then the noise spectrum is subtracted from the noisy speech spectrum, to obtain a clean speech spectrum estimate. This algorithm introduces an annoying sound called "musical noise" for listeners. Some subspace algorithms [14] are based on the tradeoff between noise reduction and speech distortion. This method is based on the decomposition of the noisy speech into two subspaces: signal plus noise subspace and the noise subspace. Wiener filtering approach [15, 16] can be fixed or adaptive in nature. In case of fixed filters, the design is based on prior knowledge of signal and noise characteristics. Adaptive filters require very less a priori knowledge and can adjust its parameters automatically. The work of Wiener filters can be extended to the work of Kalman [17] and Bucy. These approaches cannot remove the background noise completely.

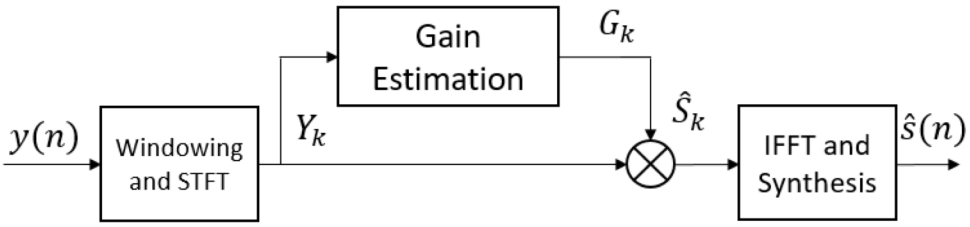


Figure 2.1 Conventional single microphone Speech Enhancement

In Statistical model-based methods, the weighting gain function depends on the cost function like Minimum Mean Square Error (MMSE), Maximum A Posteriori (MAP) or Maximum Likelihood (ML), and finally the statistical characteristics of the speech and the noise signal. Ephraim and Malah [6] spectral amplitude estimator has been very popular and is derived based on modeling speech and noise spectral components as statistically independent Gaussian random variables. The performance of this estimator is further improved by minimizing the MSE of the log-spectra [8]. Once the probability density function (pdf) is known, estimation depending on the assumptions of parameters can be carried out. If the parameter is assumed to be deterministic, it is known as classical estimation, i.e. Maximum-Likelihood (ML) approach [18]. Next, if the unknown parameter is a random variable with its own pdf, it is termed as a Bayesian estimator. This estimator allows incorporation of prior knowledge about the parameter by assigning a prior pdf. These widely used SE techniques give the enhanced speech of higher speech quality but compromises on the intelligibility. It is difficult to consider the effects of reverberation and time delays using a single microphone, as there is no spatial information. Hence in multi-channel SE algorithms, these limitations can be avoided.

2.2 Overview of Multi-Channel SE Techniques

In the multi-channel SE algorithms, the spatial information of the speech source and the noise source can be obtained. Therefore, a good estimate of the Direction Of Arrival (DOA) can be used to steer a beam towards the signal source and a null towards the noise source [19]. Beamforming techniques assume the above concept, where omnidirectional microphones are considered. Conventionally beamforming was used for narrowband applications such as radar. Beamforming algorithms can be classified into data-independent and data-dependent beamformers. Data-independent techniques use fixed parameters, whereas data-dependent beamformers update the parameters depending on the input signal. The simplest fixed beamformer is the delay and sum beamformer [20], where delayed versions of the microphones are equally combined at the output. The first adaptive beamformer used was the constant directivity beamformer [21], which obtains an invariable response in a wide frequency band. Then there are superdirective beamformers [22] for closely spaced endfire arrays under diffused noisy conditions. Traditionally used beamformers are the minimum variance distortionless response (MVDR) beamformer [23], which involves the knowledge of noise covariance matrix. We know that adaptive filter methods like LMS, NLMS [24] can be used to estimate the clean speech spectrum. Though these methods do not provide significant noise suppression, they can be used as SNR boosters. Coherence-based method deals with coherent noise suppression, where the speech from two microphones is correlated, while the noise is uncorrelated with speech [25]. Multi-channel Blind Source Separation (BSS) is based on the information of mixed signals [26]. Independent Component Analysis [27] and Independent Vector Analysis are commonly used BSS techniques for linear mixtures and convolutive mixtures respectively.

CHAPTER 3

SUPER GAUSSIAN-COHERENCE BASED DUAL MICROPHONE SPEECH ENHANCEMENT

3.1 Introduction

Amongst the entire world population, we know that 360 million people suffer from hearing loss. Over the past decade, researchers have developed many feasible solutions for hearing impaired in the form of Hearing Aid Devices (HADs) and Cochlear Implants (CI). However, the performance of the HADs degrade in the presence of different types of background noise and lacks the computational power, due to the design constraints and to handle obligatory signal processing algorithms [2-3]. Lately, HADs manufacturers are using a pen or a necklace as an external microphone to capture speech and transmit the signal and data by wire or wirelessly to HADs [28]. The expense of these existing auxiliary devices poses as a limitation. An alternative solution is the use of smartphone which can capture the noisy speech data using the two microphones, perform complex computations using the SE algorithm and transmit the enhanced speech to the HADs. There are many existing HADs applications which enhance the overall quality and intelligibility of the speech perceived by hearing impaired. Most of these applications use the single microphone of the smartphone. Recent progress includes microphone array based SE techniques for better noise suppression. But, as the number of microphones increases, so does the cost and computational power. A dual microphone methodology was considered which provide a favorable solution for improving speech quality. In this work, a two-microphone SE method is presented that can be implemented on a smartphone as an application with a user interface.

Existing methods like SGJMAP single microphone SE [10] introduce musical noise due to half-wave rectification problem [29]. A solution to this is the estimation of clean speech magnitude spectrum by minimizing the statistical error criterion, as proposed by Ephraim and Malah [6, 8]. In [9], a computationally proficient alternative is proposed for SE methods in [6, 8]. In this method, super-Gaussian extension of the joint maximum a posteriori (JMAP) estimation rule is proposed to estimate the speech. By using the Super-Gaussian model of speech, mean squared error is minimized compared to Gaussian model [10].

Coherence-based method dealing with coherent noise is appropriate for HADs and CI devices [25]. The theory behind these methods is that the speech from the two microphones is correlated, while the noise is uncorrelated with speech. Based on this, a gain function is defined to filter the noisy speech [25]. Using the coherence-based function, noise is suppressed along with distortion in the speech [4]. A weighted combination of the coherence gain function and SGJMAP SE gain resulted in better speech quality and intelligibility. The efficiency of this proposed method makes it computationally capable of implementing on smartphones to work seamlessly with HADs.

In this chapter, a parameter called 'weighting' factor is introduced in the proposed SE gain function that can be varied by the user to control the weighted combination, which in turn controls the amount of noise suppression and speech distortion. The parameter can be adjusted depending on the background noise. Various objective and subjective evaluations of the proposed method are carried out for the comparison of the method against the existing benchmark techniques considered.

3.2 Conventional SGJMAP Method

In the SGJMAP [10] method, a non-Gaussianity property in spectral domain noise reduction framework is considered for the usage of super Gaussian speech model [30, 31]. Consider noisy speech $y(n)$, with clean speech $x(n)$ and noise $w(n)$,

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

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

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

where X and W are the clean speech and noise DFT coefficients respectively. In polar coordinates, (3.2) can be written as,

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

where $R_k(\lambda)$, $A_k(\lambda)$, $B_k(\lambda)$ are DFT amplitude of noisy speech, clean speech and noise respectively. $\theta_{Y_k}(\lambda)$, $\theta_{X_k}(\lambda)$, $\theta_{W_k}(\lambda)$ are the phases of noisy speech, clean speech and noise respectively. The purpose is to estimate clean speech magnitude spectrum $A_k(\lambda)$ and its phase spectrum $\theta_{X_k}(\lambda)$. λ is dropped in further discussion for swiftness. The JMAP estimator of the amplitude and phase jointly maximize the probability of amplitude and phase spectra conditioned on the observed complex coefficient given by,

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

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

Using the super Gaussian speech model, spectral amplitude estimator allows the probability density function (PDF) of the speech spectral amplitude to be approximated by the function of two parameters μ and v. The super-Gaussian PDF [30] of the amplitude spectral coefficient with variance σ_{S_k} is given by,

$$p(A_k) = \frac{\mu^{v+1}}{\Gamma(v+1)} \frac{A_k^v}{\sigma_{X_k}^{v+1}} exp\left\{-\frac{\mu A_k}{\sigma_{X_k}}\right\} \tag{3.6}$$

where Γ denotes the Gamma function.

Taking logarithm of (3.4), and differentiating with respect to A_k gives,

$$\frac{d}{dA_k} log(p(Y_k|A_k, \theta_{X_k}) p\left(A_k, \theta_{X_k}\right)) =$$

$$\frac{-(Y_k^* - A_k e^{-j\theta_{X_k}})\left(-jA_k e^{j\theta_{X_k}}\right) + (Y_k - A_k e^{j\theta_{X_k}})(jA_k e^{-j\theta_{X_k}})}{\hat{\sigma}_{W_k}^2} \tag{3.7}$$

Setting (3.7) to zero and substituting $Y_k = R_k e^{j\theta_{Y_k}}$ simplifies to

$$\frac{2R_k}{\hat{\sigma}_{W_k}^2} - \frac{2A_k}{\hat{\sigma}_{W_k}^2} + \frac{v}{A_k \beta} - \frac{\mu}{\hat{\sigma}_{X_k}} = 0 \tag{3.8}$$

On simplifying (3.8), the following quadratic equation is obtained,

$$A_k^2 + \frac{A_k}{2\hat{\sigma}_{X_k}}\left(\hat{\sigma}_{W_k}^2 \mu - 2R_k \hat{\sigma}_{X_k}\right) - \frac{v\hat{\sigma}_{W_k}^2}{2} = 0 \tag{3.9}$$

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

$$\hat{A}_k = \left(u + \sqrt{u^2 + \frac{v}{2\hat{\gamma}_k}}\right) R_k,\ u = \frac{1}{2} - \frac{\mu}{4\sqrt{\hat{\gamma}_k \hat{\xi}_k}} \tag{3.10}$$

where $\hat{\xi}_k = \frac{\hat{\sigma}_{X_k}^2}{\hat{\sigma}_{W_k}^2}$ 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) [32]. $\hat{\sigma}_{X_k}$ is the estimated instantaneous clean speech power

spectral density. In [10], $v = 0.1$ and $\mu = 1.5$ is shown to give better results. The optimal phase spectrum and the noisy phase are assumed the same $\hat{\theta}_{X_k} = \theta_{Y_k}$.

