FOREGROUND SPEECH SEGMENTATION AND ENHANCEMENT

DEEPAK K T

Foreground Speech Segmentation and Enhancement

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

DEEPAK K T

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

NOVEMBER 2016

Certificate

This is to certify that the thesis entitled "**FOREGROUND SPEECH SEGMENTATION AND ENHANCEMENT**", submitted by **K. T. Deepak** (10610204), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati.

Prof. S. R. Mahadeva Prasanna

Professor

Dept. of Electronics and Electrical Engg.

Indian Institute of Technology Guwahati

Guwahati - 781 039, Assam, India.

To

My beloved **parents**

for their endless love, support, and encouragement

&

My guide **Prof. S. R. M. Prasanna**

for his guidance and support

&

My dear wife **Shruthi**, and my lovely kids **Aditi** and **Satvik**

for the inspiration that you gave and the sacrifices you all have made during this period

Acknowledgments

I cherish all the experience I went through in the process of this graduate program, this has left me a little wiser for the day. Also, I feel I am blessed to have made it to this wonderful campus in the laps of mother nature. I owe my gratitude to all those people who have made this dissertation possible.

My deepest gratitude is to my research supervisor Prof. S. R. M. Prasanna who gave me an opportunity to carry out research work under his guidance. His patience and support helped me to overcome many crisis situations and finish this thesis work. I am in awe of his sheer dedication and commitment to the work he does. I'll feel fortunate if he has rubbed a little of these traits on me.

I am thankful to my doctoral committee members Prof. S. Dandapat, Prof. R. Sinha, Dr. Shakuntala Mahanta, and Dr. Priyankoo Sarmah for their encouragement and valuable suggestions on my work. I am especially grateful to Prof. S. Dandapat and Prof. R. Sinha for insightful comments and constructive criticisms on the work to bring it to the current form.

I especially thank my senior Dr. H. S. Jayanna who helped me to connect with my supervisor. I am grateful to all my EMST lab senior members Dr. S. R. Nirmala, Dr. P. Krishnamoorthy, Dr. Debadatta Pati, Dr. Govind, Dr. Gayadhar Pradhan, Dr. Sumitra Shukla and Dr. Haris for mentoring me beyond the profession. My thanks to Dr. L. N. Sharma for maintaining the EMST laboratory with all the necessary facilities.

My sincere gratitude to all my fellow project mates Syed Shahnawazuddin, Biswajit Dev Sarma, Aniruddha, Abhishek Dey and Siddika who have immensely contributed to building

mandi Assamese spoken query system. I also cherish those technical discussions I had with my fellow colleague Syed Shahnawazuddin. My special thanks to Abhishek Dey for helping me to build the ASR system that is used in this thesis work.

I am indebted to my friend Nagaraj Adiga for his practical advice and research discussions that helped to enrich the ideas. I am also thankful to all my friends K. K. Ramesh, Banri, Sibsankar Padhy, Anurag, Aravind and Vikram for reading my reports and comment on my views. I would like to thank my fellow colleagues Sunil, Rohan Kumar Das, Bidisha Sharma, Ramesh Kumar Bhukya, Rajib and all other project staffs for all the rich interactions I had with them.

Most importantly, none of this would have been possible without the selfless love, affection, and support of my parents. My wife Shruthi has been my biggest source of strength while carrying research and she completely insulated me from all the household responsibilities. At times my depleted energy levels were recharged by my soothing daughter Aditi and my energetic son Satvik. I am heartily thankful to God for having given me such a wonderful family.

Deepak K T

Abstract

Speech enhancement is one of the active areas of research and a challenging task when the signal is recorded in natural environments. In a typical recording scenario using a single microphone, it is safe to assume that the desired speaker is closer to the microphone sensor, relative to other interfering acoustic sources. In this work, the speech signal from close speaking person is regarded as *foreground speech* and rest of the interfering sources as *background noise*. Due to the close proximity of the desired speaker to the microphone, compared to other background sources, there are differences in the signal characteristics.

When the speech signal is recorded in natural environments, the production characteristics tend to vary depending on the levels of interfering sources. The objective of this thesis work is to exploit such unique characteristics of speech production to temporally segment foreground speech from rest of the background and further enhance it. The high signal to noise ratio (SNR) regions of foreground speech are robust to interfering noise. The high SNR region around glottal closure instants (GCIs) in the time domain and vocal tract information in the spectral domain is used to derive certain features to segment and enhance foreground speech.

This thesis proposes a new method to extract GCIs that does not need an estimation of pitch and is evaluated for its robustness under different degraded conditions. The method is used to study the nature of signal with reference to varying speaker to microphone distance and define a typical foreground speech recording scenario.

Foreground speech segmentation and enhancement methods are proposed. The methods exploit the high SNR regions in temporal and spectral domains. The temporal domain processing involves the derivation of gross and fine weight functions. Spectral processing involves enhancement of spectral peaks that represent formants and further exploit the perceptual result to minimize the effect of spectral distortion.

In a mobile based spoken query system, there is no control on recording environment from which a user can access such system. Due to other interfering sources, there can be degradation in the performance of automatic speech recognition (ASR) and thereby impact the usefulness of spoken query system to the users. This thesis demonstrates the effectiveness of foreground

speech segmentation and enhancement as a pre-processing module to the spoken query system.

The major contributions of this thesis are as follows:

- A new method is proposed to extract GCI locations and it is evaluated for its robustness under different degraded conditions.
- Typical foreground speech recording scenario is defined through experimental analysis.
- A new paradigm of speech segmentation and enhancement is proposed that utilizes the distance of the desired speaker to microphone relative to other interfering sources.
- A novel foreground speech segmentation and enhancement using speech production and perceptual features are proposed.
- New objective measures to measure the effectiveness of enhancement methods when the reference clean speech is not available.
- The effectiveness of foreground speech segmentation and enhancement is demonstrated through improved ASR performance in a spoken query system.

Keywords: Zero band filtering, glottal closure instants (GCIs), foreground speech, background noise, spoken query system, automatic speech recognition (ASR)

Contents

List of Figures

List of Tables

List of Acronyms

ASR	Automatic Speech Recognition
BAK	Background Noise
BIBO	Bounded Input Bounded Output
BKG	Background Noise
cGC	Compressive Gammachirp Filter
DNN	Deep Neural Networks
DPI	Dynamic Plosion Index
DTMF	Dual Tone Multi Frequency
DYPSA	Dynamic Programming Projected Phase-Slope Algorithm
EBE	Excitation Based Enhancement
EGG	Electroglottograph
ENR	Epoch to Non-Epochal Ratio
ERB	Equivalent Rectangular Bandwidth
FAR	False Alarm Rate
FBE	Formant Based Enhancement
FBR	Foreground to Background Ratio
FEC	Front End Clipping
FGS	Foreground Segmentation
FMLLR	Feature Space Maximum Likelihood Linear Regression
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Model
HCI	Human Computer Interaction

HE	Hilbert Envelope
HELP	Hilbert Envelope of Linear Prediction Residual
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
IDA	Identification Accuracy
IDR	Identification Rate
IIR	Infinite Impulse Response
ILPR	Integrated Linear Prediction Residual
IVR	Interactive Voice Response
IVRS	Interactive Voice Response System
LDA	Linear Discriminant Analysis
LP	Linear Prediction
LPR	Linear Prediction Residual
MCC	Mel Cepstral Co-efficients
MFCC	Mel-Frequency Cepstral Co-efficients
MLLT	Maximum Likelihood Linear Transform
MLSA	Mel Log Spectral Approximation
MMSE	Minimum Mean Square Error
MoS	Mean Opinion Score
MR	Miss Rate
MSC	Missed Speech Clipping
NACC	Normalized First Order Autocorrelation Co-efficient
NDS	Noise Detected as Speech
OVER	Noise Interpreted as Speech
OVL	Overall Quality
PBE	Perceptual Based Enhancement
PC	Personal Computer
PESQ	Perceptual Evaluation of Speech Quality

pGC	Passive Gammachirp Filter
PRI	Primary Rate Interface
SEDREAMS	Speech Event Detection using the Residual Excitation and a Mean-based Signal
SGMM	Sub-space Mixture Model
SIGMA	Singularity in Electroglottograph by Multi-scale Analysis
SNR	Signal to Noise Ratio
SoE	Strength of Excitation
SRR	Signal to Reverberation Ratio
SS	Spectral Subtraction
TSP	Temporal and Spectral Processing
VAD	Voice Activity Detection
VFR	Variable Frame Rate
WER	Word Error Rate
WLPR	Weighted Linear Prediction Residual
YAGA	Yet Another Glottal Closure Instant Algorithm
ZBF	Zero Band Filter
ZBFS	Zero Band Filtered Signal
ZFF	Zero Frequency Filter
ZFFS	Zero Frequency Filtered Signal

List of Symbols

A_k	STSA of clean speech signal
$\hat{A}_k$	STSA of estimated speech signal from noisy signal
$log\ A_k$	Logarithmic STSA of clean speech signal
$log\ \hat{A}_k$	Logarithmic STSA of estimated speech signal from noisy signal
$\beta_\alpha(\Omega)$	Phase of all-pass function
$\exp(c_1\theta_1(f))$	An asymmetric function
$\triangle c_k$	Delta cepstrum vector at k^{th} frame
$D(k)$	Background degradation in frequency domain
$\triangle E_f$	Full-Band energy difference
$\overline{E_f}$	Running average energy of the background noise
$\triangle E_l$	Low-Band energy difference
$E(N)$	Energy of noisy speech in frame N
$E_n(N)$	Energy of noise in frame N
$e[n]$	LP residual signal
$e_h[n]$	Hilbert transform of LP residual signal
$\mathrm{ERB_N}(f_{r1})$	Equivalent rectangular bandwidth of average normal hearing subjects
f_{r1}	Asymptotic frequency
$g_c(t)$	Compressive gammachirp auditory filter output
$\lvert G_T(f)\rvert$	Fourier magnitude spectrum of the gammatone filter
$G_{cp}(f)$	Passive gammachirp filter (pGC)
h_{ZF2}	Impulse response of 2^{nd} order ZFF
h_{ZF4}	Impulse response of 4^{th} order ZFF

h_{ZB2}	Impulse response of 2^{nd} order ZBF
h_{ZB4}	Impulse response of 4^{th} order ZBF
$H_1(z)$	1^{st} stage second order ZBF magnitude response
$H_2(z)$	2^{nd} stage second order ZBF magnitude response
$H_{vt}(z)$	LP filter predicted from p^{th} order LP analysis
$H^{\alpha}(z)$	All pass filter transfer function
$h_e[n]$	Hilbert envelope of LP residual signal
$L(z)$	Linear prediction residual in z-domain
$\hat{L}$	MMSE-LSA estimator
$\hat{M}$	MMSE-STSA estimator
$\mid N \mid^2$	Estimated noise only power spectrum
ϕ_1	Initial phase
r	Value of the radius on unit circle in z-plane
$R_w(z)$	Weighted linear prediction residual in z-domain
$SNR_p(N)$	*A posteriori* SNR in frame N
$S(k)$	Desired speaker's speech in frequency domain
$\hat{s}(n)$	Estimation of the speech signal
$\hat{s}_p(n)$	Normalized envelope of *pth* filter output
$S_t(z)$	Temporally enhanced speech signal
$\hat{\sigma}_f^2$	Foreground speech power estimation
$\hat{\sigma}_b^2$	Background power estimation
$T(\omega)$	Noise masking threshold
$\tau(\omega)$	Group delay of a given signal
$\theta(\omega)$	Continuous phase spectrum
$w_g(n)$	Sigmoidal function
$w_f(n)$	Fine weight function
$w(n)$	Final weight function
$\mid \hat{X}_k \mid^2$	Estimated clean speech signal power spectrum of the k^{th} frame

$\mid X_k \mid^2$	Noisy speech signal power spectrum
$\tilde{X}_{k,i}$	Estimated spectral envelope
$X(k)$	Degraded speech signal in frequency domain
$x[n]$	1^{st} order difference signal computed from signal $s[n]$
$y[n]$	Zero frequency filtered output
$y_1(n)$	1^{st} stage ZBF output signal
$y_2(n)$	2^{nd} stage ZBF output signal

1

Introduction

Contents

Speech is an important mode of human to human communication. Seldom it is possible to record speech signal in a clean recording environment without any interfering sources. Often in most recording scenarios, the speech signal is degraded by other background acoustic sources. Listening to such degraded speech signal for a long time can be an unpleasant experience. It is, therefore, necessary to process the speech signal to identify certain regions and enhance to make it more suitable for listening and other speech-based applications.

1.1 Genesis of Proposed Work

In the current era of wireless speech communication, there is no restriction on the environment from which the speakers can access speech processing systems like speech recognition and speaker recognition. Due to this, the sensor may capture interfering noise from various other sources along with desired speaker's speech. Humans have remarkable ability to selectively listen to the speech of interest in spite of interfering noise present. Listeners feel uncomfortable to listen to such degraded speech signals for a long time. Also, depending on the levels of degradation present, it is still a major challenge to handle such speech signals using computing machines. The presence of interfering noise can degrade the performance of speech and speaker recognition systems. Hence, there is a need for enhancement of such degraded speech signal to make it suitable for human listening and also to improve the performance of speech and speaker recognition systems.

The speech signal can be recorded in a controlled environment like anechoic chambers, where there will be no interference of other sources. In such scenarios, the speech-based applications can be realized by directly extracting features as shown in Figure 1.1. However, in a typical scenario, the recording environment is uncontrolled and can have other interfering sources that degrade desired speaker's speech as shown in Figure 1.2. A pre-processing block is necessary in order to identify desired speaker's speech from rest of the background content and this usually involves voice activity detection. There are different methods available in the literature that can be categorized under pre-processing, while voice activity detection method is considered to

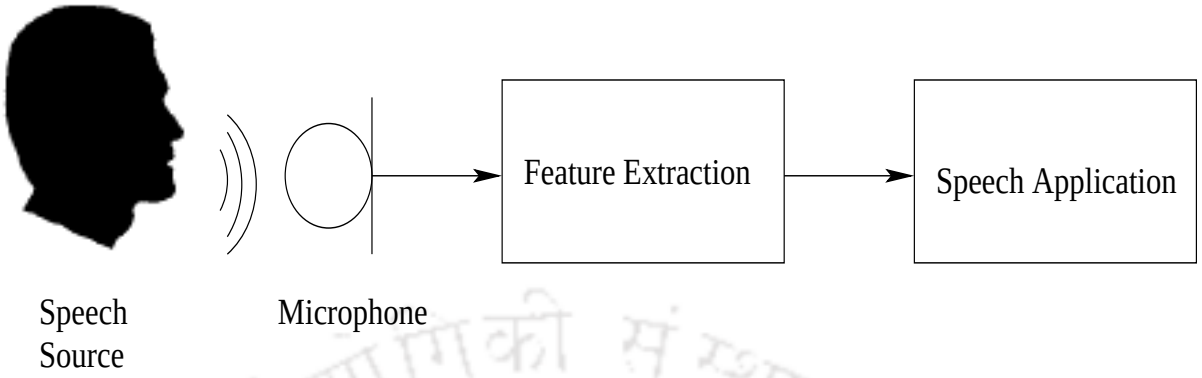

Figure 1.1: Different blocks involved in a typical speech based application when speech signal is recorded in controlled environment.

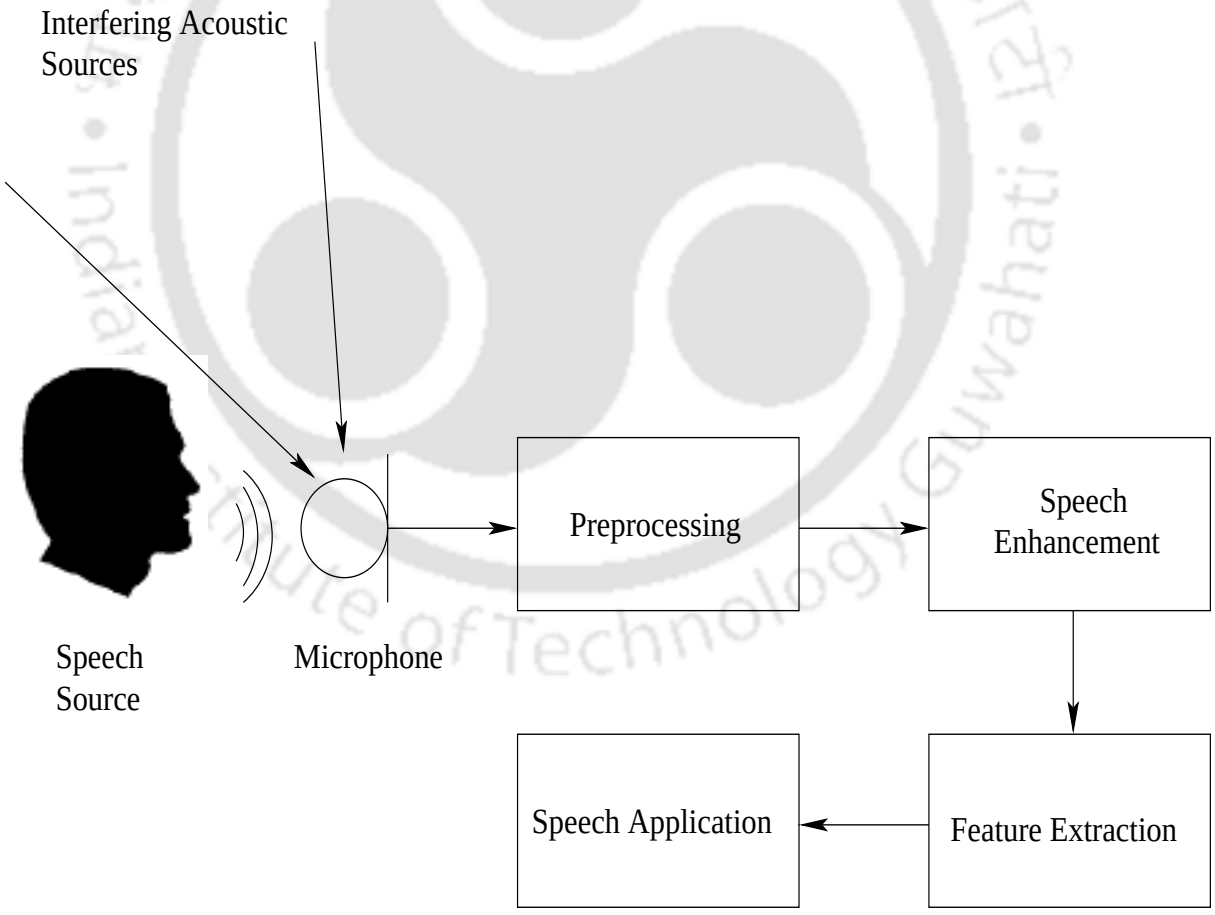

Figure 1.2: Different blocks involved in an speech based application when speech signal is recorded in uncontrolled environment.

be a pre-processing module in this work. Furthermore, the segmented speech regions need to be enhanced using speech enhancement module as shown in Figure 1.2.

There are many methods available in the literature that enhance degraded speech based on spectral characteristics of interfering noise [2–6]. The ability of such enhancement methods depends on the ability to segment interfering noise from the desired speech and further model the noise in the spectral domain. However, such methods make a certain assumption of properties from interfering noise and hence they are limited by the ability to deal with different types of interfering noises. Further, such enhancement methods work by spectral modification of desired speaker's speech in the process. One of the limitations of such methods is that the speech signal is assumed to be stationary within block size of 10 - 20 ms and such analysis of speech signal is termed as segmental. In practice, this stationary assumption may not be valid and the characteristics of speech signal vary within each production cycle [7].

The interfering noise does not affect speech signal equally in all regions. There are regions of speech signal that are robust to interfering noise and such regions have high Signal-to-Noise-Ratio (SNR)/ Signal-to-Reverberation-Ratio (SRR). Hence, analyzing signal at such high SNR/ SRR regions can help us derive the features that are robust to interfering noise. During speech production, the vocal tract system is excited by glottal pulses. The glottal pulses are obtained by the manipulation of air flow by vocal folds. There is a significant change in the pressure when air flows through vocal cords due to vocal folds coming in contact. The abrupt closing of vocal folds gives rise to an impulse like excitations that gets modulated by vocal tract system. Typically the regions around such excitations are high SNR regions and are robust to interfering noise. Also, there is some evidence that human listeners exploit spectro-temporal characteristics of the speech signal in certain regions for improved perception of the signal. In particular, it is hypothesized that such regions of the speech signal are least affected by degradations, compared to other regions [8].

The ability of speech perception mechanism of human beings to selectively focus on certain regions of speech signal enables them to derive information from those regions and fill the information in other regions [9]. Thus, there is a basis, both in speech production and perception,

that all regions of the speech signal may not be equally important for enhancing the degraded speech signal. Typically such high SNR/ SRR regions are few samples around the instants of significant excitation.

Some of the earlier works have explored in the direction of using robust speech production based features that are less susceptible to background noise [10, 11] for enhancement. The work proposed in [10] mainly involves temporal processing of degraded speech by additive random noise. However, the additive noise does not affect all the speech regions to the same extent. Therefore regions around instants of significant excitation are used as reference points to enhance the regions. The study is further extended to single and multi-channel speech enhancement case as explained in [11] by using temporal features derived using instants of significant excitation events.

In [12], a single channel speech enhancement is dealt using a combination of temporal and spectral processing. The temporal processing involves identification and enhancement of high SNR regions of the speech signal in temporal domain and subsequently, the spectral processing involves subtraction of residual noise left over. The method helps to significantly reduce musical tone noise, which is introduced due to spectral subtraction based methods. The focus of proposed work is to segment the desired speech which is spoken from a person closer to the microphone and further enhance it. The work emphasizes on speech signal recorded in practical environments, where, the desired speaker is closer to microphone sensor. New methods are explored to first segment the desired speech from rest of the interfering sources and subsequently enhance the desired speaker's speech.

1.2 Objective of the Thesis

In this work, the desired speaker's speech who is speaking close to microphone is termed as foreground speech and rest of the interfering content is categorized as background noise. Typically a person speaking through head mounted microphone or through a mobile phone in normal mode refers to close speaking scenarios. The normal mode in current work refers to

Figure 1.3: Close Speaking Scenario using Headphone.

headphone mounted on the head of desired speaker or mobile handset held close to ears while speaking as shown in Figures 1.3 and 1.4, respectively. In such scenarios, the background noise picked up by the sensor can be of comparable amplitude levels to foreground speech. Also, the background noise can have similar spectral characteristics to foreground speech when people are speaking in the background. Consequently, segmenting foreground speech from rest of the background noise that can have similar characteristics is a challenging task.

It is desirable to explore the features of foreground speech that are least affected by background noise and use them for segmentation and enhancement. The epochs are instants of significant excitations and usually, occur at glottal closure events of speech production. It is observed that such instants are least affected by background noise. The present work in this thesis demonstrates the significance of epoch based features for foreground speech segmentation and further use them to enhance the foreground regions. The work proposed in this thesis is therefore termed as **Foreground Speech Segmentation and Enhancement.**

1.2.1 Nature of the Foreground Speech Signal

The Figure 1.5 illustrates a speech signal recorded using a headphone microphone connected to a laptop in an office environment when flooring work is carried. It can be observed that there is a significant amount of noise present in the recording. Further, it can be noticed that the noise is spread in the spectral domain through all bands of frequencies and they overlap with

Figure 1.4: Close Speaking Scenario using Mobile Phone.

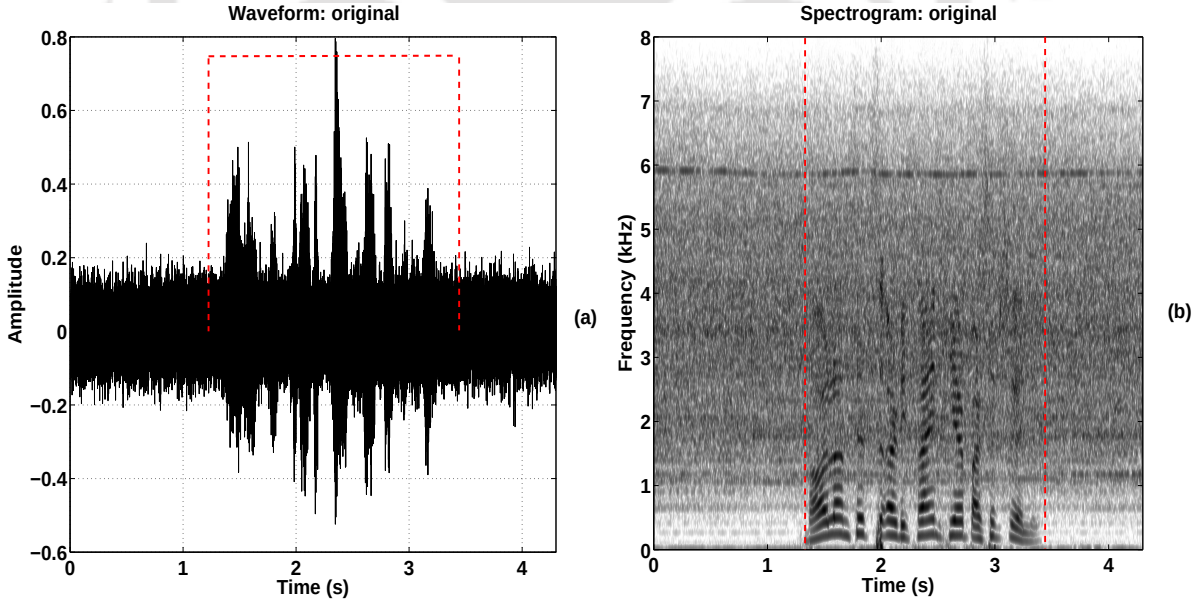

Figure 1.5: Illustration of naturally recorded speech file from a male speaker with background noise (a) time domain waveform (b) spectrogram representation of the waveform shown in (a). Segment within dotted lines shows Foreground Speech.

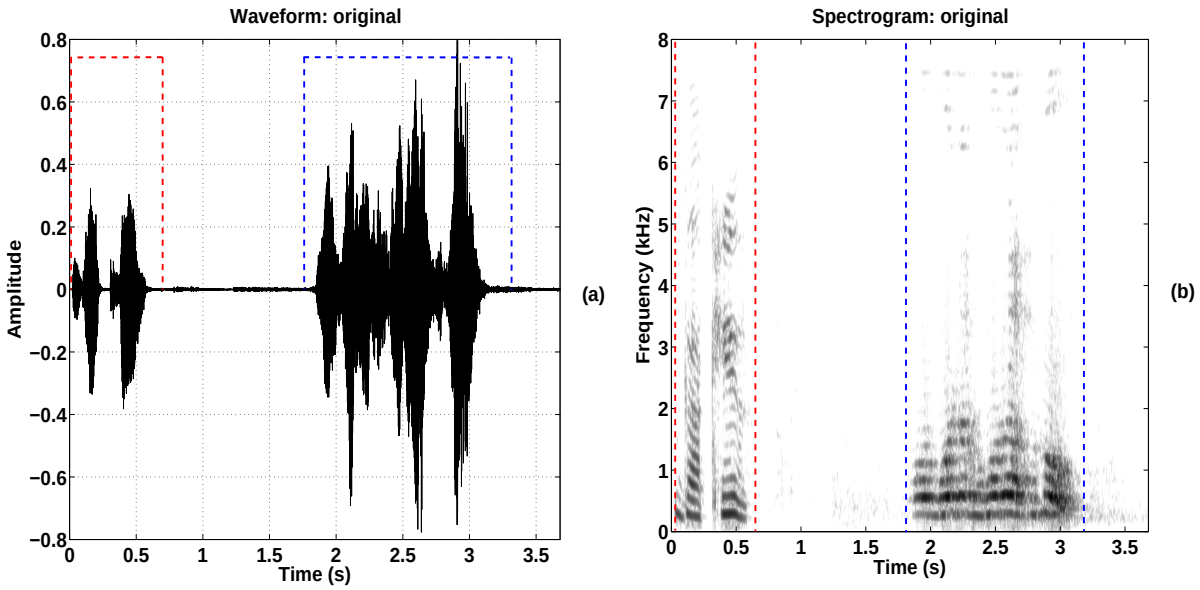

Figure 1.6: Illustration of naturally recorded speech file from a female speaker with background noise (a) time domain waveform (b) spectrogram representation of the waveform shown in (a). Segment within dotted lines shows Foreground Speech.

foreground speech. However, a normal human listener may not find it difficult to understand the information as the intelligibility is still preserved in the signal. It remains a challenging task for computing machines to segment foreground speech in the presence of such background noise. The structure of foreground speech may not be evident from the time domain signal due to the poor time resolution of the plot as shown in Figure 1.5(a). However, the harmonic structure is clearly visible in its spectrogram plots shown in Figure 1.5(b).

The Figure 1.6 shows the time and spectral domain plots of a female speaker's speech file recorded through a telephone channel using voice server. The recording contains another person's voice at the background talking and it can be noticed from Figure 1.6(a) that the amplitude levels of both foreground and background regions are comparable. It can be observed that the background noise is non-overlapping with the foreground speech in this case. Since the background noise is the voice of another person and it can be noticed from Figure 1.6(b) that both foreground and background regions are similar in their spectral characteristics. Such cases of background noise can be challenging to separate by considering all the frequency components for processing. From the Figures 1.5 and 1.6, it can be noticed that the background noise can

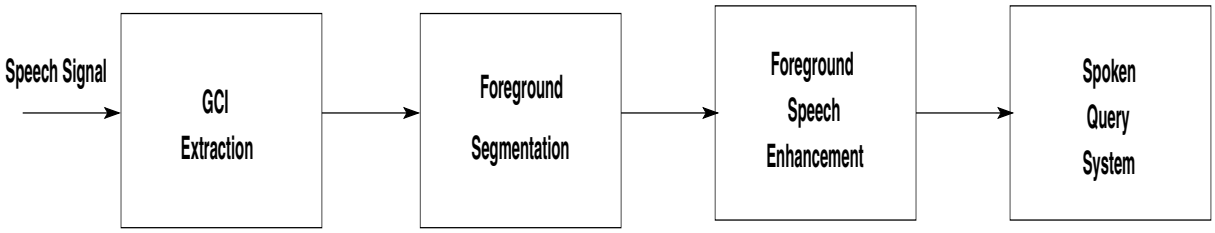

Figure 1.7: Overall Block Diagram of the proposed work

overlap with foreground speech or they can be temporally separated from foreground speech regions. Hence, the challenge lies in developing common methods which rely on foreground speech characteristics rather than characterizing many types of background noises that can interfere with the recordings.

1.2.2 Foreground Speech Processing

The Figure 1.7 shows the overall block diagram of the proposed work in this thesis. The work mainly consists of a new method to extract instants of significant excitation which predominantly consists of glottal closure instants (GCIs) from speech signal recorded in natural environments. The GCIs are robust to interfering sources and they form high SNR regions. These regions are used to derive certain features that help to distinguish foreground speech from rest of the background noise. This is taken care in foreground speech segmentation block as shown in the Figure. Further, the foreground speech is enhanced using production and perception based features. The enhanced foreground speech signal is mainly suitable for comfortable listening by humans and further used in speech based applications. In this work the usage of foreground segmentation and enhancement is demonstrated through the improved performance of spoken query system.

1.2.3 Approaches for Foreground Segmentation and Enhancement

The present work focuses on the practical scenario when the speaker speaks close to the microphone mounted to headphone or over a mobile phone. In such scenarios, there can be other interfering sources present along with desired speech. In-spite of such interfering sources,

the instants of significant excitations from the close speaking scenario is preserved and few samples around such regions are least affected. There is, therefore, a need to derive certain features using such nature of foreground speech to segment it and then enhance the same. Similarly, there are regions in spectral domain that are robust to interfering noise and they can be exploited to segment and enhance the foreground speech. The temporal segmentation of foreground speech signal should enable us to try and explore the robust aspects of production and perceptual cues to enhance the foreground speech.

1.3 Organization of the Thesis

The contents of the thesis are organized as follows:

In **Chapter 2**, review of the literature on voice activity detection and speech enhancement methods are presented. The review of existing methods motivated us to see the speech enhancement from a different direction and the motivation are presented in this chapter.

Chapter 3 discusses a brief review on epoch extraction and explains the existing zero frequency filtering method. Based on some of the drawbacks existing in zero frequency filter a motivation for a stable realization of zero frequency filter called as zero band filtering is suggested. The proposed method is evaluated under clean and noisy conditions. Also, the robustness of the proposed work is demonstrated by varying pitch and foreground scenarios.

Chapter 4 illustrates the usefulness zero band filter in the analysis of speech signal recorded at varying distances. Based on such analysis, a definition of *foreground speech* is established. Short term analysis using temporal and spectral methods further corroborates the definition of foreground speech. The excitation source and vocal tract information are unique to foreground speech and they are utilized to derive temporal features. These features help to distinguish foreground speech from rest of the background regions. The proposed foreground speech segmentation is compared with considered state-of-the-art methods for its performance.

In **Chapter 5** foreground speech segmentation is used to temporally emphasize foreground speech regions relative to other background regions. The epoch extraction using zero band filtering enables the 1^{st} stage of enhancement called as excitation based enhancement. The

excitation based enhancement alone may not be sufficient to completely eliminate the background noise and hence a novel formant peak enhancement is proposed. The perceptual based enhancement is proposed using Mel cepstral coefficients and Mel log spectral approximation filter. All 3 different stages of enhancement are evaluated in terms of subjective and objective measures. Two new objective measures are proposed to evaluate the effectiveness of foreground speech enhancement on degraded speech. The work is evaluated along with 3 other considered state-of-the-art methods for its performance.

Chapter 6 explains different modules involved for the realization of a practical Assamese spoken query system briefly. The performance of mobile based spoken query system is affected by interfering background noise. The foreground speech segmentation and enhancement is used as front end pre-processing module to attenuate background noise and further enhance foreground speech regions. The epoch extraction in telephone channel signal is not reliable using zero band filtering, therefore an alternative scheme is proposed using Hilbert envelope of linear prediction residual. The effectiveness of using foreground speech segmentation and enhancement is demonstrated through improved performance of spoken query system.

A summary of the present work is given in **Chapter 7** by listing major contributions of the present work and some directions for future research in the area of foreground speech segmentation and enhancement.

2

Processing of Degraded Speech - A Review

Contents

This chapter reviews the pre-processing methods and enhancement techniques of degraded speech signal available in the literature. In this thesis, depending on the context, speech enhancement can also be included as part of the pre-processing module. For example, in the case of speech based application, both voice activity detection (VAD) and speech enhancement can be combined together as a pre-processing module. No other interfering sources will be present when speech is recorded in a controlled environment like anechoic chamber and such a signal can be termed as clean speech. However, more often speech data is collected using microphone sensors from natural environments. Most of the environments in which the speech gets recorded are seldom under the user's control. Therefore, the speech that gets recorded can have different kinds of degradation due to the presence of other interfering sources. Such degraded speech can lead to adverse effects on the perceptual quality listened by humans and further reduce the intelligibility of spoken words. Also, the degradation of recorded speech signal can adversely impact the performance of speech recognition, speaker verification, or a spoken query system.

It is necessary to identify the relevant speech regions and further enhance them. However, the interfering sources can have similar characteristics like the desired speech signal. It is a challenge to temporally isolate the desired speaker's speech, segment it from rest of the interfering noise and further enhance it. Conventionally such tasks are handled using two important modules viz., front end pre-processing module and enhancement module. Though there are several pre-processing steps applied to degraded speech, by far VAD is one of the most commonly applied pre-processing modules in the literature. Broadly, VADs can be classified as supervised and unsupervised based methods. The VAD module alone may not be sufficient to improve the perceptual quality of degraded speech. Hence there is a need for enhancement of degraded speech and this problem has been attempted by the scientific community in different ways. Broadly, the speech enhancement methods can be classified as temporal and spectral processing in literature. This chapter mainly attempts to present a review on different VADs and speech enhancement techniques available in the literature.

2.1 Voice Activity Detection and Speech Enhancement

The speech segmentation and enhancement is of significant interest to research community due to its implications on a wide variety of applications. Hence the problem has received considerable attention from scientific community from early days. The common types of degradation to speech signal include interfering background noise, reverberation, and competing speakers at the background along with desired speaker's speech. The interfering background noise is one of the most commonly occurring degradation in the speech signal. The main focus of this thesis work is to address the solutions for interfering background sources and hence this chapter mainly reports the methods that are pertaining to background noise.

In order to improve the quality of such degraded speech signal, usually front end pre-processing module appears before speech enhancement module. There are different types of pre-processing modules available in the literature depending on the type of applications. For example, gender detection, speaker diarization, and voice activity detection [13–15] can be categorized as pre-processing of degraded speech. However, voice activity detection (VAD) is one of the widely used pre-processing modules and has several applications and requirements in speech processing area. The VAD is also called as speech/non-speech detection in the available literature. The non-speech regions include silence and noise-only regions of signal recorded, while speech regions may include speech only or speech corrupted by background noise.

The VAD essentially consists of two main steps to identify the relevant speech regions from the signal and later demarcate the boundaries of speech regions. The first step is used for feature extraction that mainly derives temporal or spectral based features which help to discriminate speech and non-speech regions and this is used for boundary decision making. However, the decision-making process can vary from simple threshold-based methods to machine learning algorithms. Therefore VADs proposed in the literature can be broadly classified as supervised and unsupervised techniques. The unsupervised methods have the advantage that these methods work with little or no training data at all [4, 16–22]. The earliest version of VAD relied on short-term energy and zero crossing features [23]. However, these features are not reliable at

low signal-to-noise-ratio (SNR) levels and such VADs fail to accomplish the task. Alternatively, the machine learning decisions are made based on large resources of labeled training data. The supervised methods require training data with labeled information, while such methods are more reliable depending on the conditions in which the training data is collected [24–28].

The supervised techniques make certain assumptions about the statistical properties of both speech and background noise and this can be an issue while handling unseen background noise. The background noise can be from different interfering sources and the noise characteristics may vary depending on the type of background source. The type of background source may vary depending on the type of application and recording environment. It is not always feasible to take all types of background sources into account in a given practical application. It is, therefore, necessary to explore methods that rely on known characteristics of speech signal rather than modeling noise. There are few methods available in the literature that makes no assumption of noise characteristics, rather they rely on known speech production based characteristics [29–33].

The interfering background noise can be separated from desired speech signal if it is temporally non-overlapping. In such scenarios, a robust VAD is sufficient to handle the case. However, the background noise can temporally overlap with desired speech signal and therefore VAD alone is not sufficient. The speech enhancement module in tandem with VAD can be an effective solution and hence it becomes imperative to study important directions traversed by speech enhancement techniques in the literature.

Most of the speech enhancement methods developed in the literature deal with additive background noise along with the desired speaker's speech. The additive background noise has been addressed by several methods starting from spectral subtraction [2] based methods. Although several other variants of this method were proposed in the literature, by far spectral subtraction method is still referred as one of the important techniques to handle additive background noise because of it's computational advantage and theoretical simplicity. All such methods including wiener filter [4,34,35] based methods mostly rely on magnitude compensation of noise components. The phase obtained from the degraded speech is retained without any

modification. It is considered that the magnitude compensation is perceptually more relevant for speech enhancement than the phase information [36].

An improvement was suggested in [4] over spectral subtraction, a wiener filter based approach is proposed using the statistical modeling of speech and noise signal in the minimum mean square error (MMSE) sense, where both speech and noise were modeled using probability density functions. The spectral subtraction based methods rely on magnitude subtraction of noise signal from the degraded speech signal. This leads to introduce spurious non-existent peaks in the magnitude spectrum. Such spurious peaks introduced in magnitude spectrum leads to undesired audible tones in the reconstructed time domain signal and this is termed as musical noise. The musical noise is more annoying for human listeners than low magnitude broadband noise [36]. It is desirable to have speech enhancement methods which do not introduce unwanted distortions.

There are few methods in the literature that takes human auditory perception into account for speech enhancement [37–39]. These methods utilize psychoacoustic phenomena called auditory masking. It is known that below a certain threshold of dominant tone within the critical band in the frequency domain, no other tones are audible to human listeners. This property is used to control the parameters of spectral subtraction to reduce the effects of musical noise and increase the intelligibility of enhanced speech [37]. Similarly, the temporal masking exists in time domain in which the two stimuli occur within a small interval of time, in which only dominant tone is audible. Forward temporal masking occurs when a masker precedes the signal (or maskee) in time and maskee is not audible by human listeners. This property is exploited to improve the enhancement quality of degraded speech in terms of auditory perception [38,39]. However, utilizing auditory phenomena like psychoacoustics for the sake of speech enhancement can render such systems to be more complex. Alternatively, few methods are explored in the literature which uses a well-understood speech production knowledge to enhance the degraded speech [10–12].

It is evident that interfering noise sources do not affect the speech signal equally in all regions [36]. There are certain regions of speech signal both in the temporal and spectral

domain that are robust to background noise. For example, glottal closure instants (GCIs) in time domain and format peak locations in spectral domain are high SNR regions in the speech signal that are relatively less prone to interfering background noise. Also, there is some evidence that humans perceive message from the speech signal by extrapolating the information available in high SNR regions to fill the gaps in low SNR regions [9]. Hence, such high SNR locations can be used as anchor points to enhance speech signal [8]. The methods explored in this direction causes the least distortion to synthesized speech [30–33, 40–45].

2.2 Features for Voice Activity Detection

Voice activity detection is extensively used in several applications including speech enhancement techniques [46–49]. VAD is used in real time applications as front-end pre-processing module for speech coders [21, 50]. Depending on the application, the complexity of task for VAD varies. For example, in the case of clean isolated word utterances, the job of VAD is to identify the end points of beginning and the end of utterance [51]. However, due to degradation of the speech signal from other interfering sources, the task of identifying relevant speech regions from non-speech regions becomes more complex. Mainly as stated in Section 2.1, VAD consists of two stages *viz.,* feature extraction and speech non-speech detection.

2.2.1 Feature Extraction

The desired properties of feature extraction include discriminative ability to identify the speech regions from rest of the non-speech regions that may include silence. Also, the extracted features have to be robust to interfering noise. Broadly the derived features for VAD in the literature can be classified as spectral, cepstral, and temporal based features.

2.2.1.1 Spectral based Features

The frequency characteristics of the speech signal are different compared to other background sources. It is, therefore, important to analyze the degraded speech signal in the frequency domain in order to extract certain features that can help to discriminate speech and

non-speech regions of the signal. The discrete Fourier transform (DFT) is one of the most important tools available for frequency analysis [52]. Though the spectral subtraction based methods are mainly used for speech enhancement purpose, however, it can be treated as a preliminary step for voice activity detection [2]. The power spectrum of noisy speech signal is subtracted from noise only components and it is given by the following relationship

$$| \hat{X}_k |^2 = | X_k |^2 - | \hat{N} |^2 \tag{2.1}$$

where, $| \hat{X}_k |^2$ is estimated clean speech signal power spectrum of the k^{th} frame, $| X_k |^2$ is noisy speech signal power spectrum, and $| N |^2$ is the estimated noise only power spectrum. The noise only power spectrum is estimated by considering average values obtained from multiple frames of noise frames in the given signal. It is assumed that speech signal is degraded by additive noise components and they are independent of speech in the power spectrum domain.

The speech signal is dynamic in nature and is known at a normal speaking rate, 5-6 phones are spoken per second [53]. On the other hand in most of the cases, nature of background noise remains stationary for a longer duration relative to the speech signal. Psychophysical study in [54] showed that the temporal information from both short windows (20−30) ms and long windows (150−250) ms are important to understanding spoken language. Methods were explored in this regard to exploiting long term features to discriminate speech from non-speech regions.

The speech and non-speech detection algorithm in [20] assumes that the discriminating feature lies in the time-varying signal spectrum magnitude. In order to utilize this property the features are extracted using long-time window in the form of multiple frames. The long-term upper bounding envelope of the spectrum over a set of $(2M + 1)$ contiguous frames, and then calculate the sum of all L sub-band SNRs, where M is the number of neighboring frames to include in the envelope estimation, and L is the number of DFT coefficients. Since it uses a long-term speech window instead of instantaneous speech frame in order to derive the spectrum envelope and hence called as long-term spectral envelope (LTSE). The speech and non-speech decision is taken based on long-term spectral divergence (LSTD) between noisy speech and

noise components. The LTSE and LSTD are computed using the following relationships

$$\tilde{X}_{k,i} = max\{X_{k-M,i}, ..., X_{k,i}, ..., X_{k+M,i}\} \tag{2.2}$$

$$LSTD(X_k) = 10log_{10}\left(\frac{1}{L}\sum_{i=1}^{L}\frac{\mid \tilde{X}_{k,i} \mid^2}{\mid \tilde{N}_i \mid^2}\right) \tag{2.3}$$

where $\tilde{X}_{k,i}$ is the estimated spectral envelope of the neighboring $(2M+1)$ frames and N_i is the i^{th} co-efficient of the average noise spectrum. The performance of this method is evaluated and compared with other considered state-of-the-art methods. It is shown that such long-term features work better in discriminating speech and non-speech regions in low SNR values.

The long-term dynamic features were derived by fusing delta cepstrum from the neighboring frames for VAD application [55]. However, an extensive study was carried in [56] using the sample variance of the long-term subband entropies, showed that the performance is significantly better compared to other methods when speech is corrupted by different types of noise.

The methods explained so far mainly utilize magnitude spectrum and do not consider phase spectrum to derive features to distinguish speech and non-speech regions. However, there is a considerable number of works that utilize phase spectrum for different speech applications [57–60]. The group delay function can be derived from Fourier transform represented in polar form as shown below

$$X(\omega) = \mid X(\omega) \mid e^{j\theta(\omega)} \tag{2.4}$$

The group delay $\tau(\omega)$ of a given signal is defined as the negative derivative of the continuous phase spectrum $\theta(\omega)$ given by

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega} \tag{2.5}$$

From Eqn. (2.5), the group delay function can be computed from the signal using the following relationship

$$\tau(\omega) = \frac{X_R(\omega)\hat{X}_R(\omega) + X_I(\omega)\hat{X}_I(\omega)}{\mid X(\omega) \mid^2} \tag{2.6}$$

It is studied in [16] that group delay function can be effectively used for speech and non-speech detection.

2.2.1.2 Cepstral based Features

The fundamental frequency F_0 represents the pitch period of the speaker in the speech signal. The F_0 is within a certain range and mostly such frequencies are unique to human speakers. Therefore the features derived using F_0 can be effectively used for distinguishing speech and non-speech regions. The cepstrum is one of the important ways of extracting fundamental frequency from the speech signal and such a feature is robust to most of the interfering noise. The cepstrum can be derived using power spectrum of the log-power spectrum and this can be used to analyze the power spectrum of the signal. The pitch of a speaker can be estimated from cepstrum given by the following relationship

$$c_k = \mid DFT(log \mid X_k \mid^2) \mid^2 \tag{2.7}$$

where c_k is cepstral co-efficients derived from k^{th} frame, $\mid X_k \mid^2$ is estimated signal power spectrum of the k^{th} frame and DFT is discrete Fourier transform.

The fundamental period derived from cepstral coefficients can be used to detect speech and non-speech regions [52,61]. The mel-frequency cepstral co-efficients (MFCC) features derived in mel-scale is one of the extensively used cepstral based feature for speech and speaker recognition systems [62]. However, the MFCC features are also used in the literature for speech and non-speech detection. In order to capture the dynamic nature of speech signal, the first and second order derivative features are derived from MFCC and they are called delta and delta-delta coefficients. The delta coefficients are defined using the following relationship

$$\triangle c_k = \sum_{i=1}^{M} k(c_{k+i} - c_{k-i}) / \left(2 \sum_{i=1}^{M} i^2 \right) \tag{2.8}$$

where a delta window of length $(2M+1)$ frames is used to extract the delta cepstrum vector $\triangle c_k$ at k^{th} frame. In [63] the MFCC feature vectors along with delta and delta-delta co-efficients are used for voice activity detection. Similarly, in [64] the speech and non-speech detection is carried using only delta co-efficients.

2.2.1.3 Temporal based Features

The temporal-based features find wide usage in most practical applications due to its suitability for real-time applications. Signal energy and zero crossing rate (ZCR) are some of the widely used features in temporal domain [52]. The basic assumption is that the desired speech regions have higher energy relative to interfering background noise. Further such features are derived using frame-based approaches, where it is assumed that speech signal is stationary within 10-30 ms. However, neither of these assumptions hold good as those features may fail at low SNR conditions. Due to the low computational complexity and being unsupervised techniques, the temporal based methods are still preferred. By far, G.729 International Telecommunication Union (ITU) VAD standard developed for fixed telephony and multimedia communications is used as a baseline system to benchmark the performance of other methods.

G.729 ITU VAD The G.729 is ITU defined standard for audio compression to encode speech signal with lower bit rate. The coding algorithm behind this standard is known as Conjugate Structure - Algebraic Code Excited Linear Prediction (CS-ACELP) at the fixed rate of 8 kb/s [21]. The G.729 is mainly used in Voice over Internet Protocol (VoIP) based applications. For example, in the case of the conference call, it is used to conserve the bandwidth of transmission. There are variations proposed to the basic version of G.729 called as G729A, G729B, and G729AB. However, the working principle remains similar in all those variants. The VAD plays a major role in order to compress the speech signal in G.729 standard. The algorithm works on frame-based approach. The decision on the status of each frame being speech or non-speech is taken based on instantaneous features derived from the current frame. Subsequently, the final decision is taken depending on the relationship with previous frames. The 4 instantaneous parameters used in G.729 VAD are line spectral frequencies, full-band energy, low-band energy, and zero-crossing rate. Two different encoders *viz.,* inactive voice and active encoders are connected in parallel. For a given speech frame either one of the encoders is active. An encoder being active depends on speech or non-speech decision of a frame. The basic set of parameters is the set of autocorrelation coefficients, which is denoted by $\{R(i)\}_{i=0}^{q}$.