The speech magnitude spectrum estimate is

$$\hat{A}_k = G_k \, R_k \tag{3.11}$$

where

$$G_k = \left[u + \sqrt{u^2 + \frac{v}{2\hat{\gamma}_k}}\right] \tag{3.12}$$

As in [33], it is considered that phase is perceptually insignificant. For the complete derivation of the gain function, we refer [10].

3.3 Coherence-based gain function.

Two microphones of Google Pixel placed apart (by about 13 cm) was considered, such that speech source and noise source are separated spatially and assumed to be at angles *0* and θ respectively [4], where $0^0 \leq \theta \leq 360^0$. The noisy speech is defined as,

$$y(n) = x_j(n) + w_j(n) \qquad (j = 1, 2) \tag{3.13}$$

where j is the microphone index, $x_j(n)$ and $w_j(n)$ are speech and noise components respectively at each microphone. The Short Time Fourier Transform (STFT) of (3.13) is defined as,

$$Y_j(\omega, l) = X_j(\omega, l) + W_j(\omega, l) \qquad (j = 1, 2) \tag{3.14}$$

where l is the frame index and $\omega = 2\pi n/N$, where $n = 0, 1, 2, \ldots, N-1$, N is the number of FFT points. ω lies in the range of $[-\pi, \pi]$. The complex coherence function between the two signals is given by,

$$\Gamma_{y_1 y_2}(\omega, l) = \frac{\Phi_{y_1 y_2}(\omega, l)}{\sqrt{(\Phi_{y_1 y_1}(\omega, l)\Phi_{y_2 y_2}(\omega, l))}} \tag{3.15}$$

where $\Phi_{uv}(\omega, l)$ is the cross-power spectral density (CSD) defined as $\Phi_{uv}(\omega, l) = E[U(\omega, l)V^*(\omega, l)]$ and $\Phi_{uu}(\omega, l) = E[|U(\omega, l)|^2]$ is the power spectral density (PSD). According to [4], the noise and speech components are uncorrelated. Therefore, CSD of the microphone signals can be written as,

$$\Phi_{y_1y_2}(\omega, l) = \Phi_{x_1x_2}(\omega, l) + \Phi_{w_1w_2}(\omega, l) \tag{3.16}$$

Dividing both sides of (3.16) by $\sqrt{(\Phi_{y_1y_1}(\omega, l)\Phi_{y_2y_2}(\omega, l))}$ and neglecting ω and l we get,

$$\Gamma_{y_1y_2} = \Gamma_{x_1x_2}\sqrt{\frac{\Phi_{x_1x_1}}{\Phi_{x_1x_1} + \Phi_{w_1w_1}}}\sqrt{\frac{\Phi_{x_2x_2}}{\Phi_{x_2x_2} + \Phi_{w_2w_2}}} + \Gamma_{n_1n_2}\sqrt{\frac{\Phi_{w_1w_1}}{\Phi_{x_1x_1}+\Phi_{w_1w_1}}}\sqrt{\frac{\Phi_{w_2w_2}}{\Phi_{x_2x_2}+\Phi_{w_2w_2}}} \tag{3.17}$$

True Signal to Noise Ratio (SNR) at the j^{th} microphone is given by,

$$SNR_j = \frac{\Phi_{x_jx_j}}{\Phi_{w_jw_j}} \qquad (j = 1, 2) \tag{3.18}$$

Considering the closeness of the two microphones, we can assume that $SNR_1 \approx SNR_2 \approx \widehat{SNR}$. It can be seen that coherence of speech is dominant at high SNR values and that of noise is dominant at low SNR values. From an approximation for the coherence value given in [25], (3.17) can be written as,

$$\hat{\Gamma}_{y_1y_2} = [cos(\omega\tau) + j\, sin(\omega\tau)]\frac{\widehat{SNR}}{1 + \widehat{SNR}} + [cos(\omega\tau\, cos\theta) + j\, sin(\omega\tau\, cos\theta)]\frac{1}{1+\widehat{SNR}} \tag{3.19}$$

where $\tau = f_s(d/c)$, d is the microphone spacing, c is the speed of sound and f_s is the sampling frequency. We make use of a suppression filter proposed in [25], where in 2 separate filters are used to suppress noise from certain range of θ values. For θ values around 90^0, the suppression filter is,

$$G_1(\omega, l) = 1 - \left|\mathcal{R}\left[\hat{\Gamma}_{y_1 y_2}(\omega, l)\right]\right|^{P(\omega)} \tag{3.20}$$

where $\mathcal{R}[.]$ is the real part and $P(\omega)$ is defined in two frequency bands as,

$$P(\omega) = \begin{cases} \alpha_{low}, & if\ |\omega| \leq \frac{\pi}{8} \\ \alpha_{high}, & if\ |\omega| > \frac{\pi}{8} \end{cases} \tag{3.21}$$

where α_{low} and α_{high} are positive integer constants such that $\alpha_{low} > \alpha_{high} > 1$. When $90^0 < \theta \leq 180^0$, the gain function becomes,

$$G_2(\omega, l) = \begin{cases} \mu, & if\ \Im\left(\hat{\Gamma}_{y_1 y_2}(\omega, l)\right) < Q(\omega) \\ 1, & Otherwise \end{cases} \tag{3.22}$$

where $Q(\omega)$ is defined as,

$$Q(\omega) = \begin{cases} \beta_{low}, & if\ |\omega| \leq \frac{\pi}{8} \\ \beta_{high}, & if\ |\omega| > \frac{\pi}{8}. \end{cases} \tag{3.23}$$

where β_{low} and β_{high} are negative constants such that $\beta_{low} > \beta_{high} > -1$. For further details on these gain functions, we refer to [25]. The final coherence based gain function is defined as,

$$G_{coh}(\omega, l) = G_1(\omega, l) G_2(\omega, l) \tag{3.24}$$

3.4 Weighted Combination of $G_k(\omega, l)$ and $G_{coh}(\omega, l)$

The block diagram of the proposed method is as shown in Figure 3.1. Windowing and STFT is applied on to the two microphone signals to convert them to frequency domain [4]. Though the coherence based gain function in (3.24) suppresses noise, the speech signal sounds somewhat mechanical. The quality of the speech can be retained by using G_k, which is the SGJMAP SE gain, but it introduces musical noise for background noise types such as babble or car noise which are non-stationary in nature. To bypass this limitation, we introduce a new gain function $G_{final}(\omega, l)$ given by,

$$G_{final}(\omega, l) = \varpi G_k(\omega, l) + (1 - \varpi) G_{coh}(\omega, l) \tag{3.25}$$

where ϖ is the weighting factor that helps the user to switch between noise suppression and speech distortion. At high values of ϖ, the final gain $G_{final}(\omega, l)$ results in good noise suppression and limited speech distortion.

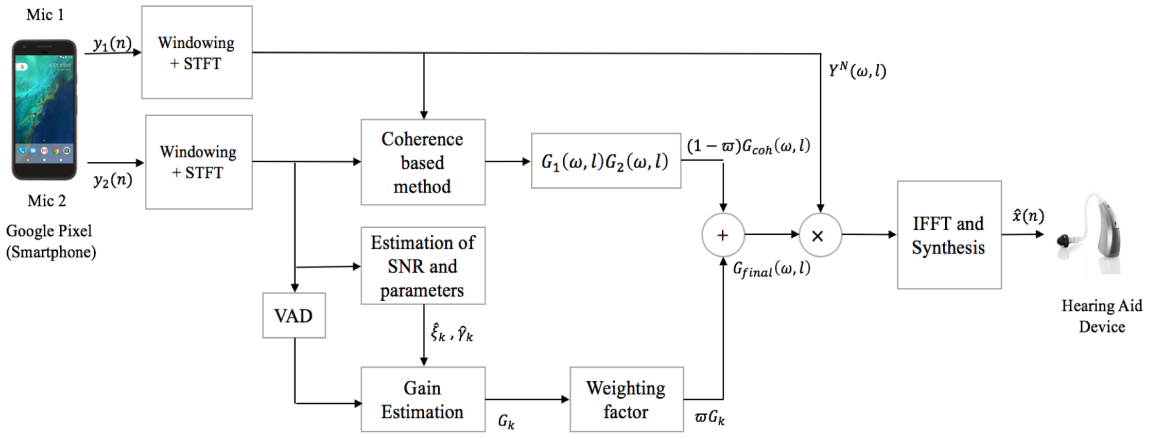


Figure 3.1 Block Diagram of Proposed SE Method

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

$$\hat{X}_k = \hat{G}_{final}\, Y_k \tag{3.26}$$

The time domain reconstruction signal $\hat{x}(n)$ is obtained by taking Inverse Fast Fourier Transform (IFFT) of $\hat{S}_k$.

3.5 Smartphone implementation to function as an assistive device to HA

In this work, Google Pixel with Android 7.1 Nougat operating system is considered as an assistive device for HA. The above device has an M4/T4 HA Compatibility rating and meets the requirements set by Federal Communications Commission (FCC) [34]. The noisy speech was considered at the sampling rate of 48 kHz and 20 ms frames with 50% overlap. The computational time for each frame is around 12 ms on Pixel. The parameter values are hard coded based on [10]. The values of v and μ are set as 0.1 and 1.5 respectively for the test results but can be varied depending on the noisy environment. The range of ϖ is from 0 to 1. At $\varpi = 0.3$ it is seen to provide better results. Fig. 3.2 shows a screenshot of the application implemented on Google Pixel. Turning on the 'OFF' button enables SE method to process the incoming audio stream by applying the proposed SE algorithm on the magnitude spectrum of noisy speech. The enhanced speech signal is then played back through the HAD or through any headphones for normal hearing people. The smartphone application consumes low power because of the computational efficiency of the developed algorithm. The audio streaming is encoded for Bluetooth low energy consumption.