The VAD features are obtained by computing the difference between instantaneous parameters and background noise parameters given by

Spectral Distortion $\triangle S$: A 11^{th} order linear prediction coefficients (LPCs) are computed using auto correlation method. The LPCs are converted to a set of line spectral frequencies (LSFs) given by $\left\{LSF_i\right\}_{i=1}^{p}$, where, $p = 10$ along with second reflection coefficient. The spectral distortion measure is computed as the summation of squares of the difference between the current frame LSFs vector and running averages of the background noise LSFs vector given as

$$\triangle S = \sum_{i=1}^{p} (LSF_i - \overline{LSF_i})^2 \tag{2.9}$$

Full-Band Energy Difference $\triangle E_f$: The full-band energy E_f is the logarithm of the normalized first auto-correlation coefficient $R(0)$ given by $10 \log_{10} \frac{1}{R(0)}$. The full-band energy difference is computed by taking the difference between current frame energy E_f and running average energy of the background noise $\overline{E_f}$ given by

$$\triangle E_f = \overline{E_f} - E_f \tag{2.10}$$

Low-Band Energy Difference $\triangle E_l$: The low-band energy is measured by passing the speech signal through a finite impulse response (FIR) low-pass filter having a cut-off frequency of 1 kHz. The low-band energy is computed using the relationship $10 \log_{10} \frac{1}{N} h^T R h$, where h is the 13 tap FIR low-pass filter impulse response and R is the Toeplitz auto-correlation matrix of size 13×13. The low-band energy difference parameter is computed by the difference between current frame low-band energy E_l and the running average of low-band energy of the background noise $\overline{E_f}$ given by

$$\triangle E_l = \overline{E_l} - E_l \tag{2.11}$$

Zero-Crossing Difference $\triangle ZC$: The normalized zero-crossing rate is computed from speech signal. The zero-crossing difference measure is computed using the difference between instantaneous zero-crossing of current frame ZC and running average of low-band energy of the background noise $\overline{ZC}$

$$\triangle ZC = \overline{ZC} - ZC \tag{2.12}$$

Decision Smoothing: The difference parameters are feature vector that lie in four dimensional space. The decision obtained form such feature is smoothed by taking previous frames status. Smoothing of the VAD features help in reliable VAD decision to mark the boundaries of speech and non-speech components of signal.

Variable Frame Rate Analysis The fixed frame rate analysis is based on the assumption of short term stationarity within (20-30) ms. Neither speech signal nor background noise is stationary, hence a variable frame rate (VFR) approaches are used for speech and non-speech analysis [65–67]. For example, the temporal durations of bursts, consonants, and vowels are not uniform and they vary in length. Hence, the fixed frame length analysis may not be suitable for handling the dynamic nature of speech signal. Based on VFR analysis a low-complexity voice activity detector is proposed in [22]. The frame rate is adapted based on *a posteriori* SNR weighted energy. The energy is computed in time domain in order to reduce the complexity. The frame is either retained or discarded based on accumulative *a posteriori* SNR weighted energy distance. The *a posteriori* SNR is defined as the logarithmic ratio of the energy of noisy speech to the energy of noise and it is computed using the following relationship

$$SNR_p(N) = log\frac{E(N)}{E_n(N)} \tag{2.13}$$

where $SNR_p(N)$ is *a posteriori* SNR in frame N, $E(N)$ is the energy of noisy speech in frame N, and $E_n(N)$ is the energy of noise in frame N. The *a posteriori* SNR weighted energy distance between two consecutive frames can be computed as

$$D(N) =\mid logE(N) - logE(N-1) \mid \cdot SNR_p(N) \tag{2.14}$$

where $D(N)$ *a posteriori* SNR weighted energy distance between two consecutive frames, $E(N)$ is the energy in frame N, and $E(N-1)$ is the energy in previous frame $N-1$. The decision to retain or discard the frame is taken based on the threshold derived using accumulative distance computed from weighted energy distance given in Eqn. (2.14). The method is evaluated under degraded conditions by adding different types of noises to clean speech in the range of 20 dB

to -5 dB. It is found that the performance of VFR based method is superior compared to other fixed frame rate approaches.

The time domain VADs have low complexity in terms of theory and computation. Due to such advantages, these VADs are still extensively used is modern day applications and especially they are relevant in the context of power scarce handheld devices. Since the time domain, VADs require low computational complexity and hence they are suitable for real-time applications.

Speech Production based Features The speech production is full of time varying events in the form of source excitation that gets modified by vocal tract articulatory movements. Hence, the speech signal is dynamic in nature and the assumption of stationarity within a frame of $(10 - 30)$ ms is not a valid assumption [52]. The excitation source information mainly in the form of the train of impulses is due to vocal fold vibration. This leads to instants of significant excitation in the speech signal and mostly the region around such instants form to be high amplitude locations. Therefore such instants are robust to interfering noise and they happen to be high SNR regions. The glottal closure instants (GCIs) are the predominant form of instants of significant excitation.

Linear prediction analysis is one of the well-known methods to deconvolve source excitation from vocal tract information. The linear prediction (LP) residual obtained from inverse filtering the speech signal through LP filter is used to derive certain higher order statistics features for speech and non-speech detection in [68,69]. Further, it is shown in [31] that different attributes of source information features derived from zero frequency filtered signal (ZFFS) [70] and integrated linear prediction residual (ILPR) [71] are effectively used for glottal activity detection purpose [31]. The zero frequency filter (ZFF) is essentially a 4^{th} order resonator filter at 0 Hz that helps to get rid of all other frequency components except for the signal components near 0 Hz. Hence, it is robust to interfering noise and it is shown in [30] that ZFFS energy can be effectively used for robust glottal activity detection.

2.3 Methods for Speech Enhancement

The VAD helps to distinguish and isolate non-overlapping regions of background noise from relevant speech content. However, the background noise overlapping with desired speech regions affects the quality and intelligibility of speech signal [62]. There are different types of degradations present when the speech signal is recorded in practical scenarios. The common types of degradations include background noise, reverberation, and competing for speech signal from other speakers. In the majority of the cases, the presence of background noise degrades the desired speech signal. Hence, the focus of the current chapter is to explore different methods that have addressed enhancement of speech signal degraded by background noise.

Depending on the type of background sources, the background noise can be categorized as stationary and non-stationary types. Typically the stationary noise sources consist of background machine, fan, and air-conditioner, while non-stationary noise sources can include babble noise, background music, and background speakers. Due to the nature of background sources, the power spectral density (PSD) does not vary with time in case of stationary noise, whereas the PSD varies with respect to time in case of non-stationary noise. The methods available in literature consider the background noise to be additive in nature. If desired speech signal $s(n)$ is corrupted by additive background noise $d(n)$ then the resultant degraded speech can be expressed as

$$x(n) = s(n) + d(n) \tag{2.15}$$

The frequency domain representation of degraded speech signal is give by

$$X(k) = S(k) + D(k) \tag{2.16}$$

where k is the index of frequency bin, $X(k)$ is degraded speech signal, $S(k)$ is desired speaker speech, and $D(k)$ is the background degradation in frequency domain. The job of speech enhancement is to estimate the speech signal $\hat{s}(n)$ close to the desired speech signal $s(n)$. Therefore the objective of a good speech enhancement is to minimize the error

$$e(n) = s(n) - \hat{s}(n) \tag{2.17}$$

where $e(n)$ is the estimation error between desired speaker's speech $s(n)$ and estimated speech signal $\hat{s}(n)$.

The problem of enhancing the quality and intelligibility of degraded speech has received considerable attention from the research community for a long time. The primary benefit is to enable smooth communication between humans [62], however, the advantage of speech enhancement is extended to speech based applications like speech and speaker recognition tasks [52]. Hence, the speech enhancement problem addressed using different approaches. Broadly speech enhancement can be classified into *spectral*, subspace based approaches, and *temporal* based methods.

2.3.1 Spectral based Enhancement Methods

The majority of speech enhancement methods available in literature fall into this category. The spectral based enhancement methods estimate the spectral magnitude of desired speech and these methods assume that humans are least sensitive to short-time phase [72, 73]. In case of spectral based enhancement, the methods available in the literature can be classified into two categories. Firstly, the methods are based on spectral modeling of noise components and subsequently subtract noise from degraded speech to obtain spectral magnitude of desired speech signal and such methods are called as *spectral subtraction* methods. The second category of methods such as minimum mean square estimation (MMSE) rely on estimating statistical parameters of signal for speech enhancement.

2.3.1.1 Spectral Subtraction

Spectral subtraction based method is one of the early approaches to speech enhancement. Due to its theoretical simplicity and computational attractiveness the spectral subtraction based methods are still relevant in present day context and they are generally used as state-of-the-art methods to compare the performance of newly proposed techniques in literature. The additive background noise is assumed to be stationary and uncorrelated to desired speech signal in such methods. Hence, based on such assumption the background noise is modeled by calculating the average magnitude spectrum which is subtracted from the overall signal in

frame based approach to estimate the desired speech from the signal. In order to improve the noise modeling the average noise magnitude spectrum is computed during the speech pauses. This helps in better estimation of noise characteristics by taking into account of time varying nature of background sources.

From the degraded speech given in Eqn. 2.16, the estimated speech signal using spectral subtraction is obtained using the following relationship

$$| \hat{S}(k) | = | X(k) | - | \hat{D}(k) | \tag{2.18}$$

where k is the index of frequency bin, $| X(k) |$ is the magnitude spectrum of degraded speech signal, $| \hat{D}(k) |$ is the average magnitude spectrum estimate of background noise, and $| \hat{S}(k) |$ is the clean speech signal estimation using spectral subtraction.

The correct estimation of average magnitude spectrum may not be always possible and further due to dynamic nature of desired speech there can be over estimation of background noise. This leads to have some negative values in the estimated magnitude spectrum $| \hat{S}(k) |$ of desired speech signal. The simplest solution to overcome this problem is to half-wave rectify the resultant magnitude spectrum. However, this gives rise to spectral peaks at certain frequencies within a frame. The random frequency peaks often switch at the frame rate of processing and this corresponds to tones in temporal domain. These undesired tones are often called as *musical noise* in spectral subtraction context [73–75]. The large variance in the estimation of average magnitude spectrum and also variability in the noise attenuation contributes to such musical noise. The presence of musical noise in the enhanced speech signal at times is more annoying than the background noise itself. There are several methods proposed in the literature to overcome this problem [6, 35, 75, 76].

In [2] a method was proposed to reduce the musical noise by using average magnitude spectrum of noise and further attenuation of signal during non-speech regions. The decision of speech and non-speech regions are taken by using an efficient VAD prior to enhancement. However, the residual musical noise is still a problem in the proposed method. The improvement over conventional spectral subtraction method is proposed in [74] to prevent over subtraction

of magnitude spectrum. The minimum threshold of magnitude spectrum can be defined, where the system does not allow it to go below certain minimum set value. The proposed method can is summarized using the following relationship

$$| \hat{S}(k) | = \begin{cases} | X(k) | - \alpha | \hat{D}(k) |, & \text{if } | X(k) | - \alpha | \hat{D}(k) | > \beta | \hat{D}(k) | \\ \beta | \hat{D}(k) |, & \text{otherwise} \end{cases} \tag{2.19}$$

where α is called the over-subtraction factor which is dependent on SNR of the given signal. The α parameter is responsible for attenuation factor of noise components from the signal.

The higher subtraction factor indicates that there is larger attenuation of noise components, however, too larger value of α tries to suppress the desired components of the signal. Therefore, it is necessary to choose the optimal value of α to reduce the musical noise and meanwhile, try to retain desired speech components in the signal. The β defines the noise spectral floor for the given signal and this factor ensures that the subtraction will not go below the certain threshold value set. It can be noticed that tuning α and β parameters is a difficult task and further these parameters depend on SNR of a given signal.

The noise characteristics vary depending on the type of background source and their frequency components differ. Hence, all spectral components of the speech signal are not equally affected by interfering background noise. Therefore suppressing background noise based on the frequency content which affects the speech signal severely helps to reduce musical noise [77]. Hence, as an extension of the system shown in Eqn. 2.19, the over-subtraction factor α is varied with respect to frequency. The frequency components having low SNR are attenuated with larger values of over-subtraction factor and hence called as non-linear spectral subtraction. Such non-linear weightings are applied for each frame of the signal.

As an extension to frequency based analysis, a method is proposed in [6] that uses sub-band analysis for spectral subtraction. The signal is divided into N number of non-overlapping bands and each band is applied with independent over-subtraction factor computed from the corresponding sub-band signal. There are several improvements proposed in the literature

that uses spectral smoothing, formant enhancement, and comb filtering approaches to enhance spectral subtraction to reduce the musical noise [78]. The self-adaptive approach was proposed in [79], where the over-subtraction factor is calculated based on *a priori* SNR estimation. However, it is still a difficult task to completely eliminate the musical noise from the output of spectral subtraction. The effectiveness of these methods in reducing musical noise diminishes at low SNR levels. Recently, a method proposed in [80] uses iterative approach by measuring the power spectral density (PSD) of noise. The iterative steps are given as follows

(i) The average power spectrum of the input noise is estimated.

(ii) The model of average noise magnitude spectrum is subtracted from the original signal. The noise floor level factor β is set high leaving larger noise components as residue and such a subtraction is termed as *weak subtraction*.

(iii) Return to step (i) using input signal as partially increased SNR from step (ii).

It is reported that the method is highly effective in reducing the musical noise. The efficacy of the proposed work depends on the number of iterations the enhancement undergoes, where each iteration helps to reduce the background noise and leads to lesser musical noise [80]. However, having more number of iterative steps can render such methods not suitable for real time scenario.

2.3.1.2 Enhancement using Auditory Masking Properties

It is found that there are several methods proposed to reduce the musical noise using spectral subtraction approach. There is a trade-off between reducing the background noise and causing least distortion to the enhanced speech signal. Based on auditory masking phenomenon not all noise components are audible to humans. Hence, it is possible to hear relatively clean speech signal in spite having noise components below noise masking threshold [81]. There are few methods proposed in the literature that utilizes such properties of human auditory system to enhance speech signal. In [37] the spectral subtraction method is extended to include auditory masking phenomena. The over subtraction factor α and noise floor parameter β

are controlled based on noise masking threshold. The α and β parameters are dependent on frequency components given by $\alpha(\omega)$ and $\beta(\omega)$, respectively. This is similar to non-linear spectral subtraction method [77]. These parameters are dependent on noise masking threshold given by the following relationships

$$\alpha_m(\omega) = F_\alpha[\alpha_{min}, \alpha_{max}, T(\omega)] \tag{2.20}$$

$$\beta_m(\omega) = F_\beta[\beta_{min}, \beta_{max}, T(\omega)] \tag{2.21}$$

where m is the frame index, $T(\omega)$ is the noise masking threshold obtained through modeling the frequency selectivity of the human ear and its masking property, α_{min}, β_{min}, and α_{max}, β_{max} are the minimal and maximal values of the over subtraction and spectral flooring, F_α and F_β are the functions leading to a maximal residual noise reduction. However, the enhancement study was limited to masking phenomena in frequency domain. In [39] the temporal masking properties of human auditory system is used for speech enhancement. Temporal masking is a time domain phenomena in which two stimuli occur within small interval of time. The masker preceding the desired signal temporally is called as forward masking which is effectively used for enhancement.

A wavelet packet transform based sub-band decomposition of noisy speech signal is used to improve the effectiveness of spectral subtraction based algorithm in [82]. The speech signal is divided into 24 critical bands using wavelet decomposition and this is called as *perceptual wavelet packet transform*. The auditory masking threshold can be computed from clean speech signal and seldom such clean speech is available in natural recording scenario. Therefore an initial estimation of clean speech signal is obtained by subtractive type algorithm. The rough clean speech signal estimated is used to derive time-frequency noise masking threshold. The weighted subtraction scheme is adopted based on the assumption that high SNR speech frames are applied with lower subtraction and low SNR frames are subjected to over subtraction. This helps to improve the speech intelligibility in the reconstructed speech using inverse wavelet packet transform.

The auditory masking properties of human audible perception is beneficial in reducing the

musical noise that is evident in spectral subtraction type methods. However, computing noise masking threshold from the degraded speech signal is a challenging task. Also, making use of psycho acoustic phenomenon for the purpose of speech enhancement can make such methods to be more complex and computationally intensive.

2.3.1.3 Minimum Mean Square Estimator

In spectral subtraction based methods there was no specific assumption on the statistical properties of both noise and speech spectral components. Methods are proposed in [4, 83] that utilizes the probability distributions of speech and noisy spectral components. It is observed in [4] that the speech and noisy spectral components as statistically independent zero mean Gaussian random variables. Also, it is known that the estimation of clean speech signal from noisy signal is perceptually mapped to original clean speech through minimum mean square error (MMSE) sense. The speech enhancement is carried using the *short-time spectral amplitude (STSA)* of the speech signal in its perception. The MMSE-STSA estimator is given by the following relationship

$$\hat{M} = E\{(A_k - \hat{A}_k)^2\} \tag{2.22}$$

where $\hat{M}$ is the MMSE-STSA estimator, A_k is the STSA of clean speech signal, and $\hat{A}_k$ is the STSA of estimated speech signal from noise signal.

The MMSE-STSA estimator for speech enhancement aims to minimize the mean square error between the short time magnitude of the clean and enhanced speech signal. It is reported that the optimality criteria gives better enhancement of speech signal with reduced musical noise. However, the optimality criteria is based on the mean square sense and does not consider the non-linear characteristics of human audible perception. Hence, a new method of speech enhancement is proposed in [83] that suits the non-linear nature of human perception called as MMSE log-spectral amplitude (MMSE-LSA). This method is an extension of MMSE-STSA, which aims at minimizing the mean square error between the logarithm of short time spectral amplitude of clean and estimated speech. The relationship of MMSE-LSA is given by

$$\hat{L} = E\{(log\ A_k - log\ \hat{A}_k)^2\} \tag{2.23}$$

where $\hat{L}$ is the MMSE-LSA estimator, $log\ A_k$ is the logarithmic STSA of the clean speech signal, and $log\ \hat{A}_k$ is the logarithmic STSA of estimated speech signal from the noise signal. The subjective analysis carried in [83] reveals that MMSE-LSA is better in terms of reducing musical noise, while there is not much improvement in the speech quality compared to MMSE-STSA.

It is known discrete cosine transform (DCT) has a better energy compaction property compared to discrete Fourier transform (DFT). Hence, the property of this energy compaction of DCT is utilized for MMSE estimator to yield better results compared to MMSE-STSA [84]. The fundamental assumption made in MMSE-STSA estimator is that both real and imaginary parts of DFT coefficients have Gaussian probability distribution function. This assumption is valid asymptotically for a longer duration of the speech frame. However, in reality, the frame size of (20-30) ms is adopted for analysis. Hence, the Gaussian distribution assumption may be valid for noisy components of signal but it may not be suitable for speech components of the signal. Therefore different probability distributions are attempted to model real and imaginary parts of DFT coefficients, in particular, Gamma and Laplacian distributions are used for modeling in [3, 85].

The ideal binary masking (idbm) techniques exploit the *a priori* SNR measurements to separate clean speech and noise components from the signal. In order to estimate the clean speech signal, a time-frequency transformation is applied with binary masking based on *a priori* SNR. The time-frequency analysis is carried by applying DFT or gamma-tone filter bank analysis in a frame-based approach [86]. The core idea is that the noise components are suppressed by a gain value of g_{min} and speech components with a gain value of g_{max}, where g_{min} and g_{max} are the gain values of ideal binary mask function. The ideal binary function is given by the following relationship

$$g(k,m) = \begin{cases} g_{max}, & \text{if } \frac{|s(k,m)|^2}{|d(k,m)|^2} > \rho(k,m) \\ g_{min}, & \text{otherwise} \end{cases} \tag{2.24}$$

where $s(k,m)$ and $d(k,m)$ are the time-frequency components of clean speech signal and noise,

respectively, $\frac{|s(k,m)|^2}{|d(k,m)|^2}$ is the signal-to-noise ratio, and $\rho(k,m)$ is the threshold value of SNR. However, in natural recording scenario seldom *a priori* SNR is available and binary masking has to be calculated on the basis of *a posteriori* SNR. Due to the error in estimation of SNR, the quality of idbm based method suffers from poor speech enhancement quality. In order to over come this problem a continuous gain function to minimize the spectral magnitude MSE approach is proposed in [86]. It is found that the quality of continuous gain function is superior compared to binary masking function.

In contrast to conventional MMSE based speech enhancement approaches a supervised approach to enhance the speech signal is proposed in [4]. A mapping function is found between clean speech signal and its estimation using deep neural networks (DNN). In order to accommodate all possible combinations of speech and noise signal, a simulated version of database is created by adding several types of noise at different levels to speech signal. A DNN architecture is then employed as a non-linear regression to model the mapping function. One of the issues with such supervised approaches is that the performance degrades for unseen noise types. In order to overcome such problem a global variance equalization is used to avoid over-smoothing [87]. It is experimentally shown that such supervised methods give an improved performance with respect to objective and subjective scores compared to conventional MMSE methods. However, large deviation in the noise signal characteristics can still degrade the performance of such methods. Further, such supervised approaches requires huge training data for supervision and they are computationally intensive.

Though there are several methods proposed in the literature as improvements over conventional MMSE-STSA and MMSE-LSA methods. Still by far MMSE based approaches proposed in [4, 83] are the most popular methods. The MMSE based methods are practically used in hearing-aid applications for speech enhancement [88]. Also, these methods are used as state-of-the-art techniques to benchmark the performance of newly proposed methods.

2.3.2 Subspace Approaches for Enhancement

Another set of methods that helps to separate the desired speech signal and noise is signal subspace filtering. The speech and noise components are decomposed into mutually orthogonal subspaces. The decomposition is possible under the assumption of a low-rank linear model for speech and interfering noise. Here, the speech signal is assumed to be correlated while noise components are uncorrelated [89]. The speech enhancement happens by removing the noise subspace and reconstructing the signal back using speech subspace. The assumption in all subspace based approaches is that every short-time speech vector $s = [s(1), s(2), ..., s(q)]^T$ can be written as a linear combination of $p < q$ linearly independent basis functions $m_i \;\; i = 1, ...p$, given as

$s = My$

where M is a $(q \times p)$ matrix containing the basis functions and y is a length-p column vector containing the weights. The decomposition of the noisy signal into signal subspace and noise subspace can be done using either the singular value decomposition (SVD) [90] or Karhuen-Loeve transform (KLT) [91]. The SVD based approaches resolves the signal and noise subspaces in terms of eigenvectors and their corresponding eigenvalues. Both are separable in eigen subspaces, where the signal reconstructed using dominant eigenvectors and their corresponding eigenvalues leads to clean speech signal. Similarly, the study was extended using quotient SVD (QSVD) in [92]. The noisy signal is decomposed into speech and noise subspaces using KLT, where the gain function derived helps to distinguish speech and noise components. Inversion of KLT obtained using the modified KLT co-efficients by gain function helps to enhance the speech signal. Since, the subspace based approaches work on the basis of stationarity assumption of noise, the performance for non-stationary types of degradations is poor. Further, most of the subspace based approaches suffers from huge computational load.

2.3.3 Temporal based Enhancement Methods

Most of the methods discussed so far target the modeling of background noise and tries to eliminate from the signal to estimate clean speech signal. However, background noise can be

from different sources and hence modeling of such noise may not always be feasible. Further, if the background noise is having non-stationary characteristics then modeling such sources becomes a difficult task. For example, if the background noise is from a background speaker or from a source having similar characteristics, most of the conventional approaches may fail to enhance desired speech signal. Hence, there is a need for alternative approaches that focus on exploiting the characteristics of desired speech signal.

There is some evidence that humans perceive information from noisy speech signal using high SNR regions then extrapolating such information to low SNR regions to fill the gaps created by interfering noise [9]. Based on this motivation, a method was proposed in [8] that effectively used high SNR regions of excitation source information to enhance the degraded speech signal. The excitation source signal in the form of linear prediction (LP) residual is modified to excite time-varying all-pole filter to obtain enhanced speech. Since, LP residual signal is a random polarity signal and has noise-like characteristics [8], modifying such signal by enhancing instants of significant excitation relative to other regions helps to reduce the interfering noise. Also, this causes least distortion to enhanced speech signal unlike spectral subtraction or MMSE based methods that introduces musical noise. In order to exaggerate instants of significant excitation the Frobenius norm of the Toeplitz matrix constructed using 2 ms frame size of LP residual is used. A similar approach was proposed in [93] that uses a different weighting scheme for LP residual using a constrained optimization criteria.

The idea of modifying the LP residual signal for speech enhancement is extended to multi microphone scenario in noisy and reverberant environment in [94]. Time delay between two different microphone recordings is computed using the cross-correlation between residual signals. In order to enhance the instants of significant excitation Hilbert envelope of LP residual weighting function is multiplied with residual signal. The enhanced LP residual signal is synchronously added together to synthesize the enhanced speech signal using multi-microphone recordings. The idea is that the interfering noise and reverberation components of multiple recordings gets added incoherently, while the desired speech components gets added coherently. This coherent addition of desired speech components by compensating for the delay helps to

exaggerate high SNR regions further and hence enhances the speech signal.

A method is proposed in [43] that uses both temporal and spectral processing for speech enhancement. The glottal closure instant (GCI) locations are identified using Hilbert envelope of LP residual and such locations are further boosted in LP residual relative to other regions. The modified LP residual is used to synthesize first stage of enhanced signal. Further, the enhanced signal is subjected through spectral subtraction to produce a better enhanced speech. The advantage of such combined temporal and spectral techniques is that the enhanced speech is of better quality and has no musical noise.

2.4 Motivation for Current Work

Most of the speech and non-speech detection studies carried in the literature assumed the statistical independent nature of noise and speech signal. Also, many methods rely on modeling the noise based on stationary characteristics to distinguish speech and non-speech regions of the signal. However, this assumption may not be valid when background noise is having speech-like characteristics. For example, the background interfering source can be speech or music and therefore it becomes difficult to distinguish relevant speech from rest of the degradation.

In natural recording scenario, the proximity of the desired speaker is closer to microphone sensor compared to other background sources. It is, therefore, necessary to utilize high SNR regions of the speech signal to identify desired speaker's speech from rest of the background interference. The epoch locations of speech signal recorded close to microphone are robust to interfering background sources. Hence, it should be possible to identify the relevant speech recorded from rest of the background sources using such high SNR regions. It is shown in [32,33] that the features derived using the robust aspects of speech production can be an effective solution to segregate relevant desired speaker's speech from rest of the background noise.

It is evident from the literature that most of the speech enhancement methods in literature try to model background noise in order to eliminate it from the signal to retain the enhanced speech signal. However, such methods introduce unwanted distortion in the form of musical noise and hence many methods have focused on reducing such distortion. The speech enhance-

ment methods that utilize psychoacoustic phenomena and subspace-based approaches do not introduce such unwanted distortion, yet such methods can be more complex. Alternatively, the speech production is a well-studied topic in the literature for the past few decades and thus can be utilized for speech enhancement. Especially high SNR regions around epoch locations in temporal domain and formant peak locations in spectral domain can be effectively used for enhancement.

Most speech enhancement approaches study additive background noise cases. In most of the natural recording scenario, the proximity of the desired speaker is closer to microphone relative to other background sources. Due to such variations in the distance between the microphone and different acoustic sources, the signal characteristics vary between desired speaker's speech and rest of the background sources. Further the speech production mechanism tend to vary depending on the interfering background noise and this is called *Lombard effect*. Hence, it is beneficial to account any such changes in signal characteristics for speech and non-speech detection and further enhance the desired speech signal.

The temporal-based enhancement approaches that modify LP residual signal and formant peaks can be a promising direction to handle such natural speech recordings in the presence of other background interfering sources. The methods introduce the least distortion to an enhanced speech signal in the form of musical noise. The GCI extraction method forms one of the important blocks in such enhancement schemes. However, Hilbert envelope of LP residual is not a reliable method for extracting GCI locations [95]. Recently a method for epoch extraction based on 4^{th} order resonator filter at 0 Hz called as *zero frequency filter (ZFF)* is proposed [70] [1]. The ZFF is one of the reliable methods existing for epoch extraction and it is robust to interfering background noise.

[1]In this work GCIs and epochs are interchangeably used.

3

Epoch Extraction using Zero Band Filtering

Contents

Objective

Based on the review, it is evident that the speech enhancement by modification of glottal closure instants (GCIs) introduces the least distortion. Therefore, it is necessary to identify the locations of GCIs accurately. The *Zero Frequency Filter (ZFF)* is a marginally stable infinite impulse response (IIR) resonant filter at 0 Hz used to extract the epoch locations reliably from speech signals. However, the output of such an ideal resonator is an exponentially increasing / decreasing function of time. The trend is removed from the filtered output by subtracting the average over 1 - 2 pitch periods to obtain *zero frequency filtered signal (ZFFS)*. Alternatively, in this chapter a Bounded Input Bounded Output (BIBO) stable realization of ZFF is proposed for epoch extraction, where, the output of such a filter is not an increasing / decreasing function of time. The advantages of using such a stable filter are that the filter output is bounded and has no precision related problem associated with the output for lengthy speech files. Also, the method does not require remove trend procedure that needs initial pitch estimation. It can be noted that it is difficult to estimate pitch from the signal recorded in degraded conditions. The proposed approach is evaluated using the CMU-Arctic database for clean and degraded conditions. Furthermore, the method is also validated in cases of the singing voice and emotional speech to demonstrate the robustness for varying pitch scenarios. The proposed method is found to be robust for a wide range of chosen parameters.

3.1 An Overview of Epoch Extraction from Speech

Epochs are instants of significant excitation present within a pitch period. Most of the time instants of significant excitation takes place during the glottal closure of the voiced speech regions [57, 70]. Due to time varying nature of both voiced excitations and vocal tract characteristics, estimating the accurate location of epochs in voiced regions is still a challenging task. When glottis closes suddenly, a puff of air excites the vocal tract system. During such events, a rapid change takes place in the speech signal that gets manifested as sharp peaks in amplitude. However, it is not easy to detect such locations directly from speech signals.

It can be imagined that [96] nature of each of the epochs is impulse-like. The train of such impulse-like excitations with varying time and amplitudes gets convolved with time varying vocal tract system. Since, neither source nor vocal tract system is known *a priori*, separating one of them from other essentially turns out to be a blind deconvolution problem. Knowing the accurate location of epochs in speech signal has several applications in speech analysis [97], pitch synchronous based speech synthesis and desired foreground speech segmentation from rest of the background noise [32]. One of the obvious derivatives of knowing accurate epoch locations is the estimation of [98] fundamental frequency ($F0$). The $F0$ parameter is used in applications like speech synthesis [99] and prosody modification [100]. It is found that the nature of ZFFS offers discriminative information between foreground speech and background noise that includes background speech [32]. It is shown that epoch locations and strength of excitation parameters are useful in segmenting foreground speech regions from rest of the background regions in noisy environments collected naturally. The foreground speech refers to the content spoken by a speaker close to the sensor. While, rest of the content is termed as background. Many such applications derived as a result of knowing epoch locations and the challenges involved in accurately estimating those locations has motivated researchers towards addressing this problem.

Most of the existing methods in the literature rely on source filter separation model. Inverse filtering operation on speech signal using estimated vocal tract filter is used to derive LP residual signal. Though, it is not possible to extract epoch information directly from LP residual due to random polarities [101] many methods exploit it implicitly or explicitly for epoch extraction [58, 71, 101–104]. Deriving LP residual from speech signal is based on the assumption that the analysis window of 20-30 msec is stationary. However, this assumption of stationary does not hold as both voiced excitation and vocal tract system can be varying within the analysis window. Also, such methods rely on the higher energy of LP residual signal at epoch locations relative to other regions and this may not be true always. Furthermore, based on the global phase characteristics of speech signal obtained from LP residual, many methods were proposed on the basis of group delay function [1, 57]. A detailed quantitative review

amongst top four methods based on group delay techniques were present in [60]. However, reported identification rates and accuracies of group delay based methods are relatively poor when compared to current state of the art methods reported in [96].

To alleviate perfect deconvolution problem associated with LP residual based methods, a method was proposed in [70]. This method does not depend on the critical ability to deconvolve vocal tract system response from voiced excitations. The method was proposed based on the analysis that significant excitations are impulse-like and their strengths are significantly larger than other regions. Hence, those significant excitations are in the form of the train of impulses that excites vocal tract system. The effect of such impulses is spread throughout the bandwidth of the speech signal under analysis in the frequency domain. It is, therefore, evident that the impulse-like excitation information is present at 0 Hz component as well. The vocal tract system has the least effect on significant excitations at 0 Hz frequency component. In order to extract this information at 0 Hz, an integrator is designed using a marginally stable 0 Hz resonator and it is called *zero frequency filter (ZFF)*. Effectively a 4^{th} order resonator realized as two cascaded 2^{nd} order resonators are used to have a steeper roll off. The nature of the filter output is either exponentially increasing / decreasing function of time. *Zero frequency filtered signal (ZFFS)* is obtained by removing the average values within each analysis window. The analysis window size depends on 1 - 2 pitch periods on average for that particular speaker. The epoch locations are obtained from ZFFS, where, the positive zero crossings exactly coincide with epoch locations. Estimation of the initial average pitch period of the speaker becomes necessary within 1 - 2 pitch periods for accurate estimation of epoch locations. Also, since the output of the filter grows / decays as a polynomial function of time, consequently, exceeds the precision range of the processor that results in noisy output for lengthy files. In order to overcome such issues present in ZFF, we propose a method based on stable infinite impulse response (IIR) resonant filter to extract epochs from the speech signal. Since such a filter allows a narrow band of frequencies around 0 Hz to pass through, we prefer to call this filter as *zero band filter (ZBF)* and the output of the filter as *zero band filtered signal (ZBFS)*.The filter is stable and the output of filter is not exponentially increasing / decreasing function of time.

Hence, it is not imperative to remove the trend from output signal in order to obtain accurate epoch locations. The positive zero crossings of the ZBFS indicates the epoch locations.

The rest of the chapter is organized as follows: Section 3.2 explains ZFF method in brief and the design analysis of proposed ZBF method and its impact on epoch location extraction. Also, the section describes the issues present in proposed method and ways to overcome the same. Section 3.3 describes the details of evaluation procedure and results obtained in terms of identification rate and accuracies. The robustness of proposed method in terms of varying pitch scenarios is described in Section 3.4.2. The summary of present work is mentioned in Section 3.5.

3.2 Zero Band Filtering of Speech

The zero frequency filtering method is briefly described here as it is necessary for explaining the proposed method [70]. Let $s(n)$ be the speech signal where the epochs have to be identified. Difference signal $x(n)$ is obtained from $s(n)$ to minimize any low frequency fluctuations present and is given as

$$x(n) = s(n) - s(n-1) \tag{3.1}$$

The difference signal is passed twice through an ideal resonator at zero frequency given by

$$y_1(n) = -\sum_{k=1}^{2} a_k y_1(n-k) + x(n) \tag{3.2}$$

and

$$y_2(n) = -\sum_{k=1}^{2} a_k y_2(n-k) + y_1(n) \tag{3.3}$$

where $a_1 = -2$ and $a_2 = 1$. This is equivalent of successively integrating the input four times, termed more commonly as filtering at zero frequency. The trend in $y_2(n)$ is removed by subtracting the average over 1 - 2 pitch periods at each sample. This results in signal $y(n)$

$$y(n) = y_2(n) - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2(n+m) \tag{3.4}$$

and is called the *zero frequency filtered signal (ZFFS)*. Here $2N+1$ corresponds to the number of samples in 1 - 2 pitch periods on average for that particular speaker. Using ZFFS, the epoch locations can be exactly located at positive zero crossings.

In order to explain the output to input relationship using a 2^{nd} order filter from Eqn. (3.2) and to represent the system transfer function in generalized form, Eqn. (3.2) can be expanded as

$$y_1(n) = -a_1 y_1(n-1) - a_2 y_1(n-2) + x(n) \tag{3.5}$$

The z-domain equivalent of Eqn. (3.5) is

$$Y_1(z) = -a_1 Y_1(z) z^{-1} - a_2 Y_1(z) z^{-2} + X(z) \tag{3.6}$$

and rearranging Eqn. (3.6) we get

$$\frac{Y_1(z)}{X(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{3.7}$$

generalizing Eqn. (3.7) by substituting $a_1 = -2r$ and $a_2 = r^2$ [105] we get

$$\frac{Y_1(z)}{X(z)} = \frac{1}{1 - 2rz^{-1} + r^2 z^{-2}} \tag{3.8}$$

where, r represents the value of the radius on unit circle in z-plane at which the poles are placed and it's equivalent time domain Equation is given as

$$y_1(n) = x(n) + 2r y_1(n-1) - r^2 y_1(n-2) \tag{3.9}$$

Eqn. (3.9) can be re arranged as follows

$$y_1(n) - r y_1(n-1) = x(n) + r y_1(n-1) - r^2 y_1(n-2) \tag{3.10}$$

Let

$$y_3(n) = y_1(n) - ry_1(n-1) \tag{3.11}$$

substituting Eqn. (3.11) in (3.10) and re arranging the Equation we get

$$y_3(n) = x(n) + r(y_1(n-1) - ry_1(n-2)) \tag{3.12}$$

using Eqn. (3.11), the above Eqn. (3.12) can be re-written as

$$y_3(n) = x(n) + ry_3(n-1) \tag{3.13}$$

Eqn. (3.13) can be expanded as the following infinite series in terms of input signal with the assumption that $y(n) = 0$ for $n < 0$

$$y_3(n) = x(n) + rx(n-1) + r^2x(n-2) + \ldots \infty \tag{3.14}$$

Eqn. (3.14) can be written as summation series shown as

$$y_3(n) = \sum_{m=0}^{\infty} r^m x(n-m) \tag{3.15}$$

Eqn. (3.11) can be re-written as

$$y_1(n) = y_3(n) + ry_1(n-1) \tag{3.16}$$

the above Eqn. (3.16) can be written in summation series similar to Eqn. (3.15)

$$y_1(n) = \sum_{l=0}^{\infty} r^l y_3(n-l) \tag{3.17}$$

substituting Eqn. (3.15) in (3.17) we get

$$y_1(n) = \sum_{l=0}^{\infty} r^l \sum_{m=0}^{\infty} r^m x(n-m-l) \tag{3.18}$$

where Eqn. (3.18) can be re-written as

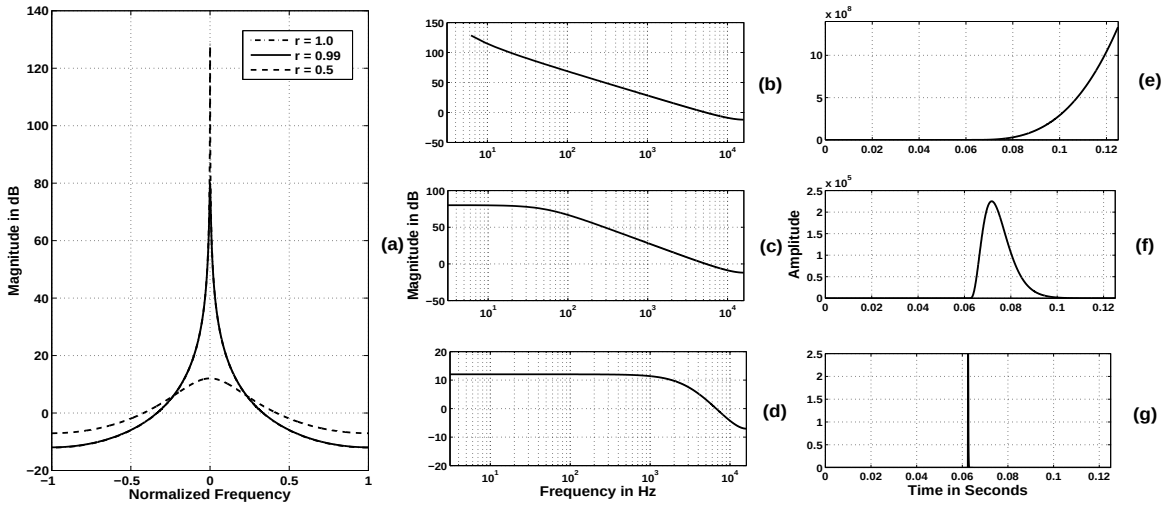

Figure 3.1: Second order filter responses (a) Magnitude response (normalized frequency axis) when poles are placed at r = 1.0, 0.99 and 0.5. Magnitude responses (semilogx) and corresponding impulse responses when poles are placed at r = 1 (b) and (e), r = 0.99 (c) and (f), r = 0.5 (d) and (g).

$$y_1(n) = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty} r^{(m+l)}x(n-m-l) \tag{3.19}$$

If $m + l = k$ in the above Eqn. (3.19), the Equation can be written as

$$y_1(n) = \sum_{k=0}^{\infty}(k+1)r^k x(n-k) \tag{3.20}$$

Eqn. (3.20) is the generalized form of representation that is derived in [106]. When $r = 1$, in Eqn. (3.20) represents the ZFF and it is evident that the impulse response of the system is diverging type. This explains the reason for the nature of *zero frequency filter* output growing/decaying exponentially.

3.2.1 Zero Frequency to Zero Band Filtering

If $r < 1$ the impulse response of the system is converging type. From Eqn. (3.8) varying value of r results in a set of filters having different magnitude responses. This is illustrated by magnitude response plots of 2^{nd} order filter for 3 different values of r at 1.0, 0.99 and 0.5 in Figure 3.1 (a). The magnitude response of the filter at $r = 0.99$ is almost same as that

of $r = 1.0$ except that it represents bounded case and gain at 0 Hz for $r = 0.99$ is relatively low compared to $r = 1.0$. Furthermore, Figures 3.1 (b), (c) and (d) represents the magnitude responses of 2^{nd} order filter when poles are placed at $r = 1$, $r = 0.99$ and $r = 0.5$, respectively. While, Figures 3.1 (e), (f) and (g) are corresponding impulse response plots, for clarity purpose, the frequency axis is represented in semi-log scale and an impulse at time instant of 0.0625 secs.

It can be noticed that as r value increases, the nature of low pass magnitude response appears to be sharper. Also, the dynamic range of filter increases as r value increases, thereby increasing the gain value at 0 Hz relative to other frequency components. When r approaches unity, the filter gets manifested as an ideal resonator and from Eqn. (3.8), it is evident that the gain of the filter is ∞ at 0 Hz. This is achieved at the cost of placing the poles on unit circle that results in the marginally stable filter. If the speech signal is passed through such a filter, the output results in exponentially growing / decaying function of time. While, the epochal information is present over the exponential trend that may not be visible directly from the plots. However, when such an exponential trend is removed from the output signal by using the average over 1 - 2 pitch periods as shown in Eqn. (3.4), ZFFS signal is obtained. Alternatively, to avoid such an issue, we prefer to apply the stable filter by placing poles within the unit circle. The advantage of using such a stable filter is that the output is converging type. The stability is obtained at the cost of finite gain at 0 Hz and an increase in bandwidth of the filter. However, if the filter magnitude response is sufficiently narrow enough that allows a band of frequencies near 0 Hz without affecting the ability to extract epochs, then such a filter is preferable. Since, such a stable filter allows a narrow band of frequencies near 0 Hz, we would like to call such a filter as *Zero Band Filter (ZBF)*.

Let h_{ZF2} and h_{ZB2} be the impulse response of 2^{nd} order ZFF and ZBF, respectively. The impulse response h_{ZF2} and h_{ZB2} can be obtained from Eqn. (3.20) as follows

$$h_{ZF2}(n) = \sum_{k=0}^{\infty}(k+1)\delta(n-k) \tag{3.21}$$

Eqn. (3.21) represents the impulse response of ZFF when $r = 1$.

$$h_{ZB2}(n) = \sum_{k=0}^{\infty}(k+1)r^k\delta(n-k) \tag{3.22}$$

Eqn. (3.22) represents the impulse response of ZBF and $r < 1$ in case of ZBF.

The voiced speech signal can be considered as the convolution between the train of impulse excitations and vocal tract system response and it is of interest to observe the filter output response to such train of impulses. Both period and amplitude of such train of impulses vary practically in the speech signal. In order to simplify the analysis, the period of such train of impulses is considered to be constant N having unit amplitude. The input signal $x(n)$ in the form of train of impulses is given by Eqn. (3.23) with period N and having unit amplitude.

$$x(n) = \sum_{j=0}^{\infty}\delta(n-Nj) \tag{3.23}$$

Using Eqn. (3.21), the ZFF output response for train of impulses given in Eqn. (3.23) is given by

$$y_{ZF2}(n) = \sum_{j=0}^{\infty}h_{ZF2}(n-Nj) \tag{3.24}$$

Similarly, using Eqn. (3.22) the ZBF output is given as

$$y_{ZB2}(n) = \sum_{j=0}^{\infty}h_{ZB2}(n-Nj) \tag{3.25}$$

When, two such filters are cascaded then the impulse response of 4^{th} order ZFF is given as

$$h_{ZF4}(n) = h_{ZF2}(n) * h_{ZF2}(n) \tag{3.26}$$

similarly 4^{th} order impulse response of ZBF is

$$h_{ZB4}(n) = h_{ZB2}(n) * h_{ZB2}(n) \tag{3.27}$$

The output response of 2^{nd} order filters are given by Eqns. (3.24) and (3.25) to such train of impulses as input when poles are placed at $r = 1$ in the case of ZFF, while the poles are

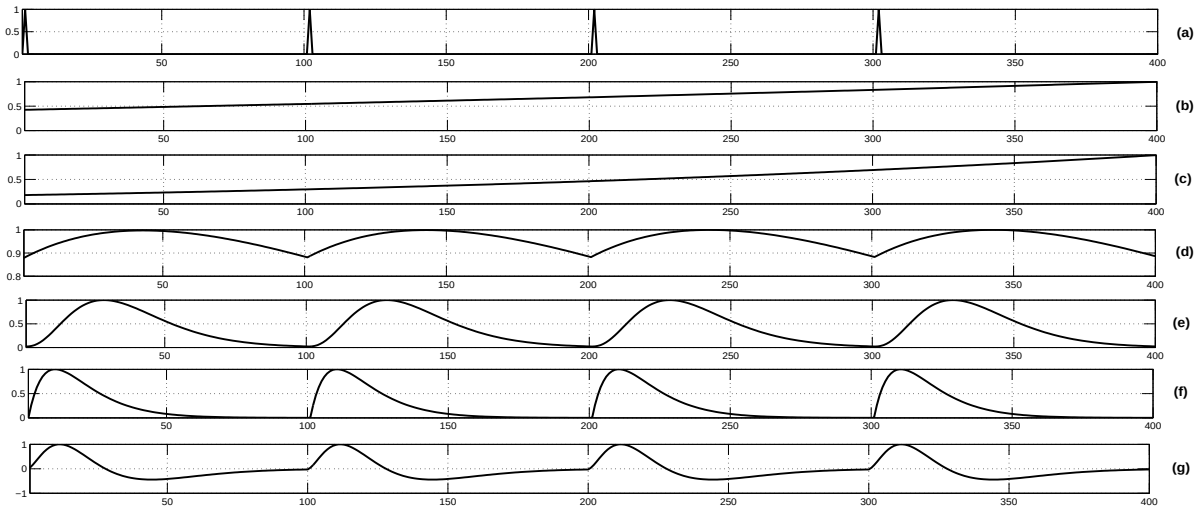

Figure 3.2: Illustration of 2^{nd} and 4^{th} order filter output response for input train of impulses when poles are placed at different r values (a) Input train of impulses and the filter response when poles are placed at (b) 2^{nd} order filter output when r = 1.0, (b) 4^{th} order filter output when r = 1.0, (c) 2^{nd} order filter output when r = 0.99, (d) 4^{th} order filter output when r = 0.99, (e) 2^{nd} order filter output when r = 0.9, (f) 4^{th} order filter output when r = 0.9.

placed within unit circle in case of ZBF. The output response can be observed in Figure 3.2 for the train of impulses as an input, where the period N is 100 samples. It can be observed from Figures 3.2(b) and (c), the output is an increasing function of time when poles are placed at $r = 1$ for 2^{nd} and 4^{th} order filters, respectively. However, it can be observed from Figures 3.2(d), (e), (f) and (g) that the output is bounded and forms the converging series. It can be further noticed that the output has faster decay when poles are placed farther away from unit circle from Figures 3.2(f) and (g), when poles are placed at $r = 0.9$ for 2^{nd} and 4^{th} order filters, respectively.

As we can notice from Figure 3.1 that placement of poles plays a critical role in fixing the magnitude response of the filter. Also, using higher order filters, one can increase the gain of the filter at 0 Hz component relative to other components. Hence, we prefer to use a 4^{th} order filter to obtain a better roll off as used in ZFF. Figure 3.4 illustrates the effect of 4^{th} order filter (realized as a cascade of two 2^{nd} order filters) for different pole placement values on the output signal when applied on speech signal from a male speaker in CMU-Arctic database, where a segment of voiced speech region is displayed.

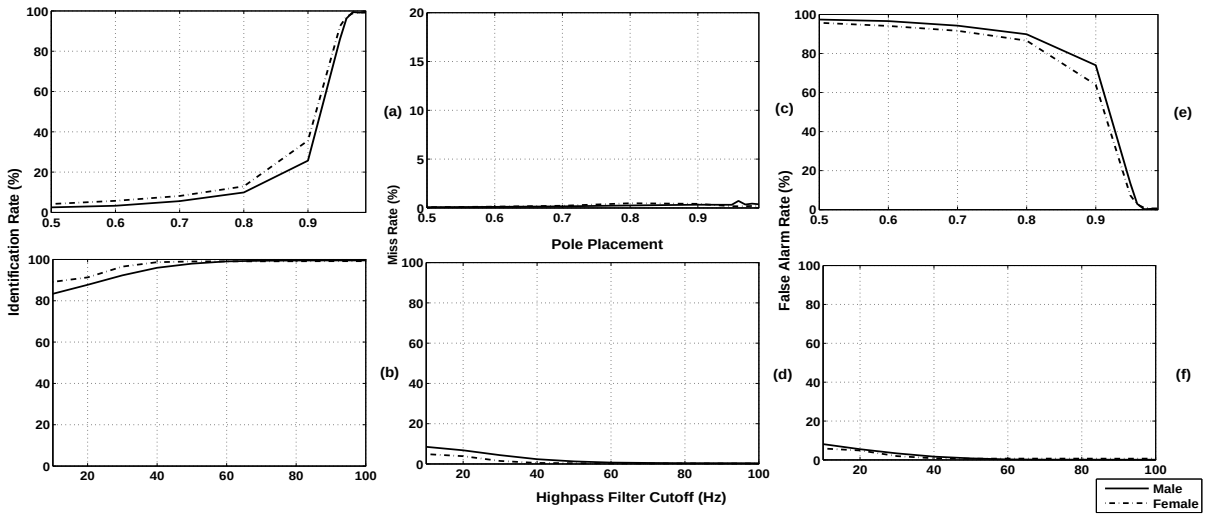

Figure 3.3: Performance evaluation of ZBF using variable parameters, (a) and (b) are Identification Rate (c) and (d) are Miss Rate (e) and (f) are False Alarm Rate for varying values of pole placement and high pass filter cutoff frequency, respectively.