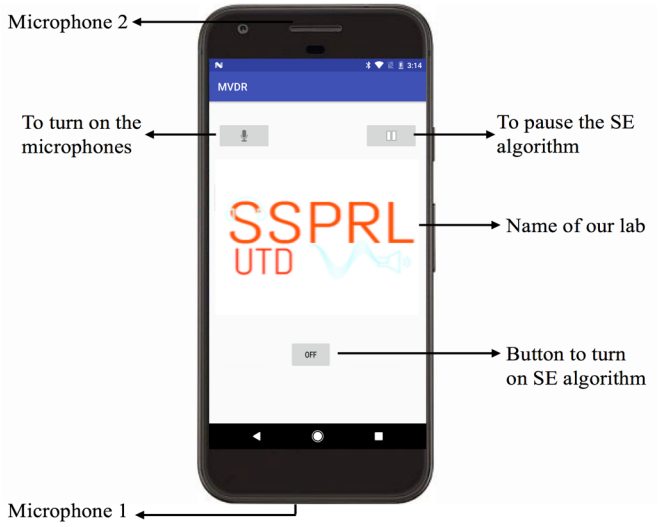


Figure 3.2 Snapshot of the developed SE application

3.6 Experimental Results and Discussion

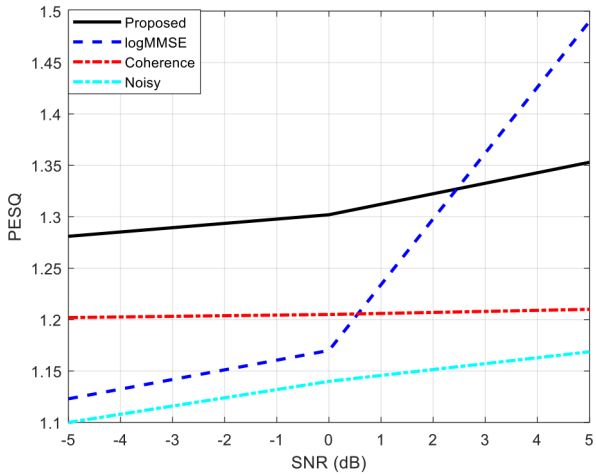
3.6.1 Objective Evaluations

Realistic recordings of machinery and babble noise are added to speech signals taken from IEEE corpus [35] and TIMIT database. For the objective measure of quality of speech, we use Perceptual Evaluation of Speech Quality (PESQ) [36]. Coherence Speech Intelligibility Index (CSII) [37] is used to measure the intelligibility of speech. PESQ ranges between -0.5 and 4.5, with 4.5 being high speech quality. CSII ranges between 0 and 1, with 1 being high intelligibility. Figure 3.3 shows the plots of PESQ and Figure 3.4 shows the plots of CSII versus SNR for three noise types. In comparison with single and dual microphone SE methods such as log-MMSE and coherence based techniques respectively, the proposed method gives better values of PESQ and CSII as shown in Figure 3.3 and Figure 3.4 respectively for machinery, babble, and traffic noise types.

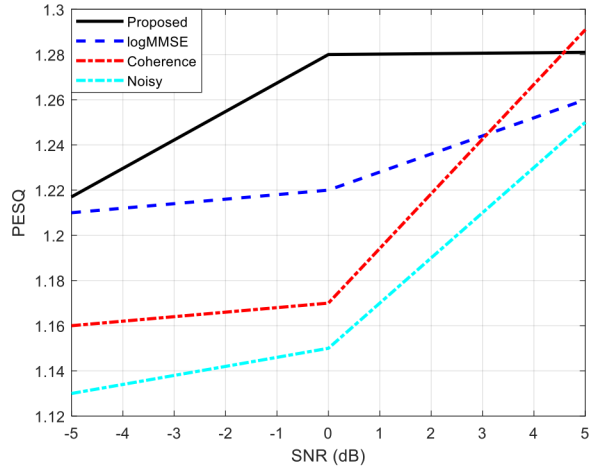
3.6.2 Subjective Evaluations

Along with objective measures, we perform Mean Opinion Score (MOS) tests on 20 normal hearing both male and female subjects. They were presented with noisy speech and enhanced speech using the proposed, log-MMSE and coherence methods at different SNR levels of -5 dB, 0 dB, and 5 dB. The subjects had to choose a suitable ϖ based on the level of comfort, but were also instructed regarding the standard value.

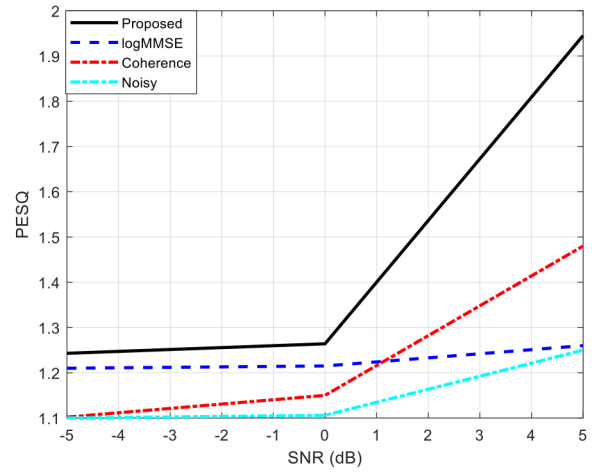
Each subject was instructed to score in the range 1 to 5 for the different audio files based on the following criteria: 5 being excellent speech quality and imperceptible level of distortion. 4 for good speech quality with perceptible level of distortion. 3 stood for fair speech quality with mediocre level of distortion. 2 for poor speech quality with lot of disturbances, causing uneven distortions. 1 having the least quality of speech and intolerable level of distortion. Subjective test results are shown in Figure 3.5, which illustrates the effectiveness of the proposed method in various background noise, simultaneously upholding the speech quality and intelligibility.



a)

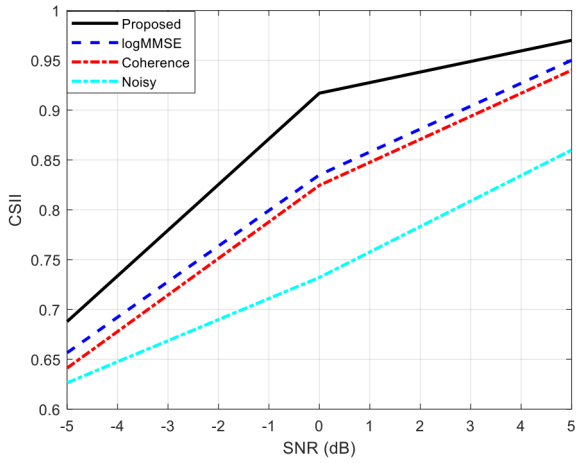


b)

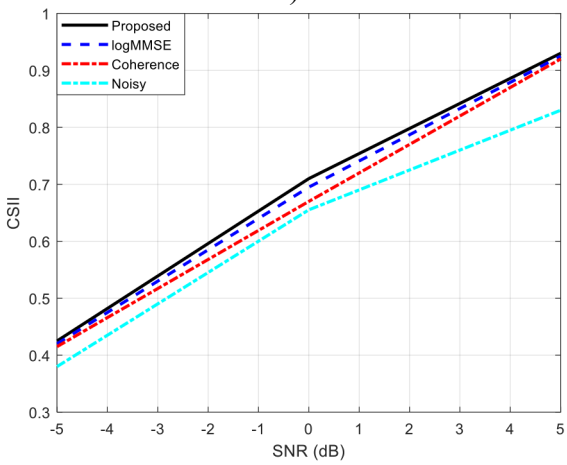


c)

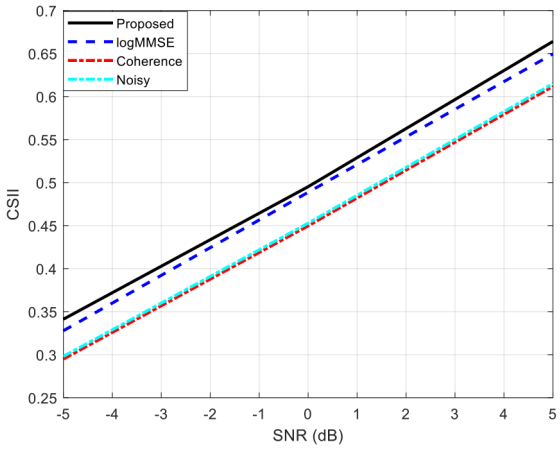
Figure 3.3 Comparison of PESQ scores a) Machinery Noise b) Babble Noise and c) Traffic Noise



a)



b)



c)

Figure 3.4 Comparison of CSII scores a) Machinery Noise b) Babble Noise and c) Traffic Noise