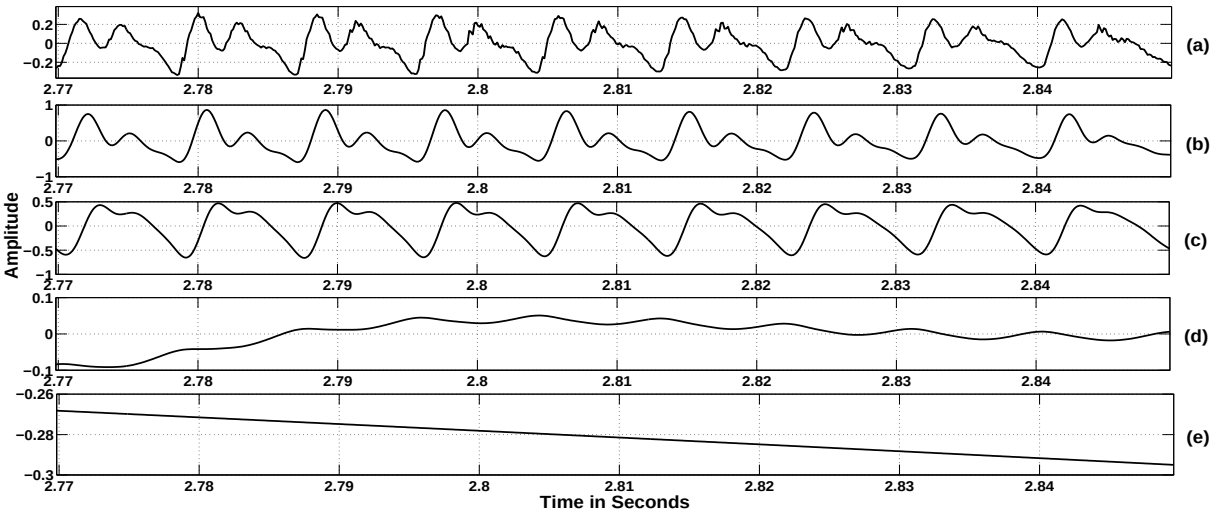

Figure 3.4: Illustration of 4^{th} order filter outputs for different values of pole placements (a) Voiced speech segment of vowel /i/ from a speech file taken from CMU-Arctic database, fourth order filter output of a voiced speech segment in (a) when poles are placed at (b) r = 0.8, (c) r = 0.9, (d) r = 0.99, (e) r = 1.0.

The Figures 3.4 (b), (c), (d) and (e) are the 4^{th} order normalized filter outputs for the segment of speech signal for different pole placements at radius r of 0.8, 0.9, 0.99 and 1.0, respectively. We can observe from Figure 3.4 that as poles are farther away from the unit circle, the filter allows more frequency components to pass relative to 0 Hz component. Hence,

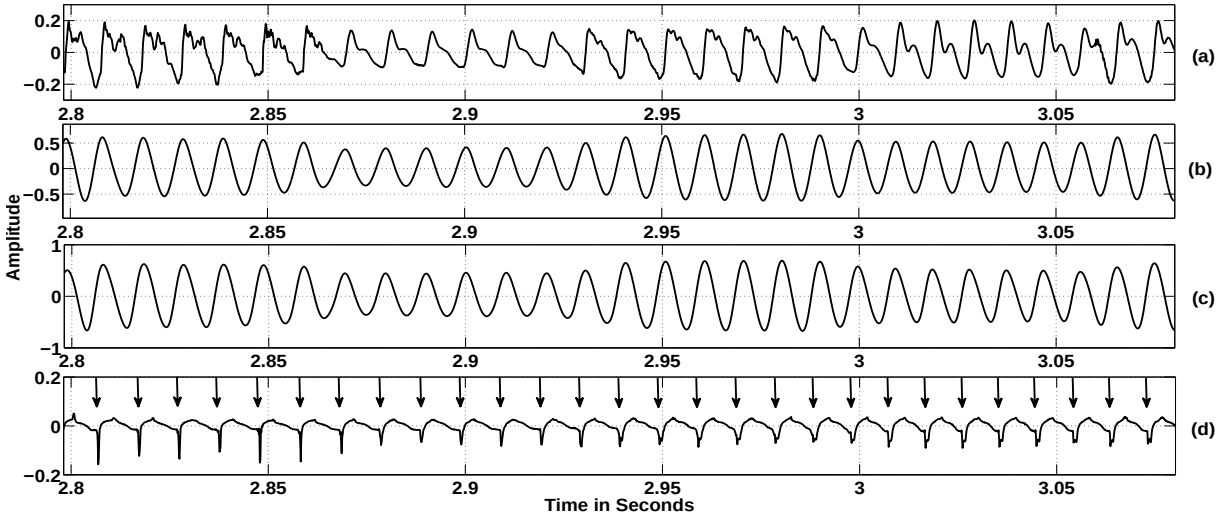

Figure 3.5: Comparison of ZFFS and ZBFS (a) Voiced speech segment of vowel /e/ from a speech file taken from CMU-Arctic database, (b) Zero Frequency Filtered output for voiced speech segment in (a), (c) Zero Band Filtered output for voiced speech segment in (a), (d) differenced EGG of voiced speech segment shown in (a) and arrows representing the actual epoch locations.

it is preferable to design a filter by placing the poles close to unit circle to avoid interference of vocal tract system response with epochal information. However, if the poles are placed on the unit circle, the filter becomes marginally stable, where, the output is either exponentially growing or decaying function of time as we can notice this from Figure 3.4(e). Hence, to avoid any such issues it is preferable to use a stable filter, while poles are placed close to the unit circle.

It can be observed from Figure 3.4(d) that when poles are placed at 0.99, the epochal information is overriding on a low-frequency fluctuation. This low-frequency fluctuation can be eliminated by highpass Butterworth filter. In order to have a sharper transition, a 4^{th} order filter is preferable. However, the parameters for ZBF and the highpass filter are chosen based on the experiments conducted for varying pole placements and highpass filter cutoff frequencies, respectively. In order to eliminate any bias in the chosen parameters, two different sets of databases belonging to male and female speakers are selected from CMU-Arctic databases for experimentation. First, the experiment involves selection of pole placement by varying its value for ZBF by assessing the performance of the method for fixed highpass cutoff frequency. The

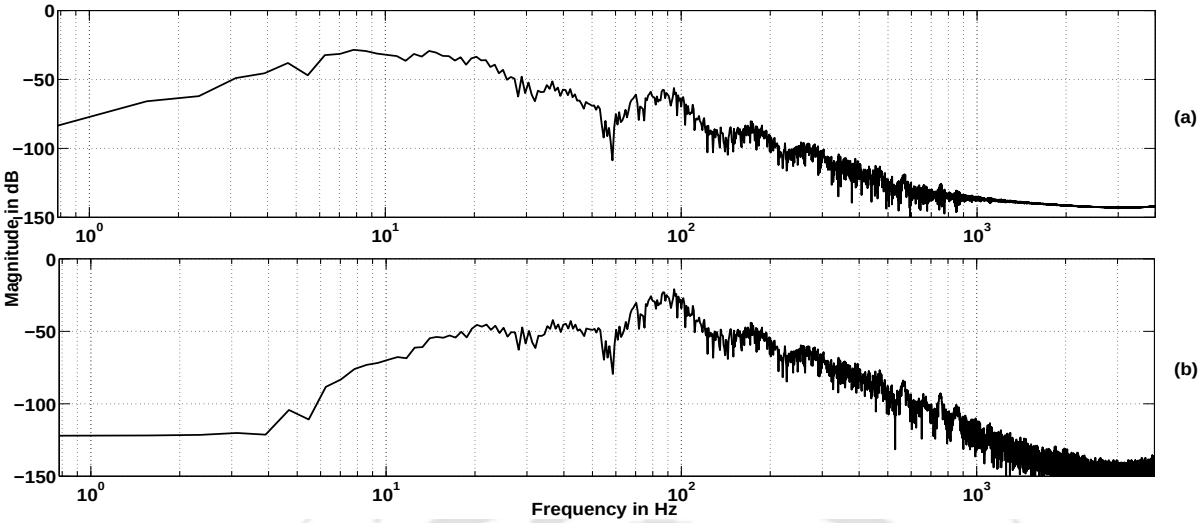

Figure 3.6: Spectrum plot of *Zero Band Filter* output for a speech file from CMU-Arctic database, (a) before and (b) after passing through 4^{th} order Butterworth high-pass filter at cutoff frequency of 80 Hz.

performance is measured in terms of identification rate (IDR), miss rate (MR) and false alarm rate (FAR) for varying values of poles from 0.5 to 0.99, while keeping the highpass filter cutoff frequency at 80 Hz. Figures 3.3 (a), (c) and (e) shows the plots of IDR, MR and FAR evaluated for two speakers databases. It can be observed that as the pole is farther away from the unit circle in the range of 0.5 to 0.9 IDR of the epoch extraction is poor because ZBF allows larger bandwidth relative to 0 Hz component. As a result, the ZBFS has many spurious positive zero crossings that lead to higher false alarms and this is evident from Figure 3.3 (e). In contrast, the IDR improves significantly when the poles are placed in the range of 0.95 to 0.99. The poles placed at 0.99 is selected because ZBF has the least bandwidth relative to other positions. The highpass filter cutoff frequency is chosen based on an experiment conducted by fixing the ZBF pole placed at 0.99 and varying cutoff frequencies from 10 Hz to 100 Hz in steps of 10 Hz.

The Figures 3.3 (b), (d) and (f) show the plots of IDR, MR and FAR evaluated for varying highpass filter cutoff frequencies. It can be observed that IDR is lower when the cutoff frequencies are in the range of 10 to 60 Hz and improves the range of 60 to 100 Hz. Highpass filter cutoff frequency of 80 Hz is selected in the proposed method so that it matches closely with lower pitch range for a male speaker. The Figures 3.6 (a) and (b) shows the spectrum plots of

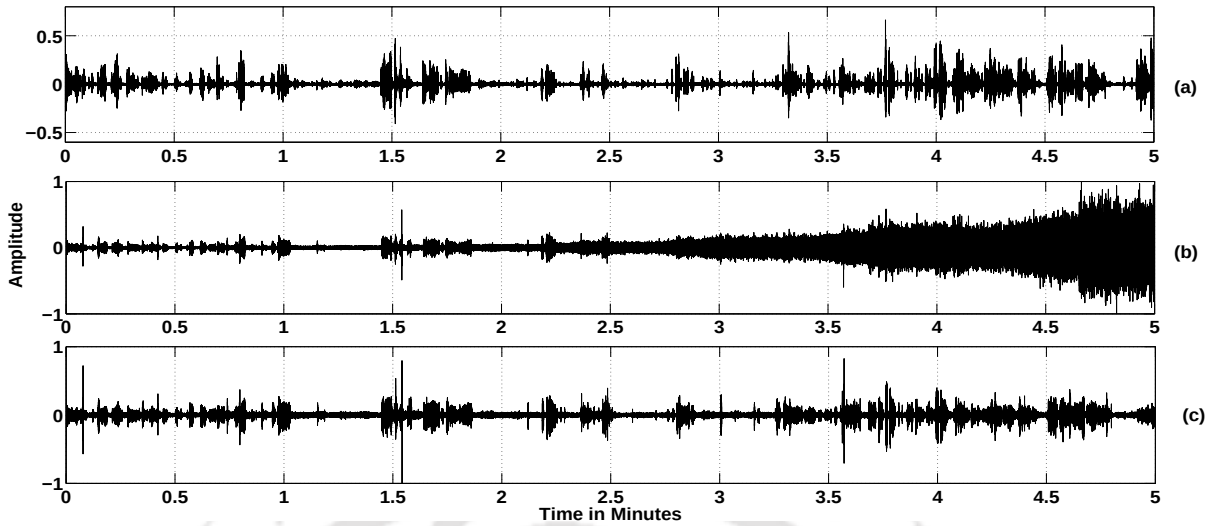

Figure 3.7: Illustration of robustness of ZBF for lengthy speech file (a) Speech waveform from a file taken from mic set of NIST 2012 database, (b) Zero Frequency Filtered output for the waveform shown in (a), (c) Zero Band Filtered output for the waveform shown in (a).

the filter output for a speech signal before and after applying the high-pass filter, respectively. We can notice that the dominant low-frequency content is suppressed by a high pass filter at 80 Hz to give emphasis to a band of frequencies that contain epochal information. Figure 3.5 shows the plots of ZFFS, ZBFS and the difference EGG for a voiced speech segment. Figures 3.5 (b) and (c) shows the plots of ZFFS and ZBFS, respectively for the voiced speech segment shown in Figure 3.5 (a) and as a reference, the differential EGG is plotted in Figure 3.5 (d). As it can be observed that ZFFS and ZBFS are almost similar and their corresponding positive zero crossings match with reference epoch locations.

Also, from Eqn. (3.20) it is evident that the output of ZFF grows / decays exponentially as a function of time. This poses a problem for removing trend from the output to obtain ZFFS in lengthy speech files. The length of files recorded can be of several minutes. Whereas, ZBF being a stable filter has no such issues when the length of speech files runs for several minutes. This is demonstrated in Figure 3.7, where, (a) shows the speech waveform of 5 minutes in length, while (b) and (c) shows ZFFS and ZBFS, respectively for the same speech signal. The speech file is taken from NIST 2012 mic database [107]. It is observed that the ZFFS is noisy above 1.6 minutes. This is further demonstrated in Figure 3.8, where a segment of voiced speech region

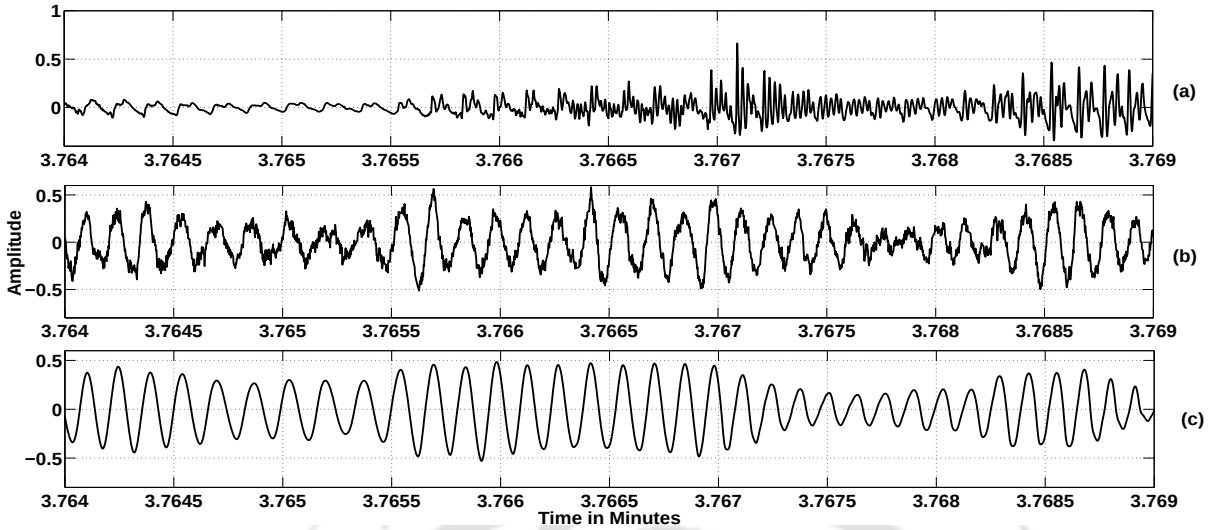

Figure 3.8: Illustration of robustness of ZBF by selecting a segment of speech from a lengthy file (a) Voiced speech segment expanded from Figure 3.7(a), (b) Zero Frequency Filtered output and (c) Zero Band Filtered output for the voiced speech segment shown in (a).

is chosen from the same speech waveform that is shown in Figure 3.7 (a), while Figures 3.8 (b) and (c) show ZFFS and ZBFS of the selected segment of voiced speech region, respectively. The voiced segment is selected from the time index of 3.764 to 3.769 minutes of the speech file. It can be observed that ZFFS output appears noisy and this may lead to spurious detection of epoch locations because of multiple positive zero crossings. While, the ZBFS output has no such issues.

3.3 Experiments and Results

The evaluation procedure is carried using clean CMU-Arctic database and subsequently, by degrading the clean speech by different kinds of additive noises at different levels taken from Noisex database. The CMU-Arctic database is freely available from Festvox website [108]. The database consists of 2 male and 1 female speakers. All have simultaneous recordings of electroglottograph (EGG) along with speech in two different channels. In order to identify the glottal closure instants from EGG signal, SIGMA algorithm [109] is used that is available in voicebox and such instants are used as reference markings against which the performance of different methods are compared. The database having BDL-US male, JMK-Canadian male and

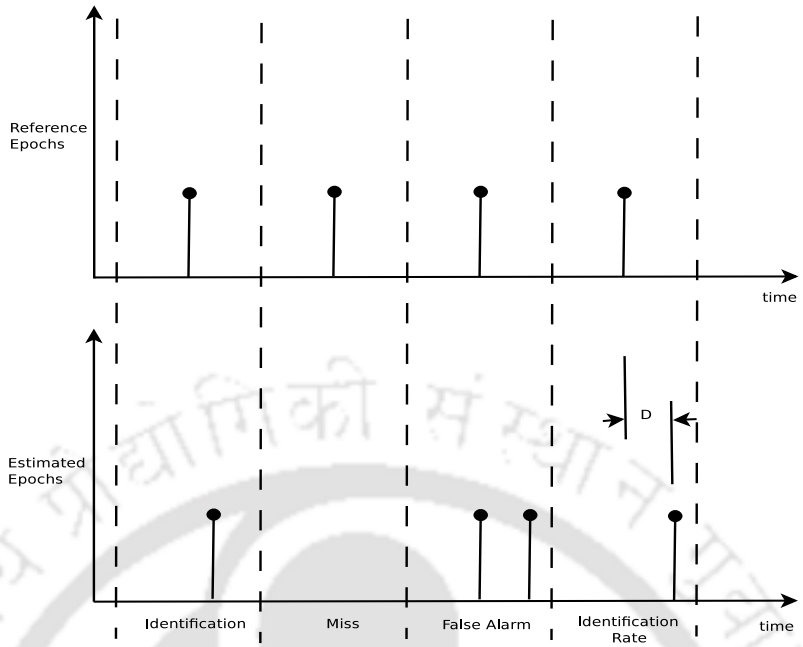

Figure 3.9: Characterization of epoch location estimates showing 4 larynx cycles with examples of each possible outcome from epoch estimation [1]. Identification accuracy is measured by D.

SLT-US female has a total of 1132 phonetically balanced sentences recorded in a closed booth in a controlled environment at the sampling rate of 32 kHz. The performance was evaluated considering the voiced regions from EGG.

The proposed method is evaluated and compared with DYPSA [1], ILPR [71], SEDREAMS [103], YAGA [104] and ZFF [70] in terms of identification rate (IDR), miss rate (MR), false alarm rate (FAR) and identification accuracy (IDA). These parameters are computed as shown in Figure 3.9 [1]. The parameters are computed with reference to glottal cycle derived from EGG as reference and is given by $1/2(c_{i-1} + c_i) \leq n \leq 1/2(c_i + c_{i+1})$, and the cycle is defined with respect to i^{th} epoch as reference, denoted by c_i. IDR represents the percentage of times a single epoch located from method under evaluation within a cycle, MR is when no epochs are located within a cycle and FAR is the number of times multiple epochs detected within a cycle in the database. However, D represents the error from the actual location of the epoch when exactly one epoch is detected by the method under evaluation and IDA is the standard deviation of error D.

The lengthy speech waveforms are simulated by concatenating many speech files from CMU-

Table 3.1: Performance comparison of epoch extraction methods under clean condition on CMU-Arctic database. IDR - Identification Rate, MR - Miss Rate, FAR - False Alarm Rate, IDA - Identification Accuracy

Method	IDR (%)	MR(%)	FAR(%)	IDA (ms)	IDR (± 0.25 ms)
DYPSA	96.66	1.76	1.58	0.59	52.46
ILPR	98.64	0.62	0.71	0.29	86.91
SEDREAMS	97.87	1.14	1.07	0.39	82.59
YAGA	98.71	0.48	0.79	0.31	90.16
ZBF	98.20	0.72	1.06	0.39	86.49
ZFF	99.04	0.18	0.77	0.36	91.26

Arctic database to evaluate the performance of ZFF and ZBF on lengthy speech files.

3.3.1 Performance Evaluation on Clean and Degraded Speech

Table 3.1 shows the results evaluated for clean speech data taken from the CMU-Arctic database. The performance is evaluated in terms of IDR, MR, FAR and IDA. The performance of the proposed method is compared with five other state of the art methods. The DYPSA algorithm relies on group delay function which is the average slope of the phase spectrum derived from Short-Time Fourier Transform (STFT) of LP residual. The negative zero crossings of group delay function correspond to epoch locations. However, it is not always guaranteed that the epochs always leads to negative zero crossings in the group delay function. The phase slope projection technique is adopted in order to project the slope in between maxima and minima in order to identify the epochs location. Furthermore, the identified epoch locations are pruned based on the distance criteria of adjacent epoch locations detected using dynamic programming [1]. As a further improvement over DYPSA, an efficient method to detect epoch locations was suggested using group delay function derived from the LP residual signal. The LP residual signal is derived using non-pre-emphasized speech signal rather than commonly used pre-emphasized speech. However, the method makes use of multi-scale product signal

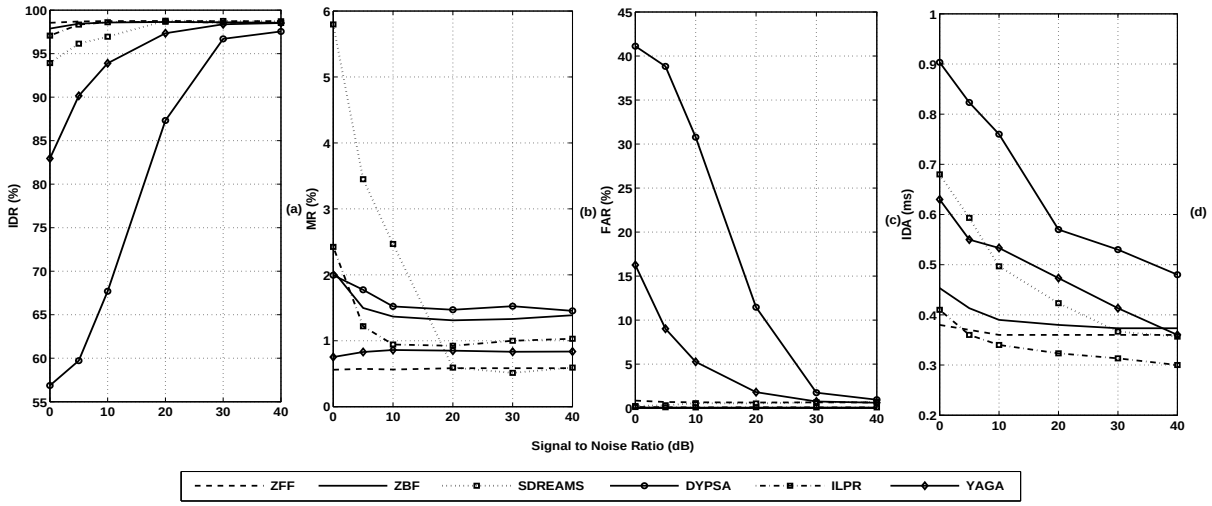

Figure 3.10: Performance comparison of epoch extraction under degraded conditions by adding white from Noisex-92 database to CMU-Arctic database (average scores of 3 different speakers) at 40, 30, 20, 10, 5 and 0 dB levels (a) IDR (b) MR (c) FAR (d) IDA for additive white noises .

derived from wavelets to obtain a train of the impulse-like signal. The impulse locations are identified by the zero crossings of group delay function obtained from LP residual signal that is further pruned by N-best dynamic programming [104]. Recently a method was proposed based on the mean based signal that is derived directly from the speech signal. The mean based signal is derived based on 1 - 2 times the pitch period as a first step to locating the epochs. However, the epoch locations are further refined using residual excitation for accurate estimation in second step in [103]. More recently a method was proposed based on inverse filtering of non-pre-emphasized speech called as integrated linear prediction residual (ILPR) using which the half-wave rectified ILPR is derived. Such a signal closely corresponds to train of impulses and closely match with epoch locations. In order to identify such locations, a temporal measure is adopted to detect the locations of transients called dynamic plosion index (DPI) [71].

The results are shown in Table 3.1 are the average scores. It can be noticed that the overall IDR and IDR within ± 0.25 ms are better in the case of ZFF and YAGA methods compared to other methods. However, the performance of the proposed method is comparable with

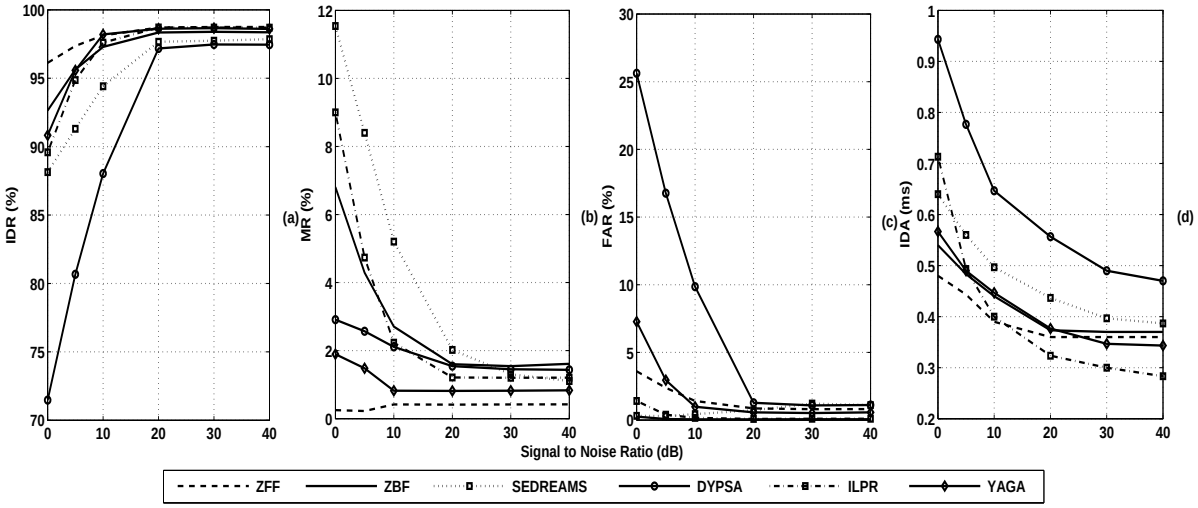

Figure 3.11: Performance comparison of epoch extraction under degraded conditions by adding babble from Noisex-92 database to CMU-Arctic database (average scores of 3 different speakers) at 40, 30, 20, 10, 5 and 0 dB levels (a) IDR (b) MR (c) FAR (d) IDA for additive babble noises .