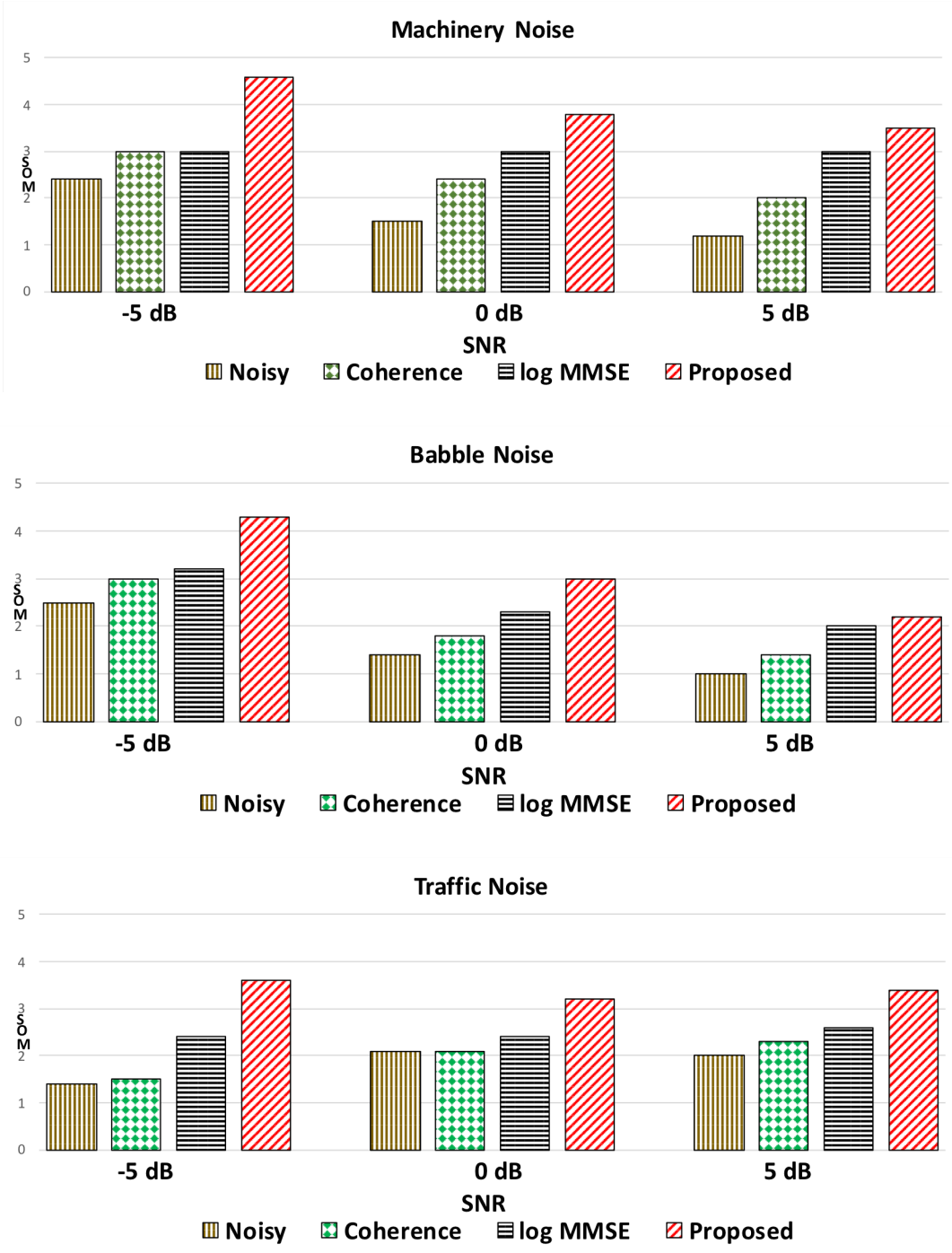


Figure 3.5 Comparison of Subjective Test scores

3.6.3 Analysis and Discussion

Time domain plots: Time domain is the analysis of signals with respect to time. In time domain, the signal is known for real numbers, against either continuous time or discrete time. A time domain plot shows how the signal changes with time. The results are shown in Fig. 3.6 using time domain plots in MATLAB. Actual recorded machinery noise is mixed synthetically with clean speech obtained from TIMIT database sampled at 16 KHz at an SNR of 0 dB.

PESQ Plot Analysis with Different values of ϖ

In this part, we see the detailed analysis of the proposed method which considers different values for the weighting parameter. In this evaluation, we vary the ϖ for 3 different Noise types (Machinery, Babble, and Traffic) at -5, 0, and 5 dB SNR. 10 clean speech files obtained from the TIMIT database are considered. We calculate the PESQ values by varying ϖ from 0 to 1 in terms of 0.1. For all the three noise type, we have seen that by setting ϖ low, can achieve maximum PESQ values. For the machinery noise file, $\varpi = 0.3$, for the babble noise file $\varpi = 0.3$ yielded maximum PESQ value for all 10 files. In presence of traffic noise, ϖ varies for different SNR to achieve maximum PESQ. These tests and results supported our claim that proposed SE method and its application is user adaptive based on the background noise. Figure 3.7 shows the surface plots for different noises for attaining optimal PESQ value

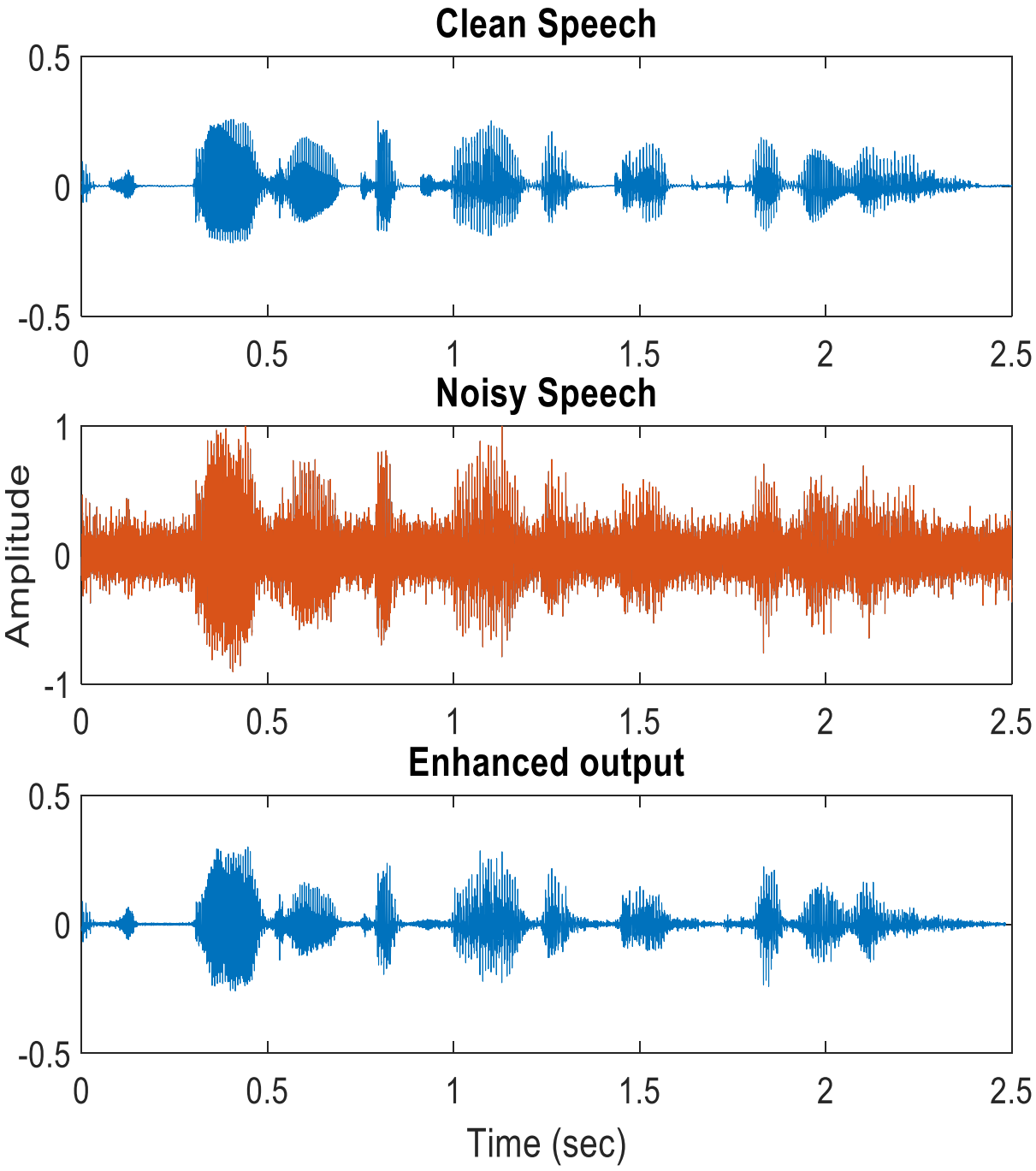
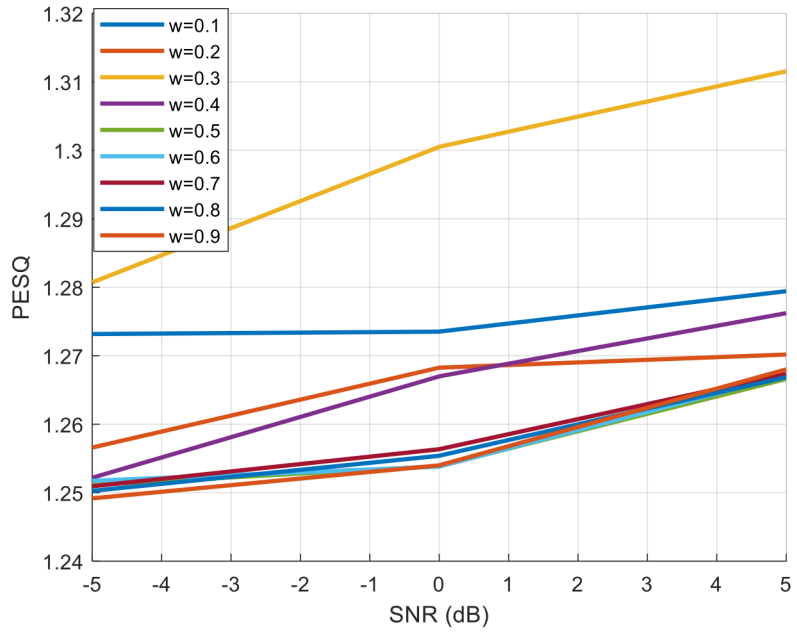
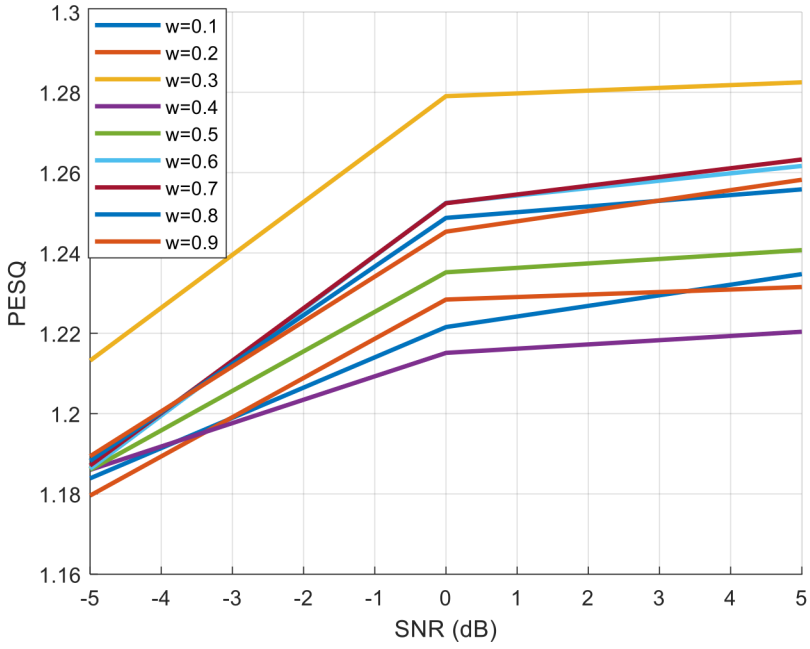