ILPR in terms IDR and IDR within ±0.25 ms. The reason for lower IDR and IDA of the proposed method compared to ZFF may be due to relatively larger passband of ZBF at 0 Hz. In terms of IDA, ILPR based method is better with an average deviation of 0.29 ms compared to other methods. Both SEDREAMS and ZBF are comparable in terms of IDA. Furthermore, to evaluate the robustness of the proposed method, the performance is computed by adding different kinds of noises to clean speech files taken from CMU-Arctic databases. The results are evaluated for 2 different noisy conditions, namely white and babble noise taken from Noisex-92 database [110]. Noise is added to clean speech at 6 different levels of 40, 30, 20, 10, 5 and 0 dB.

Figures 3.10 (a) - (d) shows the performance evaluation of DYPSA, ILPR, SEDREAMS, YAGA, ZFF and ZBF for degraded conditions by adding white noise at different levels to clean speech files in terms of IDR, MR, FAR, and IDA. The scores plotted in Figure 3.10 are the average scores obtained from all speakers considered. It can be noticed that the performance of ZFF, ZBF , SEDREAMS and ILPR based methods are highly robust to additive white noise in terms of IDR. However, there is a considerable decrease in the performance of YAGA in terms of IDR for higher levels of noise added. Furthermore the IDA of ZFF, ZBF and ILPR

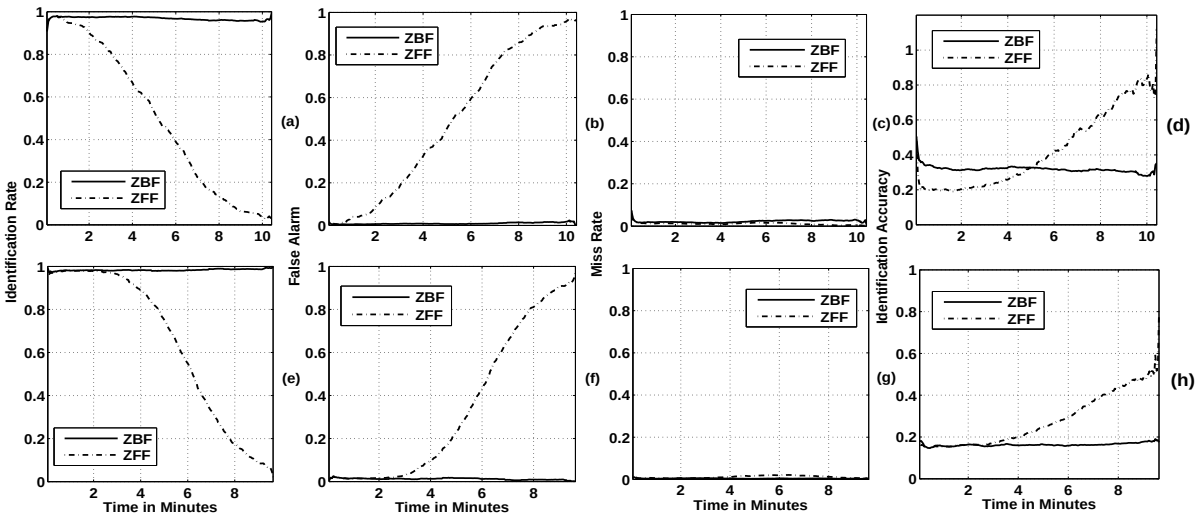

Figure 3.12: Performance evaluation of ZFF and ZBF on concatenated lengthy speech waveform generated from 200 files from CMU-Arctic database (a) Identification Rate, (b) False Alarm Rate, (c) Miss Rate and (d) Identification Accuracy, for a male speaker, (e) Identification Rate, (f) False Alarm Rate, (g) Miss Rate and (h) Identification Accuracy, for a female speaker.

remains robust for higher levels of white noise added. It can be observed that the performance of DYPSA is the most affected by the degradation of speech files by additive white noise in terms of IDR and IDA. One of the important types of noise that can interfere with the speech signal is babble noise and it is a challenge for any speech processing method as the noise can be spectrally similar to speech signal itself. Hence, the performance of different methods is evaluated for additive babble noise at various levels. Figures 3.11 (a) - (d) shows the performance evaluation of different methods at 40, 30, 20, 10, 5 and 0 dB levels. It can be observed that the performance of ZFF, ZBF, ILPR and YAGA are relatively robust at higher levels of additive babble noise compared to other methods. Though there is a slight degradation of performance of ZBF relative to ZFF at higher levels of additive babble noise, the performance of the proposed method remains robust for high levels of additive babble noise relative to other methods. The overall performance of ZBF remains robust for additive white and babble noises at different levels.

3.3.2 Performance Evaluation on Lengthy Speech Waveform

In order to demonstrate the reliability of the proposed method for lengthy speech cases, two different lengthy speech data of approximately 10 minutes were considered for each of a male and a female speaker from the CMU-Arctic database. The lengthy speech data were simulated by concatenating 200 speech files added with 20 dB white noise from each of the speaker's database, respectively. Since each of the speech files in the CMU-Arctic database is approximately 3 s in length, 200 files are concatenated to form an approximately 10 minutes speech data. Such a lengthy speech data is passed through ZFF and ZBF to obtain ZFFS and ZBFS, respectively. With prior knowledge of time stamps for each of the individual speech files, the performance is evaluated in terms of IDR, MR, FAR and IDA using EGG data as a reference. The results obtained are shown in the form of graphs in Figures 3.12 (a) - (h). Figures 3.12 (a), (b), (c) and (d) shows the plots of IDR, FAR, MR and IDA, respectively for male speaker's data and similarly Figures 3.12 (e), (f), (g) and (h) are the plots of IDR, FAR, MR and IDA, respectively for female speaker's data. The scores are normalized and smoothed by moving average filter for clarity.

As it can be noticed that the IDR exponentially falls in case of ZFF with an increase in the length of speech. However, the performance of proposed method is robust for arbitrary lengths of data. The reason for decreased IDR in the case of ZFF can be accounted by an exponential increase in the FAR due to spurious detection of positive zero crossings in ZFFS as mentioned earlier. However, it can be observed that MR does not increase with the length of speech waveform in the case of ZFF. Also, IDA decreases with the length of speech waveform in the case of ZFF, while ZBF remains robust. It can be observed that the performance of ZFF decreases approximately starting from 0.8 minutes in case of the male speaker, while it starts decreasing from 3 minutes in case of a female speaker. This may be explained from Eqn. (3.20) for $r = 1$ that the integrator output depends on the nature of input signal along with the increase in time index.

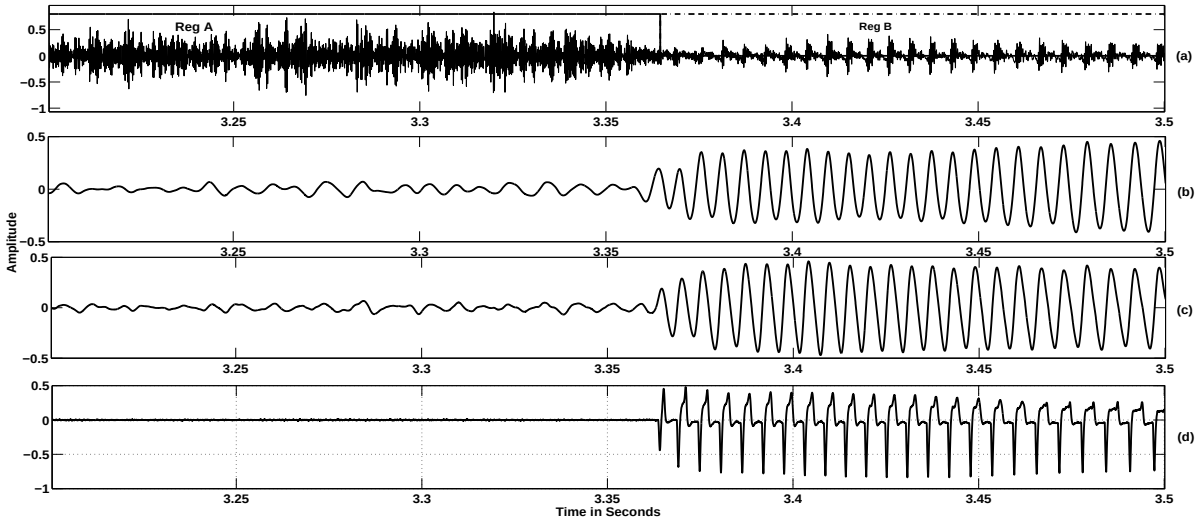

Figure 3.13: Robustness of ZFFS and ZBFS for background noise (a) Speech segment from a noisy background, where Region A (marked by solid line) consists of only background noise and Region B (marked by dotted line) consists of both background noise and foreground speech. (b) Zero Frequency Filtered output for speech waveform shown in (a), (c) Zero Band Filtered output for speech waveform shown in (a), (d) difference EGG which acts as reference epoch locations.

3.4 Robustness of Epoch Extraction using ZBF

The robustness of ZBF is further demonstrated in cases of foreground speech [32], singing voices and emotional speech. Although ZBF allows a narrow band of frequencies around 0 Hz relative to ZFF, it is shown that ZBF is equally robust to additive noises. However, the nature of degradation of speech collected from natural environments is different from additive noises. Hence, the performance of different methods is compared for speech files collected from different natural environments. Also, ZBF does not require *a priori* pitch estimation and it is of interest to evaluate the performance for varying pitch scenarios as in cases of singing voices and emotional speech.

3.4.1 Epoch Extraction in Foreground Speech

It is shown in Section 3.3.1 that ZBF is robust for epoch extraction when the speech signal is degraded by additive noise. However, it is interesting to study the performance of ZBF for epoch extraction in foreground recording scenario. Even though ZBF allows a narrow band of

Table 3.2: Performance evaluation of epoch extraction methods using data collected naturally from different noisy environments. IDR - Identification Rate, MR - Miss Rate, FAR = False Alarm Rate, IDA - Identification Accuracy.

Method	IDR (%)	MR(%)	FAR(%)	IDA (ms)
DYPSA	82.24	13.22	3.59	0.29
ILPR	89.92	8.54	1.63	0.31
SEDREAMS	73.43	21.29	5.34	0.39
YAGA	91.16	1.24	7.81	0.23
ZBF	92.62	0.43	6.98	0.37
ZFF	91.49	2.31	6.27	0.26

frequencies around 0 Hz to pass through relative to ZFF, this does not impact on the ability of ZBF to extract epoch locations in foreground speech regions. It can be observed from Figure 3.13 that both ZFFS and ZBFS are almost similar in their performance, where Figure 3.13 (a) is a speech segment from a naturally recorded speech in noisy environment and Figure 3.13 (b) and (c) are ZFFS and ZBFS of a speech segment shown in Figure 3.13 (a), respectively. It can be observed in Figure 3.13 (a) that there is presence of background noise throughout the speech segment, *Region A* consists of only background region while *Region B* consists of both background noise and foreground speech. Also, EGG is recorded in the parallel channel to provide the ground truth of epoch locations, where, Figure 3.13 (d) shows the plot of difference EGG signal that acts as a reference.

In order to demonstrate the robustness of the proposed method for *foreground speech* under degraded conditions, speech was collected at different noisy environments along with simultaneous recordings of EGG in the parallel channel. Data was collected from 9 male and 4 female speakers. The recording set up had a laptop, headphone and EGG electrodes carried to respective locations to record the data from different subjects. Locations included mechanical workshop, home (TV switched on with loud volume), generator room, traffic, dining hall and vehicle noises. The performance was evaluated using EGG as the ground truth and it is given in Table 3.2. A total of 26366 epochs were present in the data collected. It can be observed

from Table 3.2 that the performance of ZBF is significantly better in terms of IDR compared to SEDREAMS, DYPSA and ILPR methods and comparable to ZFF and YAGA methods. Furthermore, it can be noticed that the performance of ZBF is slightly better to ZFF in terms of IDR. It can be noticed from Eqn. (3.4) that the ZFF output is sensitive to accurate estimation of pitch period from the speech signal. However, it is a challenging task to correctly estimate pitch period in noisy cases and this impacts the results of ZFF in noisy cases. Since the ZBF is independent of pitch period estimation and this may explain the reason for better IDR in the case of ZBF relative to ZFF. However, YAGA, ZFF, and DYPSA are better in terms of IDA compared to other methods.

3.4.2 Varying Pitch Scenarios

3.4.2.1 Epoch Extraction in Emotional Data

Epoch extraction in emotional speech data is of significant interest amongst speech synthesis community. It is shown that the excitation source information features derived using epoch extraction methods are significantly different across different emotions. Also, due to large variations in pitch, epoch extraction is a challenging task [111]. In order to demonstrate the robustness of the proposed method for epoch extraction in such emotional data, different emotional speech files containing angry, disgust, fear and happy emotions were chosen from German emotional speech database for evaluation [112]. The database consists of both speech and parallel EGG recordings that are used for reference marking of epochs.

3.4.2.2 Epoch Extraction in Singing Voice

Epoch extraction in singing voice is challenging due to variations in pitch. In order to evaluate the performance of epoch extraction methods, a singing database was created from 10 different singers consisting of 5 male and 5 female professional singers. All the singers had an average experience of approximately 15 years of practicing Indian Hindustani classical singing and their experiences varied from 5 to 30 years. Each of those singers has sung 2 different songs in order to capture the variations in the pitch. The songs recorded are of different “raag”, “taal” and scale, also the lyrics of each of the songs were different and they consist of Hindi and

Table 3.3: Performance evaluation of epoch extraction by ZFF and ZBF using angry, disgust, fear and happy emotional speech files from German emotional database. IDR - Identification Rate, MR - Miss Rate, FAR = False Alarm Rate, IDA - Identification Accuracy.

Emotion	Method	IDR (%)	MR(%)	FAR(%)	IDA (ms)
	DYPSA	83.32	4.59	12.08	0.54
	ILPR	85.88	10.62	3.50	0.49
	SEDEREAMS	61.19	20.08	18.73	0.51
Angry	YAGA	90.93	4.43	4.64	0.46
	ZBF	91.73	1.49	6.77	0.44
	ZFF (Non Adaptive)	88.96	0.05	10.97	0.35
	ZFF (Adaptive)	92.81	0.00	7.19	0.37
	DYPSA	84.03	7.26	8.71	0.52
	ILPR	89.73	5.22	5.05	0.38
	SEDEREAMS	71.25	16.33	12.42	0.49
Disgust	YAGA	86.72	5.39	7.88	0.43
	ZBF	93.26	3.57	3.17	0.42
	ZFF (Non Adaptive)	85.51	7.26	8.71	0.34
	ZFF (Adaptive)	95.74	1.49	2.77	0.38
	DYPSA	84.61	9.31	6.08	0.55
	ILPR	87.08	8.41	4.51	0.45
	SEDEREAMS	64.70	24.33	9.97	0.58
Fear	YAGA	87.84	6.61	5.55	0.41
	ZBF	94.46	1.18	4.36	0.48
	ZFF (Non Adaptive)	88.01	3.26	8.73	0.34
	ZFF (Adaptive)	96.20	2.33	1.73	0.44
	DYPSA	85.77	9.48	4.76	0.49
	ILPR	84.03	12.58	3.40	0.51
	SEDEREAMS	70.25	17.10	12.65	0.50
Happy	YAGA	90.01	4.96	5.03	0.43
	ZBF	92.60	3.23	4.16	0.44
	ZFF (Non Adaptive)	87.43	3.19	9.38	0.33
	ZFF (Adaptive)	94.47	1.55	3.98	0.39
	DYPSA	84.30	7.66	7.9	0.52
	ILPR	86.68	9.20	4.00	0.45
	SEDEREAMS	66.84	19.46	13.44	0.52
Average	YAGA	88.87	5.34	5.77	0.43
	ZBF	93.01	2.36	4.61	0.44
	ZFF (Non Adaptive)	87.47	3.44	9.44	0.34
	ZFF (Adaptive)	94.80	1.34	5.66	0.39

Table 3.4: Performance evaluation of epoch extraction by ZFF and ZBF using singing database. IDR - Identification Rate, MR - Miss Rate, FAR = False Alarm Rate, IDA - Identification Accuracy.

Method	IDR (%)	MR(%)	FAR(%)	IDA (ms)
DYPSA	71.90	26.00	2.09	0.26
ILPR	73.56	26.01	0.40	0.18
SEDREAMS	69.95	21.86	8.19	0.31
YAGA	81.50	16.34	2.16	0.23
ZBF	83.08	14.59	2.43	0.29
ZFF (Non Adaptive)	80.24	14.39	5.36	0.22
ZFF (Adaptive)	84.76	13.73	2.51	0.26

Assamese languages. The singers were made to wear EGG electrodes around the neck along with headphone to capture EGG data along with singing voice, while the songs were recorded in a professional studio environment. The portions of the songs were selected that had significant variations in the pitch and all methods were evaluated using 17 minutes of singing files that consisted of 201808 epochs.

In order to evaluate the performance of ZFF in the case of varying pitch scenarios, two different versions of ZFF are used in the current study. Firstly, in the case of non-adaptive ZFF, the average pitch period is measured for the entire speech file that is used as an input parameter for trend removal procedure. However, measuring average pitch for the entire speech file may not be suitable for varying pitch scenarios that can result in spurious positive zero crossings or missing the instants of significant excitation and therefore requires pitch estimation on a short-term basis. Consequently, the updated average pitch ($F0$) is computed using ZFFS for every 20-30 ms of non-overlapping frames by picking the maximum peak from STFT of the corresponding frame [111, 113]. Furthermore, the updated $F0$ is used as an input to design a low pass filter from which the modified ZFFS signal is obtained and is called as adaptive ZFFS [111]. It can be observed from Tables 3.3 and 3.4 that the performance of ZBF is significantly better than SEDREAMS, ILPR and DYPSA methods in terms of IDR. However,

the performance of ILPR, YAGA and ZFF are better in terms of IDA. Furthermore, it can be observed from Tables 3.3 and 3.4 that the performance of ZBF is better than non-adaptive ZFF in terms of IDR while ZFF is having better IDA. The reason for better IDR in the case of ZBF compared to non-adaptive ZFF may be because the parameters are fixed and they are not dependent on *apriori* $F0$ estimation, while the poor estimation of $F0$ can lead to an increase in MR or FAR. This is evident from Tables 3.3 and 3.4 that there is an increase in MR. The miss detection is eliminated significantly using adaptive ZFF and as a result, there is a significant improvement in terms of IDR. However the performance of ZBF without *apriori* pitch estimation is significantly better than SEDREAMS, ILPR, and DYPSA for varying pitch scenarios such as emotional and singing voice.

3.5 Summary

In this chapter, a BIBO stable realization of ZFF is proposed and the advantages of using a stable filter are that the method does not require *a priori* estimation of pitch. Furthermore this method eliminates the necessity of removing trend from the filter's output to obtain ZBFS. The method is validated using CMU-Arctic for both clean and degraded conditions. It is found that the performance of ZBF is comparable to ZFF. Also, the method is robust for lengthy files recorded for several minutes without any precision related problems associated with the filter output. In order to demonstrate the robustness of ZBF for lengthy files, several speech files of a CMU-Arctic database are concatenated to form a lengthy file to which the performance was evaluated. It is found that the performance of ZBF does not vary with the length of the files. Furthermore, the robustness of ZBF is demonstrated in cases of emotional and singing voices, where, the performance is comparable to adaptive ZFF, while it is significantly better than SEDREAMS, ILPR, and DYPSA for varying pitch scenarios. The performance of ZBF is comparable to ZFF in the case of foreground speech recording scenario in presence of other interfering sources.

In this chapter, the speaker is closest to the microphone, however, it calls for a detailed analysis of ZBF characteristics with the varying distance between speaker and microphone. It

is necessary to establish the definition of foreground speech and background noise. Also, it is important to temporally segment the relevant foreground speech from rest of background regions for further enhancement and these works are addressed in next chapter.

4

Foreground Speech Analysis and Segmentation

Contents

Objective

The signal characteristics of speech collected over microphone depend on the distance between the speaker and sensor, and also on the presence of other background acoustic sources. Even though the presence of other background acoustic sources affect the signal characteristics in both cases, the foreground scenario offers some advantages due to the proximity of the desired speaker to the microphone which results in the better manifestation of speech characteristics. The fundamental question is when the collected speech is termed as foreground speech? In this work a definition of foreground speech is established by considering glottal closure instants (GCIs) as the basis. The GCIs are the instants of significant excitation and the regions around them are least affected by other interfering sources compared to other regions of the speech signal. Hence, an attempt is made to substantiate foreground speech based on the characteristics of GCIs.

The signal analysis requires to deal with all frequency components and this can be a complicated task. Alternatively, if the nature of speech and degradation can be observed at a particular frequency, then a method can be developed by analyzing only the characteristics of that particular frequency component. The zero band filtering (ZBF) allows only the signal components around zero frequency and significantly attenuates all other frequency components. Hence, the analysis of zero band filtered signal (ZBFS) characteristics can be helpful to derive the definition of foreground speech. Further, the nature of ZBFS can be used to derive certain features to segment the relevant foreground speech from rest of the background content. In this chapter, the excitation source based features derived from ZBFS are combined with vocal tract information based feature to segment foreground speech. Since the objective of foreground speech segmentation (FGS) is similar to voice activity detection (VAD), the performance of foreground speech segmentation is evaluated along with other considered state of the art methods.

4.1 Overview of Foreground Speech

The speech-based applications like speech recognition and speaker verification systems can be accessed from remote locations. Hence, most of the times the recording environment is not under the user's control and can have other interfering sources along with the desired speaker's speech. The other interfering sources are categorized as background noise and their presence affects the perceptual quality and intelligibility of speech [114]. Such speech signal can be annoying for listening over longer durations. The presence of background noise decreases the performance of speech and speaker recognition tasks. In most scenarios, the desired speaker will be close to the microphone sensor relative to other interfering sources that may include background speakers. The speech signal that is recorded from closely speaking desired speaker's speech, hereafter termed as *foreground speech* has unique characteristics compared to other interfering sources that are farther away from microphone [32]. It may, therefore, be beneficial to study such attributes of foreground speech to derive features for further processing. However, the fundamental question is when do we call the collected speech as *Foreground speech*? It is essential to establish the definition of foreground speech. After this, it becomes easier to distinguish *distant speech* signal based on certain characteristics of speech signal [115]. The current work conducts a systematic study to establish the definition of foreground speech. After this, the characteristics of foreground speech are explained by performing short term analysis.

There are methods available in the literature that can distinguish between degraded speech regions from rest of the background noise based on some statistical assumptions or signal characteristics of interfering noise [5]. Since the recording happens in natural environments, there can be several kinds of interfering sources and it is difficult to characterize all types of noise sources. Furthermore, the performance of such systems suffers when the interfering source is unseen. Though it is possible to explicitly model the noise signal using multiple sensors, the focus of current work remains single sensor recording scenario. Hence, it is required to evolve methods that mainly relies on features derived from foreground speech production aspects

rather than relying on noise characteristics.

The studies reveal that not all foreground speech regions are equally affected by interfering noise. The small regions around the instants of significant excitation are robust to such interfering noises and they form high signal to noise ratio (SNR)/ signal to reverberation ratio (SRR) regions [40]. Hence, it is beneficial to use instants of significant excitation as anchor points to identify such foreground speech regions and to further enhance those regions. There are some studies reported in the literature that use instants of significant excitation to characterize distant speech [113]. However, it will be interesting to study the effect on recorded speech signal when the desired speaker is at varying distances from the microphone sensor. Though the regions around epoch location remain least corrupted in the speech signal, nevertheless, the SNR levels around such regions may decrease due to the influence of other interfering sources as the distance between speaker and microphone increases. It may be possible to evolve a definition for foreground speech on the basis of epoch extraction.

The rest of the chapter is organized as follows: Section 4.2 describes the details of speech data collection in the natural environment for analysis. The Section 4.3 explains different methods applied for foreground speech analysis based on which a definition for foreground speech is derived. Section 4.4 describes the details of foreground speech segmentation using excitation source and vocal tract based features. Section 4.4.4 describes the details of the evaluation procedure and results obtained. The summary of the present work is mentioned in Section 4.5.

4.2 Speech Data Collection for Analysis

The speech signal is recorded using 5 male and 5 female speakers in a semi controlled environment like an AC machine room in which the AC machine can be turned ON or OFF using a switch. The speech data is collected in an Air Conditioner (AC) machine room along with Electroglottograph (EGG) signal recorded simultaneously. The Figure 4.1 shows the AC machine room in which the speech signals were recorded from different speakers. However, the recording environment is a semi controlled environment in which the AC machine can be turned

Figure 4.1: Speech recording in AC machine room along with Electroglottograph signal.