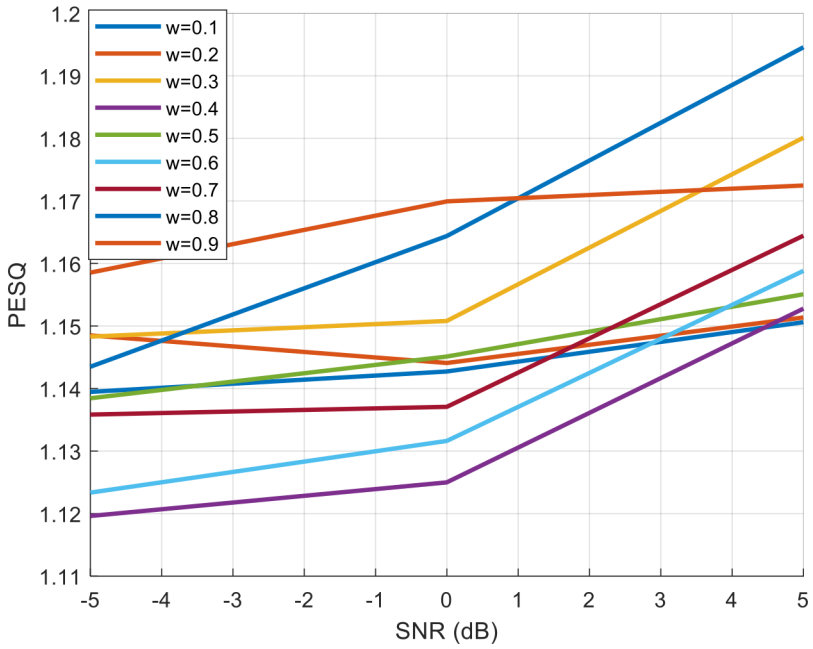
Figure 3.6 Time domain plots of Clean Speech, Noisy Speech and Enhanced Output using Machinery Noise at -5 dB SNR



a) Machinery Noise



b) Babble Noise



c) Traffic Noise

Figure 3.7 Choice of weighting parameter for attaining optimal value of PESQ
a) Machinery Noise b) Babble Noise c) Traffic Noise

3.7 Chapter Conclusion

A Noise Dependent Super Gaussian - Coherence based dual microphone SE algorithm was developed with a weighting factor in the gain function. The obtained gain allows HAD user to control the amount of noise suppression and speech distortion. The proposed algorithm can run on a smartphone device in real time, which works as an assistive device for HA. The weighting parameter permits the smartphone user to control the amount of noise suppression (quality) and speech distortion (intelligibility). The objective and subjective evaluations verify the proposed method to be an apt option to use for hearing aid application in the real-world noisy environment.

CHAPTER 4

INFLUENCE OF MINIMUM VARIANCE DISTORTIONLESS RESPONSE BEAMFORMER ON LOG SPECTRAL AMPLITUDE ESTIMATOR

4.1 Introduction

Over the last few decades, researchers have developed beamforming algorithms, which can be classified into fixed and adaptive beamformers. Fixed beamformers have static filter coefficients and signal independent spatial response [23]. The isotropic model, which is a first-order approximation of real noise fields, is commonly used in these beamformers and the noise field is not known. The beamforming filter coefficients can be changed, leading to the second class of adaptive beamformers. Among several SE techniques, adaptive beamformers are commonly used to improve the performance of the algorithm. The MVDR beamformer [38-41] has wide range applications for extraction of desired speech signals in noisy environments. The MVDR beamformer, known as Capon beamformer [42], dating back to 1980s, minimizes the output power of the beamformer under a single linear constraint on the response of the array towards the preferred signal. This spatial filtering process plays a critical role in extracting the signal of interest, suppressing ambient noise, and separating multiple sound sources. MVDR beamformer requires less a priori knowledge, which makes it practical for implementing it as a smartphone-based SE application for HADs.

The proposed algorithm is a combination of MVDR beamformer and Minimum mean square error Log spectral amplitude estimator (Log-MMSE) SE gain function, for suppressing noise and extracting the desired speech. This method is computationally efficient and helps in achieving minimal speech distortion for the hearing-impaired. Performance of the proposed method is

compared against standard techniques of SE for speech quality and intelligibility. Subjective evaluations show promising results of the real-time application.

4.2 Proposed SE Gain Function

In the smartphone application, signals captured by the two microphones is composed of both clean speech and the background noise. Figure 4.1 shows the block diagram of the proposed method implemented on android based smartphone. We consider a signal model with the first microphone as the reference point, the signal received by the n^{th} microphone (n = 1, 2) can be written as,

$$y_n(t) = s_n(t) + w_n(t) = s(t - \tau_n) + w_n(t) \tag{4.1}$$

where $y_n(t)$, $s_n(t)$ and $w_n(t)$ are noisy speech, clean speech and noise signals respectively picked up by the n^{th} microphone at time t. Let τ_0 be the relative time delay between the two microphones given by $\tau_0 = \delta/c$ with δ as the spacing between the two microphones and c being the speed of sound in air. The signals are considered to be zero mean and real, noise signal $w_n(t)$ are assumed to be uncorrelated with $s_n(t)$.

For efficient performance, the signals are transformed to frequency domain and are re-written as,

$$\begin{aligned} Y_n(\omega) &= S_n(\omega) + W_n(\omega) \\ &= e^{-j(n-1)\omega\tau_0 \cos(\theta_d)} S(\omega) + W_n(\omega) \end{aligned} \tag{4.2}$$

where $Y_n(\omega)$, $S_n(\omega)$, $W_n(\omega)$ are the Fourier transforms of $y_n(t)$, $s_n(t)$, $w_n(t)$ respectively. $\omega = 2\pi f$ is the angular frequency. Equation (4.2) can be rearranged into vector form as follows,

$$\begin{aligned} Y(\omega) &\triangleq [Y_1(\omega) \quad Y_2(\omega)]^T \\ &= d_{\theta_d}(\omega) S(\omega) + W(\omega) \end{aligned} \tag{4.3}$$

where the superscript $^{\mathrm{T}}$ is the transpose operator and $S(\omega) = [S_1(\omega)\ S_2(\omega)]^T$, and $W(\omega) = [W_1(\omega)\ W_2(\omega)]^T$

$$d_{\theta_d}(\omega) \triangleq \begin{bmatrix} 1 & e^{-j\omega\tau_0 \cos(\theta_d)} \end{bmatrix}^T \tag{4.4}$$

is the steering vector and the noisy signal $W(\omega)$, is defined similar to $Y(\omega)$. θ_d is the angle of incidence of the source at the plane of microphones. Since the signals are assumed to be uncorrelated, the correlation matrix of $Y(\omega)$ can be determined by the method explained in [23].

4.2.1 MVDR beamformer

The goal of beamforming is to extract the desired speech signal, by applying a linear filter to noisy speech. The output of the beamformer is given by,

$$\begin{aligned} Z(\omega) &= \sum_{n=1}^{2} H_n^*(\omega)\, Y_n(\omega) = H^H Y \\ &= h^H(\omega) d_{\theta_d}(\omega) S(\omega) + h^H(\omega)\boldsymbol{w}(\omega) \end{aligned} \tag{4.5}$$

where $Z(\omega)$ is the output of the beamformer, $h^H(\omega)d_{\theta_d}(\omega)S(\omega)$ is the filtered speech signal, and $h^H(\omega)\boldsymbol{w}(\omega)$ is the residual noise.

The MVDR beamformer output can be obtained by minimizing the variance on either side of (4.5), or the residual noise with the constraint that the signal from the desired direction is without any distortion. In this work, we consider the minimization of variance of the residual noise.

$$\min_{h(\omega)} E[|h^H(\omega)\boldsymbol{w}(\omega)|^2] \text{ subject to } h^H(\omega)d_{\theta_d}(\omega) = 1 \tag{4.6}$$

E[.] denotes mathematical expectation. Using a Lagrange multiplier to adjoin the constraint to the objective function, then differentiating with respect to $h(\omega)$, and equating the result to zero, (4.6) can be reduced to,

$$h(\omega) = \frac{\Gamma_{\boldsymbol{w}}^{-1}(\omega) d_{\theta_d}(\omega)}{d_{\theta_d}^{H}(\omega) \Gamma_{\boldsymbol{w}}^{-1}(\omega) d_{\theta_d}(\omega)} \tag{4.7}$$

where $\Gamma_{\boldsymbol{w}}(\omega) = \Phi_{\boldsymbol{w}}(\omega)/\phi_{W1}(\omega)$ is the pseudo-coherence matrix of the noise with $\Phi_{\boldsymbol{w}}(\omega) = E[\boldsymbol{w}(\omega)\boldsymbol{w}^{H}(\omega)]$ and $\phi_{W1}(\omega) = E[|W_1(\omega)|^2]$.

4.2.2 Gain function based on Log-Spectral Amplitude Estimator

In the Log-MMSE method, speech and noise models are considered to be statistically independent Gaussian Random Variables [43]. The aim is to minimize the mean squared error of log magnitude spectra between estimated and true speech. The input is taken to be the output of the MVDR beamformer $z(n)$, which contains filtered speech signal $s'(n)$, and some residual noise $w'(n)$,

$$z(n) = s'(n) + w'(n) \tag{4.8}$$

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

$$Z_k(\lambda) = S'_k(\lambda) + W'_k(\lambda) \tag{4.9}$$

Where S′ and W′ are the input speech and noise DFT coefficients. In polar coordinates, (4.9) can be written as,

$$R_k(\lambda) e^{j\theta_{Z_k}(\lambda)} = A_k(\lambda) e^{j\theta_{S'_k}(\lambda)} + B_k(\lambda) e^{j\theta_{W'_k}(\lambda)} \tag{4.10}$$

Where $R_k(\lambda)$, $A_k(\lambda)$, $B_k(\lambda)$ are magnitude spectra of noisy speech, input signal and noise respectively. $\theta_{Z_k}(\lambda)$, $\theta_{S'_k}(\lambda)$, $\theta_{W'_k}(\lambda)$ are the phase spectra of noisy, input speech and noise respectively. Looking at the estimator $\hat{A}_k$, which minimizes the distortion measure as explained in [8], the mean-square error of the log-magnitude spectra is given by,

$$E\left\{\left(log\, A_k - log\, \hat{A}_k\right)^2\right\} \tag{4.11}$$

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,

$$log\, \hat{A}_k = E\{log\, A_k\, |Z_k(\lambda)\} \tag{4.12}$$

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

$$\hat{A}_k = exp\, (E\{log\, A_k|\, Z_k(\lambda)\}) \tag{4.13}$$

Solving the above expectation, the final estimate of speech magnitude spectrum according to [8] is given by,

$$\hat{A}_k = \frac{\xi_k}{\xi_k+1}\, exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right\}R_k \tag{4.14}$$

$$\triangleq G_k R_k$$

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) [32]. $\sigma_{S'_k}$ is the estimated instantaneous clean speech power spectral density. The optimal phase spectrum is the noisy phase spectrum itself $\theta_{S_k} = \theta_{Y_k}$.The final clean speech estimate is,

$$\hat{S'}_k = G_k\, Z_k \tag{4.15}$$

The time domain reconstruction signal $\hat{s}'(n)$ is obtained by taking inverse Fourier Transform of $\hat{S'}_k$.

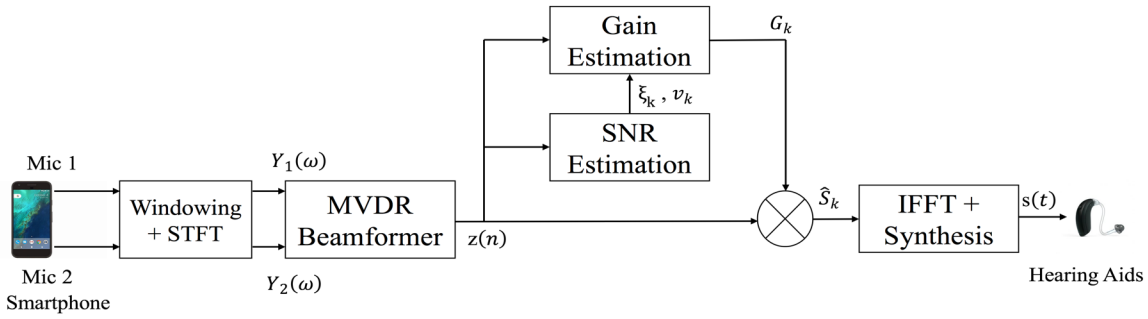


Figure 4.1 Block Diagram of the Proposed SE Method

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

In this chapter, Google Pixel running Android 7.1 Nougat operating system is considered as an assistive device. Two microphones (13 cm apart) on the smartphone capture the audio signal, process the signal and transmit the enhanced signal to the HADs. The smartphone device considered has an M4/T4 HA Compatibility rating and meets the requirements set by Federal Communications Commission (FCC). Android Studio [44] is used for implementation of the SE algorithm on the smartphone. An inbuilt android audio framework was used to carry out dual microphone input/output handling. The input data is acquired at 48 KHz sampling rate and a 10ms frame, with FFT size to be 512 is considered as the input buffer. Figure 4.2 shows the screenshot of the proposed SE method implemented on Pixel smartphone. When the button is in "ON" mode, the microphone will record the audio signal and playback to the HADs without any processing. There is another button present on the screen to apply the developed SE algorithm to enhance the audio stream. The enhanced output signal is then played back to the HADs. Initially, when the SE algorithm is turned on, the algorithm uses approximately 3 seconds to estimate the noise variance.

Hence, we assume there is no speech activity during this time. The smartphone application is computationally efficient and consumes less power.

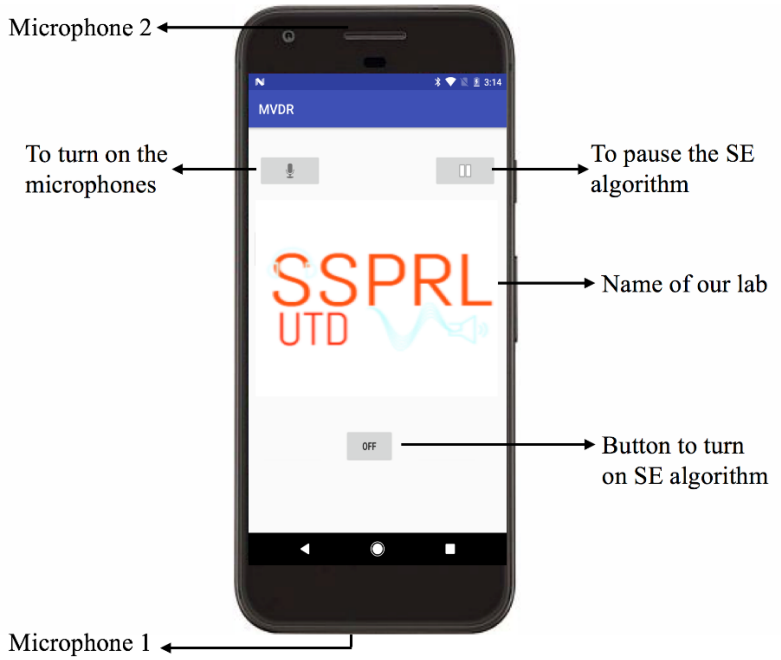


Figure 4.2 Snapshot of developed SE method

4.4 Experimental Results

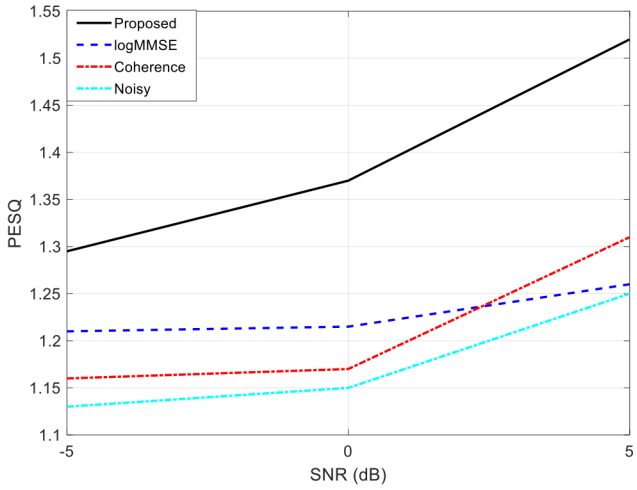
4.4.1 Objective Evaluation

The performance of the proposed method is evaluated by comparing with dual microphone coherence [25] and Log-MMSE [8] methods, promising results are seen. The objective evaluations are performed for 3 different noise types: machinery, multi-talker babble, and traffic noise. The plotted results are the average over different speech signals from the TIMIT database. The audio files are sampled at 16 kHz, and 10 ms frames with 50% overlap are considered. Perceptual

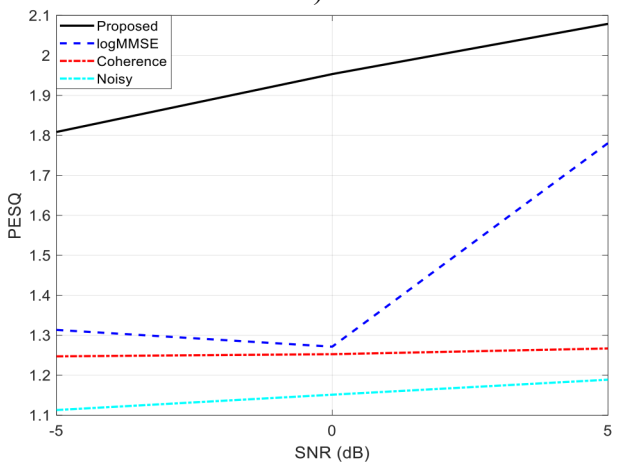
evaluation of speech quality (PESQ) [36] and short time objective intelligibility (STOI) [45] are used to measure the quality and intelligibility of the speech respectively. PESQ ranges between -0.5 and 4.5, with 4.5 being high perceptual quality. Higher the score of STOI better the speech intelligibility. Figure 4.3 shows the plots of PESQ and Figure 4.4 shows the plots of STOI versus 3 different SNR for the 3 noise types. PESQ and STOI values show substantial improvements over other methods for all three noise types considered. Objective and Intelligibility measures state the fact that the proposed SE method suppresses more noise with minimal speech distortion.

4.4.2 Subjective Test setup and Evaluation

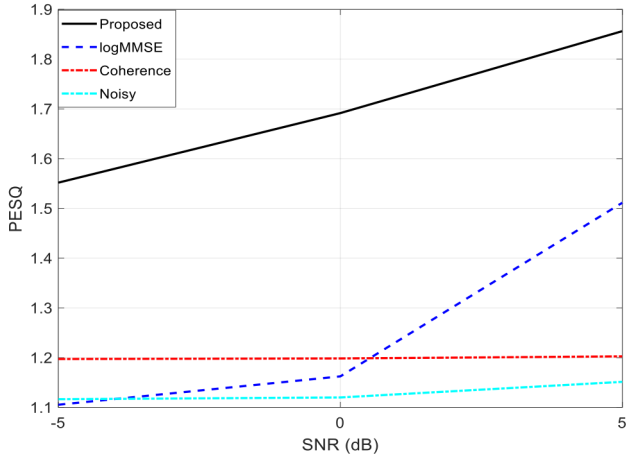
Subjective measures give information about the practical usability of our application in real-time. Thus, Mean Opinion Score (MOS) tests [46] was performed on 10 normal hearing subjects including 5 male and 5 female adults. They were presented with noisy speech and enhanced speech using the proposed, coherence and Log-MMSE methods at different SNR levels of -5 dB, 0 dB, and 5 dB. The audio files were played on headphones for the subjects. Each subject was instructed to rate between 1 and 5 for each audio file based on the following criteria: 5 being excellent speech quality and imperceptible level of distortion. 1 having the least quality of speech and intolerable level of distortion. This test provided a good comparison between the proposed method and other existing methods. Subjective test results in Figure 4.5 illustrate the effectiveness of the proposed method in reducing the background noise, simultaneously preserving the quality and intelligibility of the speech.



a)

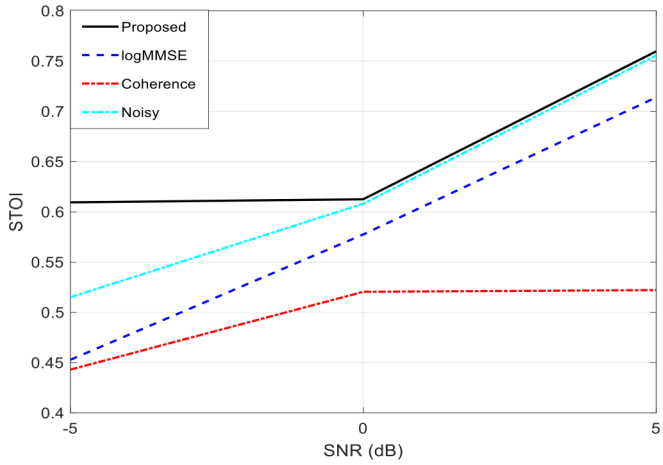


b)