ON or OFF using a switch. Also, further switching ON the AC machine introduces a significant amount of background noise that can interfere with the speech signal. It can be noticed that the AC machine room door is open and there is no resistance for other acoustic sources outside the room. The room shape is rectangular and it's size is approximately 15 m × 7 m × 12 m (length × breadth × height). This can be one such typical natural environment in which the speech based applications may need to perform. In order to capture the reference source information that is used for evaluation, the speaker is made to wear Electroglottograph (EGG) electrodes around the neck to simultaneously record EGG signal along with speech signal.

The recording setup is as shown in Figure 4.1, where, a simple headphone and EGG electrodes are connected to the laptop through EGG device. It may be noted that the study is limited to close speaking microphone and may not be necessarily extended to omnidirectional microphone scenario. The speech and EGG signals are recorded using wavesurfer [116] in a laptop from EGG device connected through a universal serial bus (USB). The speech and EGG signals are recorded at 48 kHz sampling rate and later downsampled to 8 kHz for analysis. In order to maintain the gender balance, the speech and EGG signals are recorded from 5 male and 5 female speakers. Since the objective is to study the characteristics of speech signal when

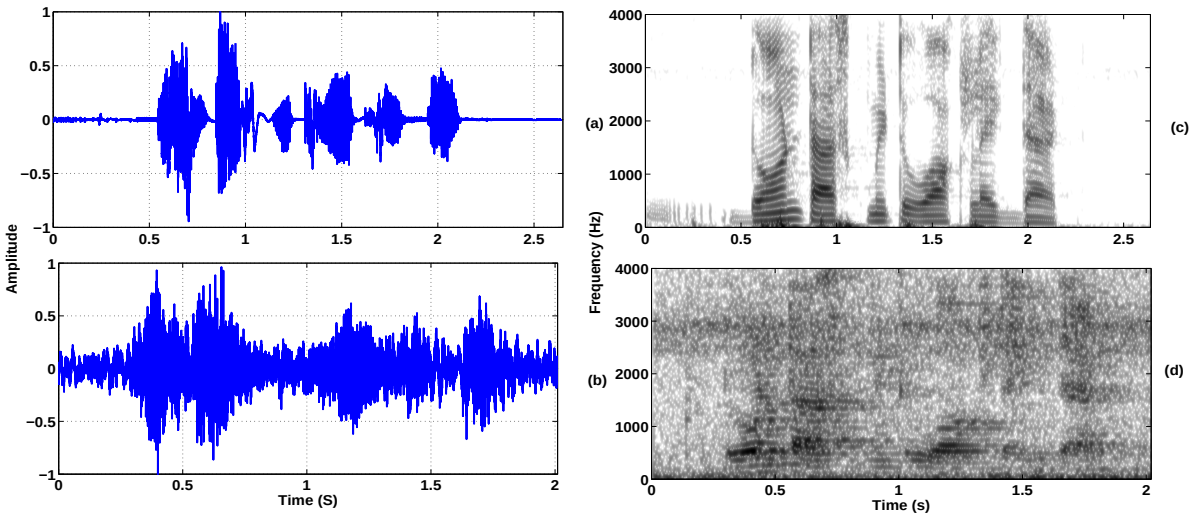

Figure 4.2: Illustration of speech signal recorded (a) at 1 inch from speaker (c) corresponding magnitude spectrum (b) at 30 inches from speaker (d) corresponding magnitude spectrum.

recorded at varying distances between the speaker and the microphone, the speech signals are recorded using the headphone microphone with varying distances from mouth to the sensor. The microphone sensor is kept at varying distances from 1 - 30 inches at the resolution of 1 inch. Though the scale is used to measure the distance between speaker and microphone, it can be noted that the microphone is handheld and hence the distance can be accounted in an approximate sense. The same holds good for the elevation that accounts for the height of microphone from the ground level. However, the speaker is asked to utter a fixed sentence *"don't ask me to walk like that"* and all such 30 utterances at varying distances are recorded in a single session to rule out any session variability in the study. Also, the elevation of the microphone is maintained to be at the same height as that of the speaker's mouth. The constant angle of recording with respect to mouth was maintained throughout the recordings to alleviate the effects of changes in recording the utterances. The recording environment can be controlled in AC machine room to simulate relatively clean and noisy environments by switching AC machine OFF and ON, respectively. For analysis, two sets of speech data are collected when AC machine is switched OFF and ON, respectively.

4.3 Foreground Speech Analysis

The Figure 4.2 illustrates the temporal and spectrogram plots of speech utterances recorded when the microphone is placed at distances of 1 inch and 30 inches from the mouth of the speaker, respectively. The AC machine is switched OFF while recording such utterances from a male speaker. It can be noticed that closely recorded speech signal is relatively less corrupted by other interfering sources compared to the speech signal recorded at a distance of 30 inches as shown in Figure 4.2 (d). Also, further it can be observed that the formant and harmonic structure is intact in case of close speaking scenario, while, it appears to be smeared in the case of distant recording. This is understandable as the speaker is farther away from the microphone and hence there is a reduction in SNR levels that can affect the temporal and spectral characteristics.

There are some studies reported in the literature to compensate for the effect of distance from speech signal which is useful in distant speech processing [113]. The current work focuses on accounting for differences in temporal and spectral characteristics of the speech signal for varying distances. However, it is difficult to characterize the speech signal that is corrupted by interfering sources in its original form directly. Rather, it is easier to analyze the speech signal by separating the source and vocal tract filter response. One of the important attributes of source feature in the speech signal is the presence of epochs. However, most of the epoch extraction methods proposed are reliable when the speaker is speaking closer to microphone even with the presence of other interfering sources [95]. Hence, such epoch extraction methods can form the basis of which the foreground scenarios can be distinguished from distant speech.

In the current work, zero band filtering is used to characterize the speech signals recorded at varying distances in terms of epoch extraction performance. The epochs present in speech signal are train of impulses in time domain and they are instants of significant excitation. They can get located by passing the speech signal through zero band filter (ZBF) [117]. The negative to positive zero crossings of zero band filtered signal (ZBFS) corresponds to epoch locations. The ZBFS of the speech signal $s(n)$ that contains epochs is defined by the following relations.

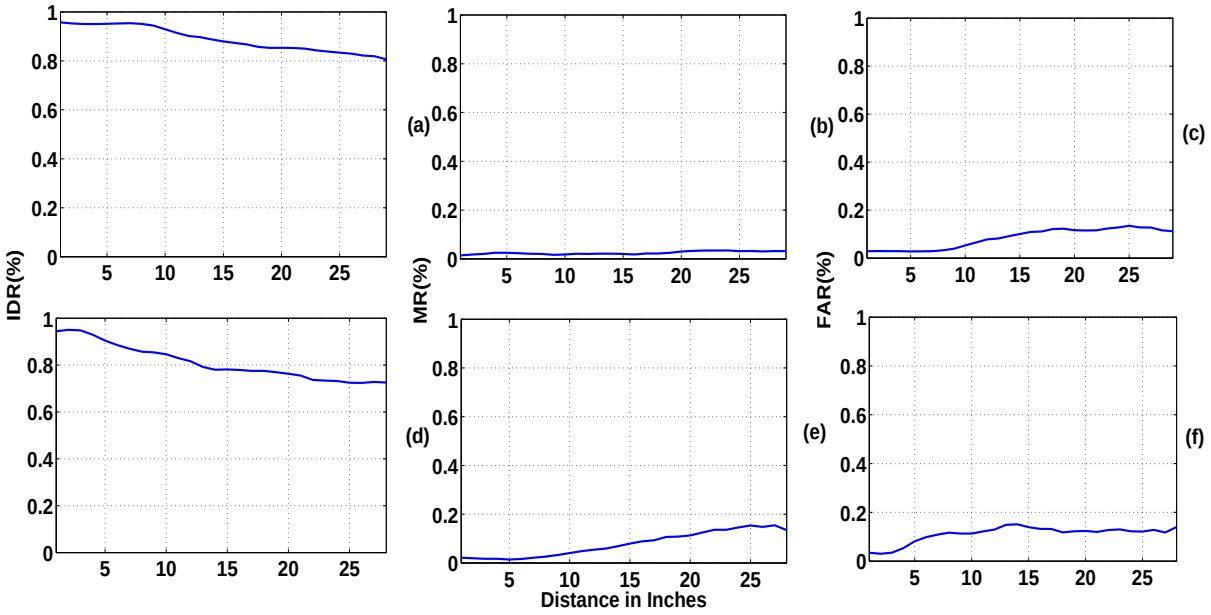

Figure 4.3: Performance evaluation of epoch extraction at varying distant speech recordings when speech files are recorded in an relatively clean environment and noisy environment (a) IDR (b) MR (c) FAR when speech signals are recorded in clean environment (d) IDR (e) MR (f) FAR when speech signals are recorded in noisy environment.

The first order difference of the speech signal is given by

$$x(n) = s(n) - s(n-1) \tag{4.1}$$

The difference signal $x(n)$ is to de-emphasize low frequency fluctuations and emphasize high frequency components present in signal and is passed twice through the transfer functions as given in [117]

$$H_1(z) = \frac{Y_1(z)}{X(z)} = \frac{1}{1 - 2rz^{-1} + r^2z^{-2}} \tag{4.2}$$

$$H_2(z) = \frac{Y_2(z)}{Y_1(z)} = \frac{1}{1 - 2rz^{-1} + r^2z^{-2}} \tag{4.3}$$

where, r represents the value of radius on unit circle in z-plane at which the poles are placed, and r value should satisfy $0 < r < 1$ for stability and its value is taken as 0.99 in ZBF. In time domain, the filter outputs can be represented in the form of summation series given by

$$y_1(n) = \sum_{k=0}^{\infty}(k+1)r^k d(n-k) \tag{4.4}$$

$$y_2(n) = \sum_{k=0}^{\infty} (k+1) r^k y_1(n-k) \tag{4.5}$$

The output $y_2(n)$ is filtered using a Butterworth 4^{th} order high-pass filter having a cutoff frequency 80 Hz in order to attenuate low-frequency fluctuations. The resulting ZBFS is used to extract epoch locations. As demonstrated in Section 3.2 the positive zero crossings of ZBFS help to locate epoch locations. The ZBFS is used to extract the epochs from the speech signal and is validated against the reference epoch locations obtained from EGG signal using SIGMA [109] algorithm. The epoch extraction performance is validated in terms of IDR, MR, and FAR with reference to each glottal cycle obtained from EGG signal. The Figure 4.3 shows the epoch extraction performance for two sets of speech recordings at varying distances, each obtained when AC machine was switched OFF and ON, respectively. It can be noticed that IDR gradually decreases as the distance increases between the speaker and microphone. This can be attributed to the fact that there is an overall drop in SNR levels throughout the speech regions including the regions around epochs. This can be due to other interfering sources when the speaker is distant away from the microphone. Though the regions around epoch locations are least affected in such distant recordings, still there is a reduction in SNR levels compared to speech recordings when the speaker is closer to the microphone.

The Figure 4.3(a) shows the epoch extraction performance on speech signals that are recorded at varying distances when AC machine is switched OFF. Also, further it can be noticed from Figure 4.3(d) that the decrease in epoch extraction performance is steeper when AC machine is switched ON. It can be observed that there is a steeper increase in MR and FAR as the distance between speaker and sensor is above 3 inches. Hence, it is reliable to extract epochs when the distance between speaker and microphone is well within 1-3 inches. The desired speaker's speech regions recorded at such distances of 1 - 3 inches is therefore termed as *foreground speech*. One of the important characteristics of foreground speech signal is that the epochs can be reliably extracted even in the presence of interfering background noise. Also, the speaker is closer to the microphone in foreground scenario, while rest of the background noise sources are relatively far away from the microphone.

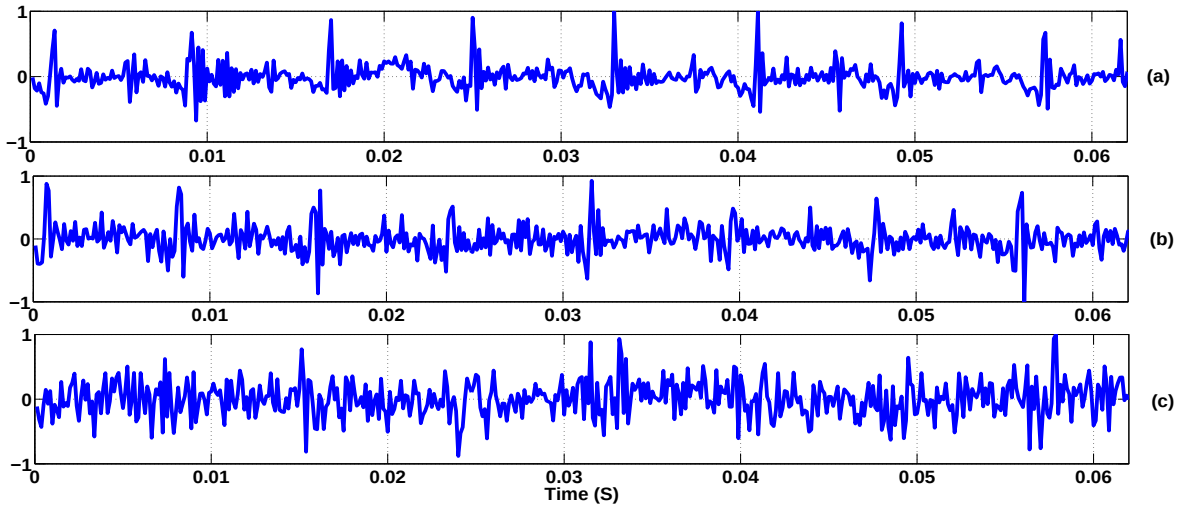

Figure 4.4: Illustration of LP Residual obtained from a segment of speech signal from the same acoustic unit when the microphone is placed at (a) 1 inch (b) 12 inches and (c) 30 inches from mouth of the speaker, respectively.

4.3.1 Short Time Analysis of Foreground and Distant Speech

The focus of this section is to study the short time analysis of foreground and distant speech using both temporal and spectral methods, respectively. It is of interest to know the source level information that can be accounted for the differences between foreground and distant speech scenarios. This is illustrated in Figure 4.4, where the segment of LP residual signal is taken from the same acoustic-phonetic unit from 3 different utterances recorded at varying distances of 1 inch, 12 inches, and 30 inches, respectively. The LP residual is obtained by inverse filtering process using LP coefficients calculated by 12^{th} order LP analysis using a 20 ms frame size with a frame shift of 10 ms. It can be noticed that there are considerable differences in the LP residual of speech signals recorded at varying distances from microphone, especially at instants of significant excitation. The instants are clearly visible as sharp peaks in case of close speaking scenario as shown in Figure 4.4 (a). However, it can be observed from Figure 4.4 (b), as the distance of microphone increases to 12 inches, the sharpness of the peaks at epoch location is reduced with respect to another region within a glottal cycle. This may be ascribed to the fact that, as the distance from the speaker increases from the microphone, the interfering sources have the profound effect on speech signal by reducing the overall SNR level. Nevertheless, it

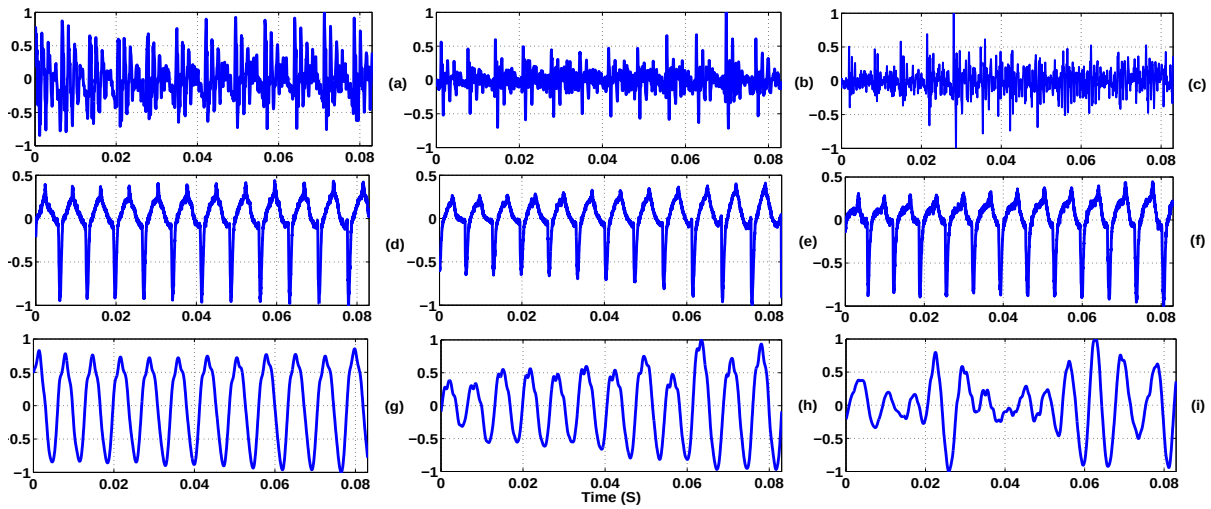

Figure 4.5: The nature of ZFFS, where, (a), (b) and (c) are speech segments taken from the same acoustic unit when speech signal is recorded at 1 inch, 12 inches and 30 inches, respectively, (d), (e) and (f) are the corresponding difference EGG signals, while (g), (h) and (i) are its corresponding ZFFS segments.

can be observed that epoch locations still remains to be high SNR regions compared to other regions in LP residual signal. This is further illustrated using Figure 4.4 (c), where the speech signal is recorded when the speaker is at 30 inches from the recording microphone. The effect of reduction in the peakiness of epoch locations is further exaggerated as the distance between speaker and microphone increases. This may lead to less reliable extraction of epochs.

Figures 4.5 (a), (b) and (c) shows a short time segment of speech signal chosen from the same acoustic-phonetic units, recorded at 1 inch, 12 inches, and 30 inches, respectively. The Figures 4.5 (d), (e) and (f) shows the corresponding difference EGG plots, where the epochs are shown in terms of sharp negative peaks. The ZBFS obtained by passing the speech signal through ZBF is shown in Figures 4.5 (g), (h) and (i). The positive zero crossings of ZBFS corresponds to epoch locations. It can be observed that the epochs can be reliably obtained from speech signal spoken from 1-inch distance to microphone and it can be observed that ZBFS is sinusoidal in nature. However, as the distance between speaker and microphone increases, there is distortion in ZBFS and this leads to spurious false alarms or increased miss rate as shown in Figure 4.5 (i). Hence, the epoch extraction performance is expected to decrease with increased distance between speaker and microphone. This is consistent with the studies

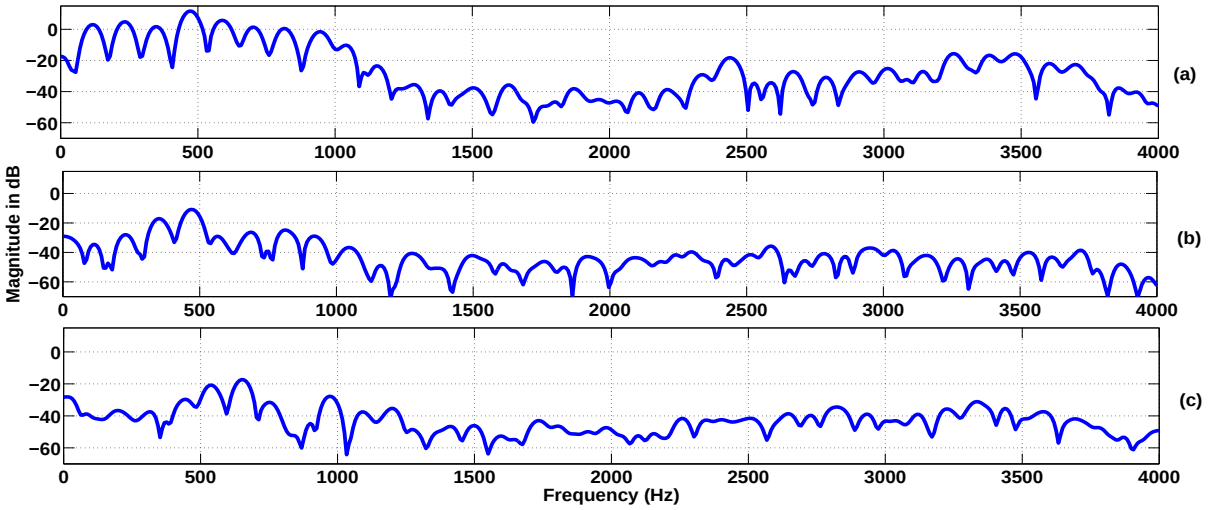

Figure 4.6: Illustration of magnitude spectrum obtained from a segment of speech signal from the same acoustic unit when the microphone is placed at (a) 1 inch (b) 12 inches and (c) 30 inches from mouth of the speaker, respectively.

conducted in Section 4.3 using ZBFS.

It is observed in Section 4.3 that there are differences in spectrogram plots between foreground and distant speech. This is further illustrated in Figure 4.6 using log magnitude spectrum obtained from a 50 ms speech segment taken from the same acoustic-phonetic unit from 3 different utterances recorded at distances of 1 inch, 12 inches and 30 inches from mouth to microphone sensor, respectively. It can be observed that the magnitude spectral dynamic range is relatively larger in the case of close speaking scenario compared to distant speech recordings. Also, it can be observed that the harmonic structure is better preserved in case of close speaking scenario compared to distant speech recordings. The spectral analysis reveals that there are differences in spectral characteristics based on the distance between microphone and speaker.

4.3.2 Foreground Speech Analysis in Noisy Environments

Even though foreground speech signal is recorded from a close distance of 1 - 3 inches from microphone, still there can be degradation of foreground speech due to interfering background noise. The overall SNR levels of foreground speech signal reduce due to the presence of significant amount of background noise. However, epochs being the instants of significant excitation and regions around such instants are the least affected by interfering noise. This is further

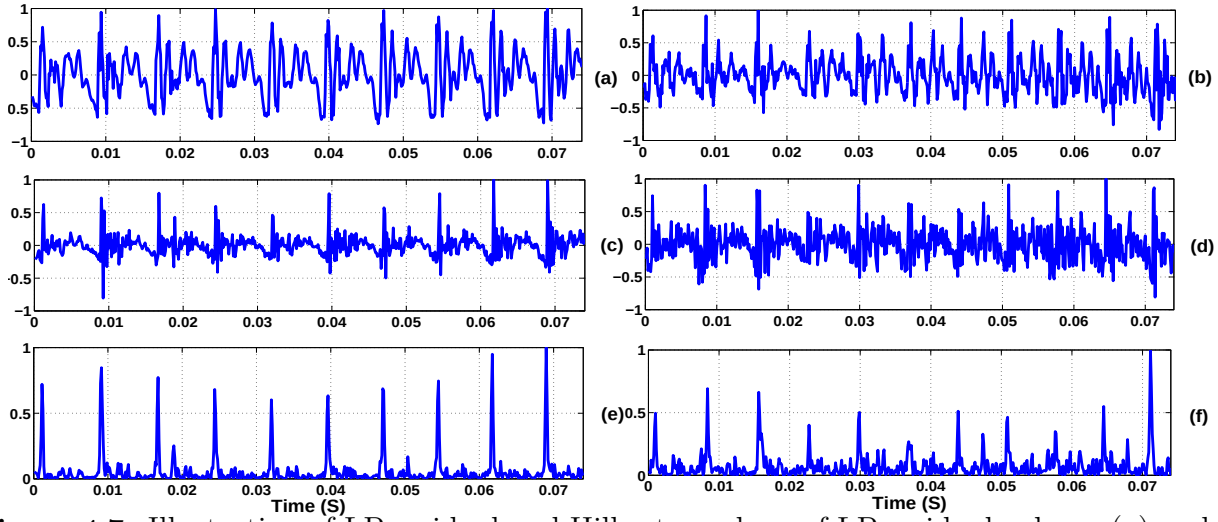

Figure 4.7: Illustration of LP residual and Hilbert envelope of LP residual, where, (a) and (b) are speech segments recorded when AC machine is switched OFF and ON, respectively. (c) and (d) are corresponding LP residual signals, (e) and (f) are corresponding HE of LP residual signals.

illustrated using Figure 4.7, where, Figures 4.7 (a) and (b) shows the segment of speech signal recorded at 1-inch a distance between the speaker and the microphone, where, AC machine is switched OFF and ON, respectively. It is of interest to analyze source excitation separately in order to study the characteristics of foreground speech. The linear prediction analysis is one of the important methods to deconvolve source excitation and filter. The LP residual is computed as explained in Section 4.3 which can be used for source excitation analysis. The LP residual signal computed from speech segments shown in Figures 4.7 (a) and (b) can be observed in Figures 4.7 (c) and (d), respectively. It can be noticed that epoch locations have sharp characteristics in both segments. However, there is a reduction in peak amplitude level at epoch locations when there is a significant amount of background noise. The LP residual signal is a random polarity signal and hence it is difficult to analyze such signal directly. Alternatively, it is easier to analyze the epoch locations using Hilbert Envelope (HE) of LP residual signal defined as the magnitude of the analytic signal [101]. The analytic signal is the complex temporal representation of the real signal and is given by

$$\tilde{e}(n) = e(n) - je_h(n) \tag{4.6}$$

where, $e_h(n)$ is the Hilbert transform of $e(n)$. The HE of LP residual $e(n)$ is computed as

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{4.7}$$

The HE is a unipolar signal and they are as shown in Figures 4.7 (e) and (f). It can be noticed that the epoch locations are exaggerated relative to other regions within a glottal cycle. From the observation of LP residual and HELP residual signal, it can be noticed that the epoch locations are intact even when there is a significant amount of background noise. Due to such characteristics of foreground speech, the epoch locations can be robustly extracted.