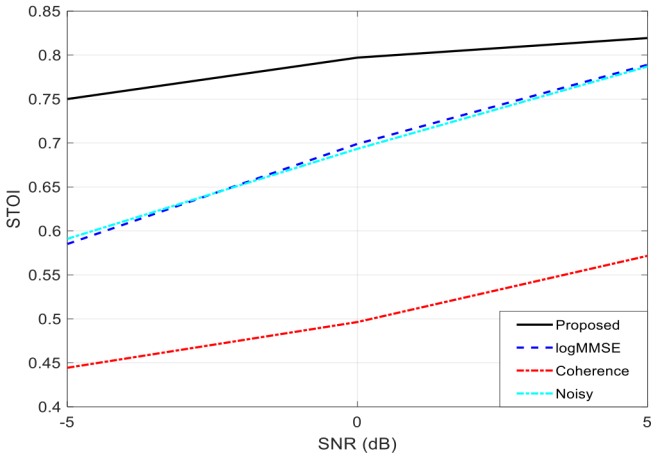


c)

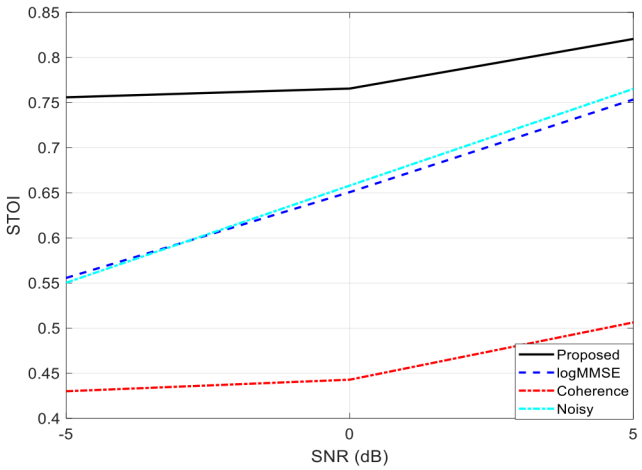
Figure 4.3 Comparison of PESQ scores for (a) Machinery noise, (b) Babble noise and (c) Traffic noise



a)

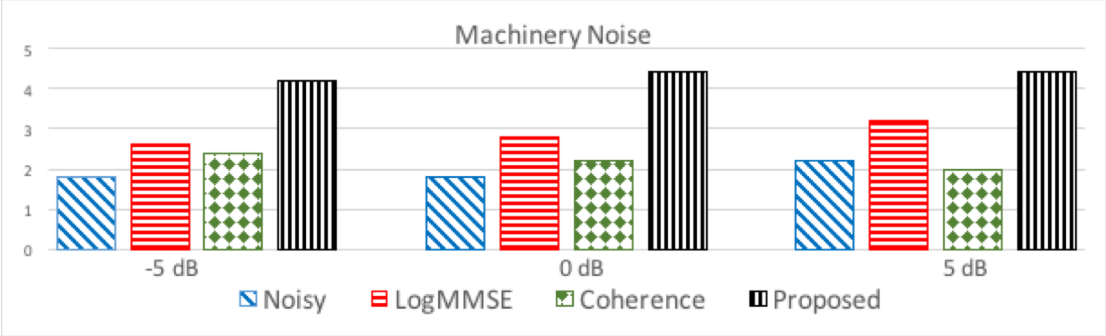


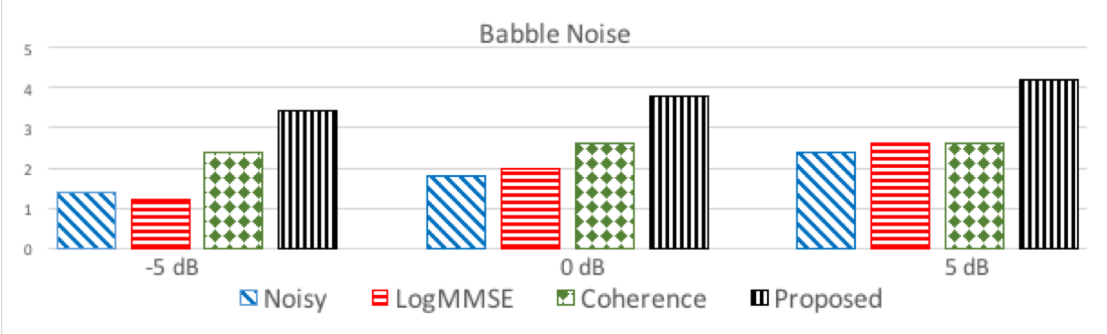
b)



c)

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





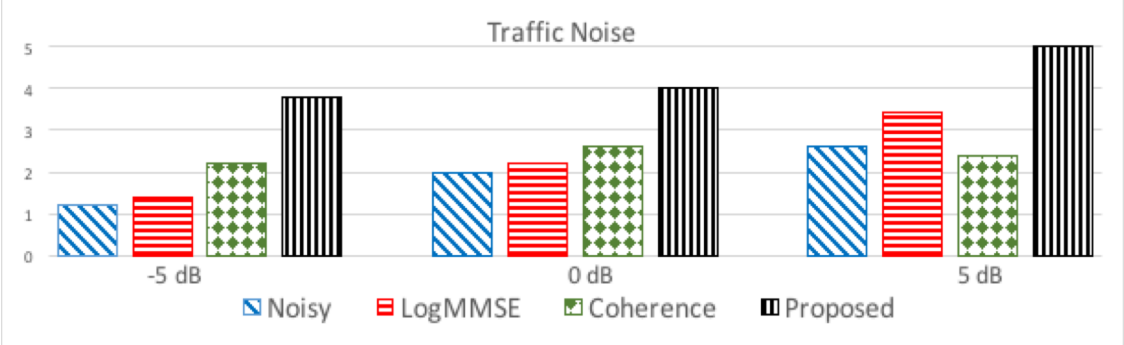


Figure 4.5 Comparison of Subjective results

4.4.3 Time-domain plots

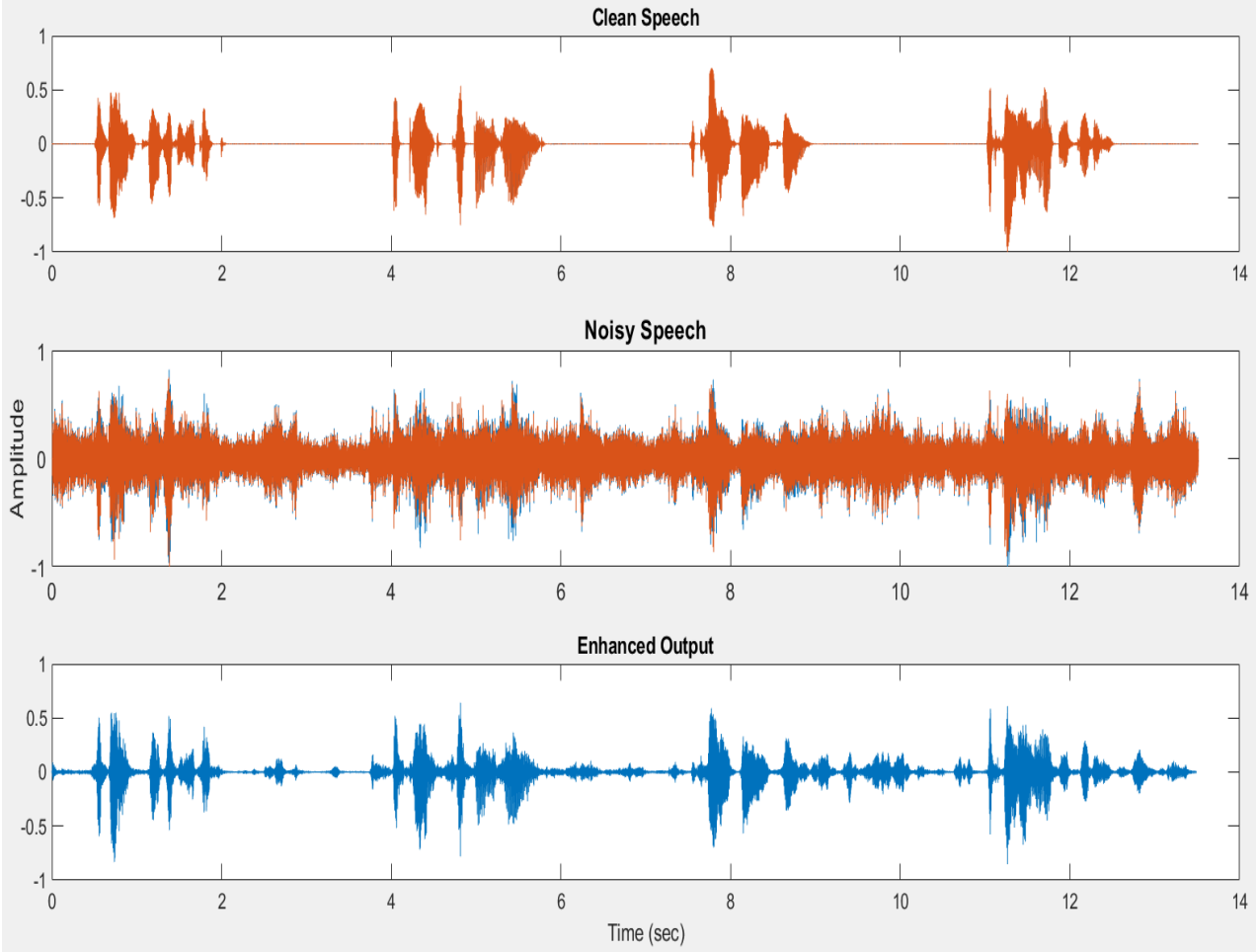


Figure 4.6 Time domain plots of Clean Speech, Noisy Speech and Enhanced Output using Babble Noise at 0 dB SNR

4.5 Chapter Outcomes

An MVDR beamformer based dual microphone SE algorithm was developed and implemented on a smartphone as a real-time application. This method can act as an assistive device for HADs. Objective and Subjective evaluations verify that the proposed method can be used as a solution to enhance the speech in real-world noisy environments.

CHAPTER 5

CONCLUSION

In this thesis, dual microphone Speech Enhancement techniques are developed and are implemented on Android-based smartphones to work as an assistive device to Hearing Aids.

In Chapter 3, a dual microphone SE was proposed that makes use of the coherence-based function to suppress noise with minimal speech distortion. The proposed SE method makes use of the gain function of the new super-Gaussian Joint Maximum *a Posteriori* (SGJMAP). The combined SE method is implemented on an Android-based smartphone to work in real-time. The weighted union of these two gain functions strikes a balance between noise suppression and speech distortion. A weighting parameter introduced in the derived gain function allows the smartphone user to control the weighting factor based on different background noise and their comfort level of hearing.

In Chapter 4, the two-microphone minimum variance distortionless response (MVDR) beamformer was used as an SNR booster to the SE algorithm. The computational efficiency of the algorithm allows this setup to be implemented on a smartphone. When the microphone array and the speech source take the end-fire setup, best performance can be observed when compared to other algorithms that use only single microphone. The limitation of this approach is, it works well only in the end-fire case.

The objective and subjective results of the proposed methods show significant improvements and prove the usability of the developed application in real-world noisy conditions.

REFERENCES

[1] Y. Ephraim and I. Cohen, “Recent Advancements in Speech Enhancement,” in the Electrical Engineering Handbook, CRC Press, 2006, ch. 15, pp. 12-26.

[2] T. J. Klasen, T. Van den Bogaert, M. Moonen and J. Wouters, "Binaural Noise Reduction Algorithms for Hearing Aids That Preserve Interaural Time Delay Cues," in IEEE Transactions on Signal Processing, vol. 55, no. 4, pp. 1579-1585, April 2007.

[3] Yu-Ting Kuo, Tay-Jyi Lin, Wei-Han Chang, Yueh-Tai Li, Chih-Wei Liu and Shuenn-Tsong Young, "Complexity-effective auditory compensation for digital hearing aids," 2008 IEEE International Symposium on Circuits and Systems, Seattle, WA, 2008, pp. 1472-1475.

[4] 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.

[5] P. C. 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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] Abd El-Fattah, M.A., Dessouky, M.I., Abbas, A.M. et al. Int J Speech Technol (2014) 17: 53.

[17] 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.

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

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

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

[21] Ward D.B., Kennedy R.A., Williamson R.C. (2001) Constant Directivity Beamforming. In: Brandstein M., Ward D. (eds) Microphone Arrays. Digital Signal Processing. Springer, Berlin, Heidelberg

[22] Bitzer J., Simmer K.U. (2001) Superdirective Microphone Arrays. In: Brandstein M., Ward D. (eds) Microphone Arrays. Digital Signal Processing. Springer, Berlin, Heidelberg

[23] C. Pan, J. Chen and J. Benesty, "Performance Study of the MVDR Beamformer as a Function of the Source Incidence Angle," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 1, pp. 67-79, Jan. 2014.

[24] 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.

[25] 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

[26] S. Makino, T-W. Lee, H. Sawada, "Blind Speech Separation," Springer Signals and Communication technology, 2007.

[27] P. Common, "Independent component analysis, a new concept?," Signal Processing, vol. 36, pp. 287-314, 1994.

[28] B. Edwards, "The future of Hearing Aid technology," Journal List, Trends Amplif, v.11(1): 31-45, Mar 2007.

[29] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc of IEEE Conf. on Acoustic SpeechSignal Processing, pp. 208-211, Washington D.C, 1979.

[30] 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.

[31] 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.

[32] 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.

[33] P. Vary, "Noise suppression by spectral magnitude estimation— mechanisms and theoretical limits," Signal Processing, vol. 8, no. 4, pp. 387–400, 1985.

[34] https://support.google.com/pixelphone/answer/7022290 - hac

[35] “IEEE recommended practice for speech quality measurements,” IEEE Trans. Audio Electroacoust., vol. AE-19, no. 3, pp. 225–246, Sep. 1969.

[36] 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.

[37] P. Loizou, “Speech Enhancement: Theory and Practice”, Boca Raton, FL: CRC Press, 2007.

[38] C. Pan, J. Chen and J. Benesty, "On the noisereduction performance of the MVDR beamformer innoisy and reverberant environments," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014, pp. 815-819.

[39] H. Cox, R. M. Zeskind and M. M. Owen, “Robust Adaptive Beamforming”, IEEE Trans. Acoust., Speech and Signal Process., vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[40] H. Cox and R. M. Zeskind, “Reduced Variance Distortionless Response (RVDR) Performance with Signal Mismatch”, in Conf. Record of the 25th Asilomar Conf. on Signals, Systems and Computers, IEEE, Nov. 1991, vol. 2, pp. 825–829.

[41] M. E. Lockwood, D. L. Jones, C. R. Lansing, W. D. O’Brien, Jr., B. C. Wheeler and A. S. Feng, “Effect of Multiple Nonstationary Sources on MVDR Beamformers”, in Conf. Record of the 37th Asilomar Conf. on Signals, Systems and Computers, IEEE, Nov. 2003, vol. 1, pp. 730–734.

[42] J. Capon, “High resolution frequency-wavenumber spectrum analysis,” Proc. IEEE, vol. 57, pp. 1408–1418, Aug. 1969.

[43] 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), Bethesda, MD, 2017, pp. 32-35.

[44] May. 2018. [Online]. Available: https://developer.android.com/studio/intro/index.html

[45] C. H. Taal, R. C. Hendricks, R. Heusdens, and R. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Trans. Audio, Speech, Lang. Process. vol. 19, no. 7, pp. 2125–2136, Feb. 2011.

[46] Subjective performance assessment of telephone- band and wideband digital codecs, ITU-T Rec. P.830, 1996.

BIOGRAPHICAL SKETCH

Nikhil Shankar is from Bangalore, India. He completed his Bachelor of Engineering in Electronics and Communication at Visveswaraya Technological University in May 2016. He worked as an intern at Defense Research and Development Organization (DRDO), Bangalore, from January 2016 to May 2016 where he designed and fabricated a printed dipole antenna. He began his Master's program in Electrical Engineering at UT Dallas in August 2016. He started working in Statistical Signal Processing Research Laboratory (SSPRL) in September 2016. His research includes, developing single and dual microphone Speech Enhancement techniques, Beamforming algorithms and implementation of these algorithms on Smartphone for Hearing Aid applications. He began working as a Teaching Assistant in the Department of Electrical and Computer Engineering at UT Dallas in the spring semester of 2018.

He has published five papers during his research at SSPRL under the guidance of Dr. Issa Panahi.

CURRICULUM VITAE

Nikhil Shankar

Email: nikhillalith@gmail.com

EDUCATION

Master of Science (Thesis), Electrical Engineering GPA: 3.59
The University of Texas at Dallas August 2018
Bachelor of Engineering, Electronics and Communications Engineering GPA: 3.7
Visveswaraya Technological University, India June 2016

SKILLS

Languages/Tools: C, MATLAB, Xcode, Android SDK, Visual Studio, Audacity, Keras, Tensor flow, Familiar with C++, Objective C and Python.
Courses: Digital Signal Processing I, Digital Signal Processing II, Introduction to Speech Processing, Speech Perception Lab, Random Processes, Detection and Estimation Theory, Digital Communication Systems, RF and Microwave Circuits.

RESEARCH EXPERIENCE

Graduate Research Assistant **Sept 2016- Present**
Statistical Signal Processing Research Laboratory (SSPRL) UT Dallas, Richardson, TX

- Development and Real-Time implementation of single and dual channel Speech Enhancement algorithms on smartphone for hearing aids.

Graduate Teaching Assistant **Jan 2018 – Present**
Erik Jonsson School of Engineering and Computer Science UT Dallas, Richardson, TX

- TA for Linear Algebra and Signals & Systems Lab courses. Lead Laboratory sessions, Grade Quiz and Assignments. Guide Students in solving problems and understanding concepts.

RESEARCH PROJECTS

1. **Real-time Blind Source Seperation using Independent Vector Analysis (IVA) (Mar 2018 - Present)**

- Developed and implemented a real-time IVA algorithm on android based smartphone using two microphones.
- Implemented a Generalized Cross Correlation based DOA using feed forward neural network to reduce the computational time.

2. **Application of Beamformer as a pre-filter for Speech enhancement in real time (Oct 2017 – Feb 2018)**

- Developed and implemented a Minimum Variance Distortionless Response (MVDR) beamformer algorithm after determining DOA.

- It is used as a SNR booster for speech enhancement algorithm in real-time applications, using two microphones of the smartphone.

3. Noise dependent Super Gaussian-Coherence based dual microphone speech enhancement (Aug – Sept 2017)

- The coherence between speech and noise signals was used to obtain a Speech Enhancement (SE) gain function.
- A non-Gaussianity property in spectral domain noise reduction framework is considered for the usage of speech model.

4. Voice Activity Detection (VAD) using Artificial Neural Networks (ANN) (June – July 2017)

- Speech features such as Spectral flux, ZCR, Energy entropy and Spectral Centroid were extracted to serve as input nodes to the ANN. The

model was trained using back propagation algorithm in Keras with Tensor flow.

- This was a Classification problem, i.e. to detect speech or noise frame when noisy speech is the input.

5. Formant Frequency based speech enhancement technique (Apr – May 2017)

- Developed a method based on Log spectral amplitude estimator (logMMSE) to improve speech intelligibility in the noisy real world acoustic

environment using the priori information of formant frequency locations.

- Objective Evaluation (PESQ, CSII and STOI) and Subjective tests (MOS) were carried out for speech quality and intelligibility assessment.

6. Super Gaussian Joint Maximum a Posteriori (SGJMAP) based single microphone speech enhancement (Jan – Mar 2017)

- A SGJMAP gain function was derived with a "trade-off" parameter allowing the smartphone user to customize their listening preference, by

controlling the amount of noise suppression and speech distortion in real-time.

- The above algorithm was coded in both MATLAB and C and implemented on both iOS and Android based smartphones for hearing aids.
- Clinical testing of the application was carried out on hearing impaired subjects in anechoic chamber and sound booth.

7. Real-time Audio Compression for hearing aids (Sept – Dec 2016)

- MATLAB code for audio compression using filter banks was converted to C.
- The C code was simulated and implemented on a smartphone to work in real-time for hearing aids application.

COURSE PROJECTS

1. System Identification for Real Data using Adaptive filter algorithms like LMS and NLMS.

- A Mathematical model of the unknown system was determined using the adaptive filter algorithms.

PUBLICATIONS

1. **N. Shankar**, G. S. 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.
2. **N. Shankar**, Abduallah Kucuk, C. Karadagur Ananda Reddy, G. S. Bhat and I. Panahi, "Influence of MVDR beamformer on speech enhancement based Smartphone application for Hearing Aids," – 40th International conference of the IEEE Engineering in medical and Biology.
3. 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.
4. 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), Bethesda, MD, 2017, pp. 32-35.
5. 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.