The Sections 4.3, 4.3.1, and 4.3.2 illustrated the nature of speech signal when it is recorded in foreground and distant scenarios. The unique characteristics of foreground speech should motivate us to derive certain features that help to distinguish foreground speech from rest of the background interfering sources and further enhance foreground speech. The interfering background source may also include background speaker.

4.4 Foreground Speech Segmentation

The acoustic background noise picked up by the recording microphone can be a speech like. It is, therefore, difficult to distinguish the desired foreground speech from rest of the background noise. However, it is important to temporally segment the foreground speech regions from rest of the background noise for subsequent processing to enhance the foreground regions.

4.4.1 Excitation Source based Features

The nature of the resulting ZBFS characteristics is different for background acoustic sources compared to foreground speech. It can be noted from [32] that the nature of ZBFS is nearly periodic and has higher amplitude levels for foreground speech, while the ZBFS appears to be dispersed in case of background speech and other acoustic sources. Also, the amplitude levels of ZBFS is significantly low in background noise regions compared to foreground speech regions. This can be attributed to the fact that the proximity of foreground source is closer to the microphone than other acoustic sources. This is illustrated in Fig. 4.8, where, Figures 4.8(a)

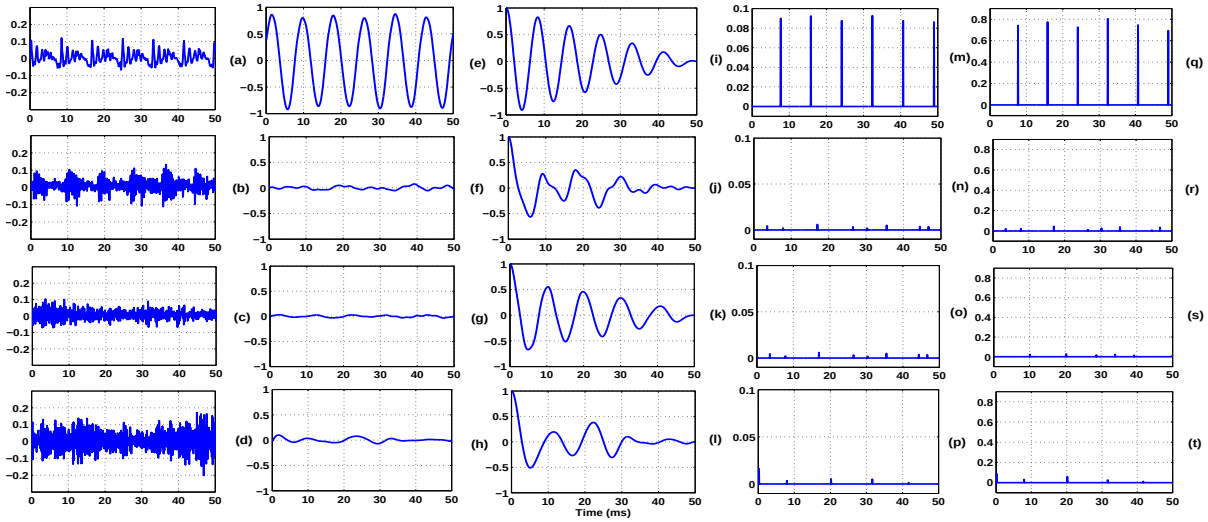

Figure 4.8: 50 milliseconds of (a) foreground speech (b) background speech (c) background music with vocals (d) background noise. Respective, ZBFS((e) - (h)), normalized autocorrelation sequence using ZBFS ((i) - (l)), strength of excitation using ZBFS ((m) - (p)), and ZBFS energy sampled at glottal closure instants ((q) - (t)).

and (b) show 50 *ms* segments of foreground speech and background speech chosen from the same recording. The signal is recorded at 16 kHz sampling rate using headphone microphone connected to a laptop in a living room. The speaker is closer to the microphone while talking and a television is playing speech in the background. Since the signal is recorded in the natural environment using single microphone sensor there is no reference clean speech signal available to measure signal to noise ratio (SNR). Hence, a similar measurement can be made using foreground to background ratio (FBR) that essentially computes the ratio of normalized foreground speech power to background power and it is expressed in decibels (dB). The detailed description of FBR is given in Section 5.2.3.

The segments shown in Figures 4.8(a) and (b) are taken from the signal having FBR as 6.27dB. The Figures 4.8(e) and (f) illustrate the corresponding ZBFS, while, Figures 4.8(a) and (b), respectively, while, Figures 4.8(i) and (j) shows the normalized autocorrelation sequence computed from ZBFS. The slope value measured at positive zero crossings of ZBFS indicates the strength of excitation (SoE) [30]. The Figures 4.8(m) and (n) shows the SoE computed from ZBFS segments shown in Figures 4.8(e) and (f), respectively. The Figures 4.8(q) and (r)

illustrates normalized ZBFS energy computed using the segments shown in Figures 4.8(e) and (f), respectively. The ZBFS normalized energy is computed using a frame size of 5 ms with one sample shift. However, for illustration purpose, the ZBFS energy is sampled at glottal closure instants to compare them with SoE. It can be noticed that the signal contours of SoE and ZBFS energy follows a similar trend and hence offers the same discriminative abilities to distinguish between foreground and background speech regions. The similar trend of these two signals can be attributed to the fact that ZBFS energy depends on SoE and hence they are related.

The Figure 4.8(c) shows 50 *ms* segment of background music with vocals chosen from the signal having FBR as 5.21 dB. The signal is recorded from a foreground speaker when the television is playing music with vocals in the background. The corresponding ZBFS, normalized autocorrelation sequence computed from ZBFS, SoE derived using ZBFS and ZBFS normalized energy plots are shown in Figures 4.8(g), (k), (o) and (s), respectively. Further, Figure 4.8(d) shows 50 *ms* segment of foreground speech and the background noise chosen from another recording. The speech signal is recorded in foreground scenario from a speaker in an office environment when the mosaic polishing machine is operating at the background. The segment is chosen from the signal having FBR as 7.38 dB. The corresponding ZBFS, normalized autocorrelation sequence computed from ZBFS, SoE derived using ZBFS and ZBFS normalized energy plots are shown in Figures 4.8(h), (l), (p) and (t), respectively.

It can be observed that the ZBFS derived from foreground speech region appears to be nearly periodic while it appears distorted for background speech, music and noise regions. This fact is further illustrated using normalized autocorrelation sequence plots. It can be noticed that the value of the first largest peak (excluding the center) normalized with respect to the center value in the autocorrelation sequence defined as normalized autocorrelation coefficient (NACC) is larger in the case of foreground speech regions compared to background regions. The reason for such a change in ZBFS characteristics can be ascribed to the fact that foreground source is closer to microphone sensor and hence the excitation source is least affected by interfering sources. The amplitude levels of ZBFS is directly related to epoch strength of the excitation source and hence energy at foreground speech regions is relatively larger compared to background

regions. Hence, such discriminative features derived from ZBFS can be used to segment the foreground speech from rest of the background noise that may include speech like sources. This is further illustrated by choosing a full sentence spoken by foreground speaker in Figure 4.9, where, Figure 4.9(a) shows the speech signal recorded from a male speaker in office while the mosaic polishing machine is operating in the background which has FBR value 7.38 dB. The signal is recorded using a headphone microphone connected to a laptop for recording. It can be observed that there is a significant amount of noise present in the background. The NACC and ZBFS energy computed from the signal is shown in Figures 4.9(b) and (c), respectively. The two features derived from ZBFS offers discriminative information between foreground and background regions. The NACC and ZBFS energy are relatively higher in the case of foreground speech regions relative to background noise regions. Hence, theses two features can be used to segment the foreground speech from the recorded signal.

4.4.2 Vocal Tract Information based Feature

The features discussed so for are derived using excitation source information and does not make use of vocal tract articulatory gestures that are unique to foreground speech. One way of capturing the vocal tract information is by measuring the spectral envelope. However, it is suggested that the temporal dynamics of the spectral envelopes as more reliable means for carrying the linguistic context of the speech message which can be obtained through modulation spectrum energy of speech signal. The temporal envelope of speech is dominated by low-frequency components that are in the similar range to the dynamics of speech production, in which the articulators move at such rates [118]. It is studied that the linguistic information of speech signal lies in the range of 2 to 16 Hz and centered at 4 Hz. Subsequently, filtering slow and fast varying trajectories of spectral envelopes can be useful in alleviating the effects of interfering background sources. Hence, modulation spectrum energy can be used as a feature to segment the foreground speech from rest of the background noise. In this work, the modulation spectrum energy is extracted from the signal as given in [119]. However, the signal is divided into 18 subbands using compressive gammachirp auditory filter (cGC) [120,121]. The auditory

filter designed is level dependent and nonlinear that emulates psychophysical data on masking and two-tone suppression. The cGC consists of two filters *viz.,* a passive gammachirp filter (pGC) and a dynamic filter which is an asymmetric function that shifts in frequency with stimulus level. The cGC filter is realized using the following relationship

$$g_c(t) = at^{n_1-1}\exp(-2\pi b_1 \mathrm{ERB_N}(f_{r1})t) \times \exp(j2\pi f_{r1}t + jc1\ln t + j\phi_1) \tag{4.8}$$

where a is amplitude; n_1 and b_1 are parameters defining the envelope of the gamma distribution; c_1 is the chirp factor;f_{r1} is a frequency referred to as the asymptotic frequency, since the instantaneous frequency of the carrier converses to it when t is infinity; $\mathrm{ERB_N}(f_{r1})$ is the equivalent rectangular bandwidth of average normal hearing subjects; ϕ_1 is the initial phase; and $\ln t$ is the natural logarithm of time. The Fourier magnitude spectrum of cGC is given by

$$|G_c(f)| = a_\Gamma . |G_T(f)| . \exp(c_1\theta_1(f)) \tag{4.9}$$

$$|\theta_1(f)| = \arctan\left(\frac{f - f_{r1}}{b_1 \mathrm{ERB_N} f_{r1}}\right) \tag{4.10}$$

where $|G_T(f)|$ is the Fourier magnitude spectrum of the gammatone filter; $\exp(c_1\theta_1(f))$ is an asymmetric function because θ_1 is an antisymmetric function centered at the asymptotic frequency f_{r1} and a_Γ is constant. Further the asymmetric function $\exp(c_1\theta_1(f))$ is decomposed into lowpass and highpass asymmetric filter functions to represent passive and dynamic components separately. The resulting compressive cGC filter $|G_{cc}(f)|$ is

$$|G_{cc}(f)| = [a_\Gamma . |G_T(f)| . \exp(c_1\theta_1(f))] . \exp(c_2\theta_2(f)) \tag{4.11}$$

$$|G_{cc}(f)| = |G_{cp}(f)| . \exp(c_2\theta_2(f)) \tag{4.12}$$

The compressive gammachirp is composed of a level independent passive gammachirp filter (pGC) $G_{cp}(f)$ which represents the passive basilar membrane and a level dependent highpass asymmetric function that simulates active component in the cochlea. The Amplitude envelope

is computed from each individual subband outputs obtained from filter bank. Each filter's output is the first halfwave rectified and then subjected through a lowpass filter having a cutoff frequency $28Hz$. The amplitude envelope obtained is downsampled by a factor of 100 and normalized by an average value obtained from the respective filter bank output. The modulations of the normalized envelope signals obtained are further analyzed using Discrete Fourier Transform (DFT). The DFT is computed using $250ms$ Hamming window with shift of $12.5ms$ which essentially captures dynamic properties of the signal. However, the $2-16Hz$ components from each such channels are summed together to obtain the modulation spectrum energy signal. The modulation spectrum energy can be computed using the following relationship

$$m(i) = \sum_{p=1}^{18} \sum_{k=k1}^{k=k2} \mid \hat{S}_p(k,i) \mid^2 \tag{4.13}$$

where, i is the frame index, p represents the critical band filter and $k1$, $k2$ represents frequency index of $4Hz$ and $16Hz$, respectively. $\hat{S}_p(k,i)$ is obtained using the following relationship

$$\hat{S}_p(k,i) = \sum_{n=0}^{N-1} \hat{s}_p(n + i \times F) w(n) e^{-j2\pi nk/N} \tag{4.14}$$

where, $\hat{s}_p(n)$ represents the normalized envelope of pth filter output, F is the frame shift, $w(n)$ is a Hamming window, and N is the number of points used for computing the DFT.

The modulation spectrum energy computed for each frame are upsampled to 16000 $samples/s$. The modulation spectrum energy computed from the signal shown in Figure 4.9(a) is illustrated in Figure 4.9(d). It can be noticed that the modulation spectrum values are significantly higher in foreground speech regions compared to background noise regions. This can be attributed to the fact that vocal tract articulatory gestural movements operate in an exclusive frequency range of $2-16\ Hz$ compared to other acoustic sources. Hence, the modulation spectrum energy can be used to distinguish the foreground speech regions from rest of the background content.

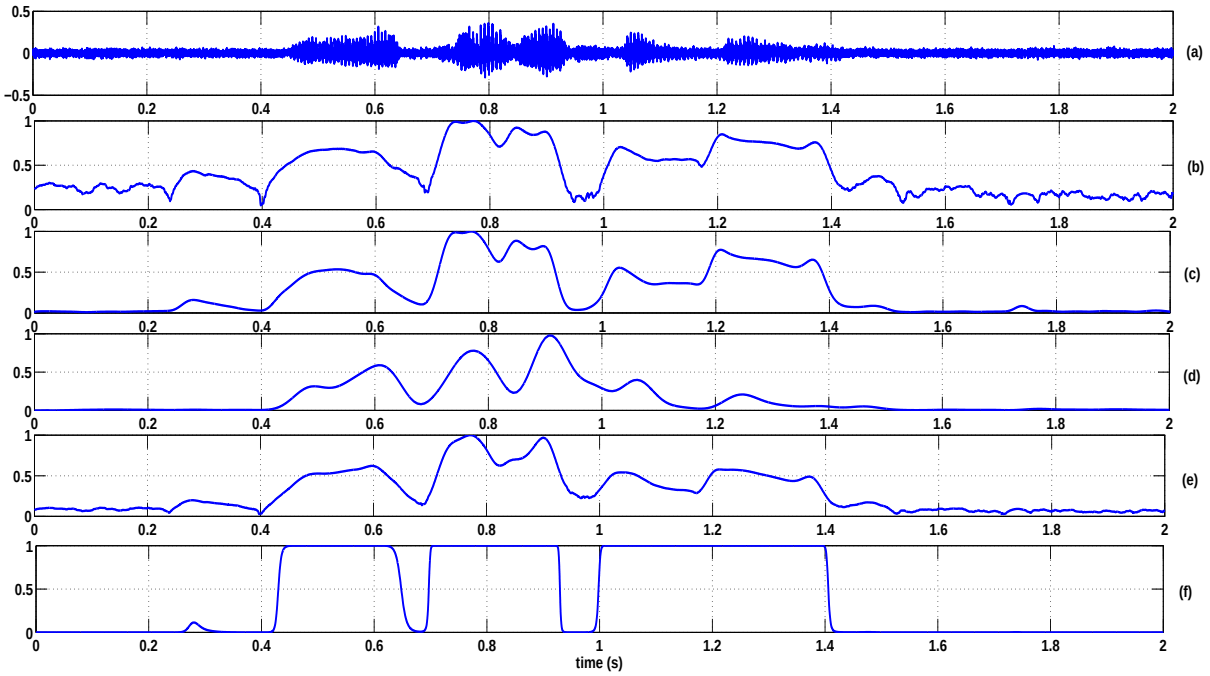

Figure 4.9: The figure illustrates the foreground segmentation using the combined evidence from different features (a) speech signal recorded in foreground scenario, (b) normalized first order autocorrelation coefficients derived from ZBFS, (c) short term energy derived from ZBFS, (d) modulation spectrum energy derived from speech signal, (e) combined evidence , and (f) combined evidence passed through sigmoidal function to segment the foreground speech regions from rest of the background noise.

4.4.3 Combined Evidence

The excitation source based features such as NACC and ZBFS energy are combined with vocal tract based modulation spectrum in order to identify the foreground speech regions. In order to illustrate the concept NACC, ZBFS energy, and modulation spectrum energy are computed with 1 sample shift and each such signal is amplitude normalized. Let us denote such normalized sequence of NACC, ZBFS energy, and modulation spectrum energy as $N_c(n)$, $Z_E(n)$, and $M_E(n)$, respectively, where, n represents the number of samples in a given signal. The features are temporally added and normalized with respect to maximum value of the entire sequence to obtain the combined evidence given by $E(n) = (N_c(n)+Z_E(n)+M_E(n))/\max(N_c(n)+Z_E(n)+M_E(n))$. The combined evidence of all three features is shown in Figure 4.9(e). However, it is difficult to set the threshold directly on such signal for foreground segmentation. Alternatively,

the gross level feature is obtained by passing $E(n)$ through the sigmoidal function given by

$$w_g(n) = (1 - w_{gm}) \frac{1}{1 + \exp\left(-\lambda(E(n) - T_h)\right)} + w_{gm} \tag{4.15}$$

where, $w_g(n)$ is the sigmoidal function of $E(n)$, λ is slope parameter set to 20, T_h is the threshold derived from mean value of the signal $E(n)$ and w_{gm} is minimum value of the sigmoidal function which is set to 0 in this case. The $w_g(n)$ function forms the gross level feature which mainly helps to segment the foreground speech in the presence of background noise. This is illustrated using Figure 4.9(f) where the foreground regions are further enhanced relative to other background regions. The weight function $w_g(n)$ derived from three features can be used as gross level feature to temporally enhance the foreground speech.

4.4.4 Performance Evaluation of Foreground Speech Segmentation

The foreground speech segmentation can be compared to voice activity detection (VAD) as they have a similar objective. Hence, the performance is compared with considered two state of the art VADs. One of the VAD considered for performance evaluation is the latest G.729 ITU-T VAD standard developed for fixed telephony and multimedia communications [21]. The G.729 VAD uses multiple boundaries for voice activity decision. More recently a variable frame rate approach was proposed on the basis that speech signal is not stationary in a short period of fixed frame rate [22]. Therefore a variable frame rate (VFR) approach is followed based on *a posteriori* SNR-weighted energy distance. In order to compare the performance of all three different methods 10 different TIMIT speech files consisting of 5 female and 5 male speakers are considered for evaluation [122]. The speech files are modified by appending $50ms$ silence on either side of the speech signal to closely simulate the natural recording scenario. Also, it can be noted that the ground truth of VAD is manually marked in TIMIT files and they form the reference for evaluation. The noisy speech recordings are simulated by additively combining clean speech signal with 4 different types of noise at different levels. The 4 different types of noise considered are mosaic machine noise, hostel mess recording (babble noise), background music with vocals and background speech. In order to compare the performance of all three

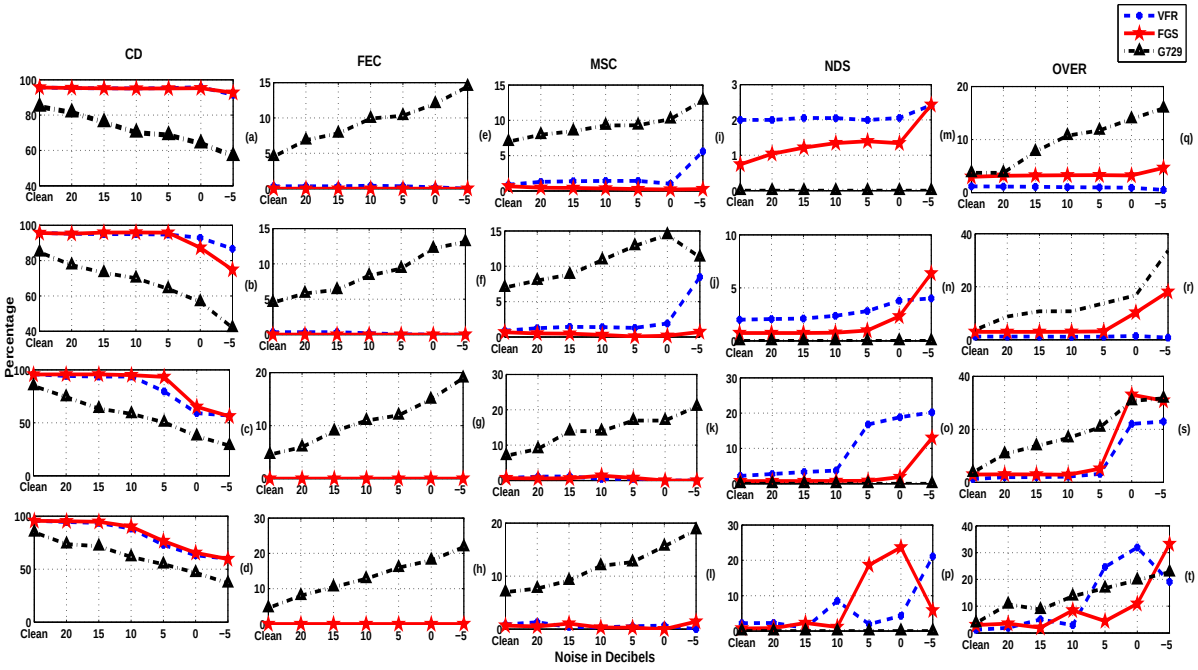

Figure 4.10: Illustration of performance evaluation of VAD using foreground segmentation, VFR VAD and G.729 VAD for 4 different noise types, ((a) - (d)) is CD, ((e) - (h)) is FEC, ((i)-(l)) is MSC, ((m) - (p)) is NDS, ((q) - (t)) is OVER plots with different additive noises of machine noise, babble noise, background music with vocals and background speech, respectively.

methods, the following parameters are used

- Correct VAD decision (CD): Correct decisions made by the VAD.
- Front End Clipping (FEC): Clipping introduced while passing from noise to speech activity.
- Missed Speech Clipping (MSC): Clipping due to speech misclassified as noise.
- Noise Detected as Speech (NDS): Noise interpreted as speech within a silence period.
- OVER: Noise interpreted as speech due to the VAD flag remaining active in passing from speech activity to noise.

All the parameters are expressed in terms of percentage. The Figure 4.10 illustrates the performance of three different methods in the form of graph plots. The Figures 4.10(a), (e),

(i), (m) and (q) show the performance of different methods in terms of CD, FEC, MSC, NDS and OVER, respectively for additive background mosaic machine noise. It can be noted that the noise is added at 20, 15, 10, 5, 0 and -5 dB. It can be observed that the performance of foreground segmentation and VFR VAD are equally robust for background machine noise even at low SNR levels. However, the performance of G.729 VAD degrades as the noise levels increase. The reason for degradation in case of G.729 is due to increase in FEC, MSC, and OVER. Similarly, Figures 4.10(b), (f), (j), (n) and (r) show the performance of three methods for additive background noise recorded in a large hostel mess environment when it is crowded (babble noise). It can be observed from the plots that the performance of foreground segmentation and VFR VAD remains robust to the additive noise level of $5dB$ and followed by degradation in performances. However, it can be noticed that the performance of VFR VAD is better compared to foreground segmentation and G.729 VAD at lower SNR levels.

The deterioration in the performance of foreground segmentation is mainly due to the increased OVER rate at low SNR levels, particularly at 0 and $-5dB$ levels. The Figures 4.10(c), (g), (k), (o) and (s) show the performance curves for additive background music with vocals at different levels. It can be observed from the plots that the foreground segmentation performance is better than other two methods at low SNR levels. Similarly, Figures 4.10(d), (h), (l), (p) and (t) show the performance of 3 different methods for additive background speech at different levels. It can be noticed that there is degradation in performance of all 3 methods at low SNR levels. However, the performance of foreground segmentation is marginally better compared to VFR VAD at low SNR levels. The signal characteristics of background music with vocals and background speech are similar to foreground speech signal. Hence, analyzing the signal at zero frequency rather than an entire set of frequencies helps in better discriminating the foreground speech from rest of the background noise as in the case of foreground speech segmentation. As explained in Section 4.4, the features derived from ZBFS such as ZBFS energy and NACC are used as features for foreground segmentation. Also, modulation spectrum energy centered at $4Hz$ is also used in combination with features derived from ZBFS. The performance evaluation shows that foreground segmentation is robust to interfering background noise.

4.5 Summary

In this chapter, a detailed analysis of close and distant speech recordings using a single microphone is carried using temporal and spectral based methods. Based on analysis it is found that the speech recordings have different signal characteristics depending on the distance between microphone and speaker. A definition of foreground speech is established based on the close speaking scenario. The foreground speech analysis motivated to derive certain production based features to segment foreground speech from rest of the background regions. This chapter proposed a foreground speech segmentation method using ZBFS. The ZBFS features, namely, the normalized first order autocorrelation coefficient and strength of excitation showed significant discrimination among foreground speech and background degradation regions. The modulation spectrum energy is also used as a production based feature that is useful to distinguish foreground and background regions. A method is developed using three different features for foreground speech segmentation. The effective foreground speech segmentation helps to temporally separate foreground speech regions and background noise regions, however, the foreground speech region has to be further enhanced to make it suitable for listening and other speech-based applications.

5

Foreground Speech Enhancement

Contents

Objective

Based on the foreground speech analysis it is evident that the speech signal characteristics are different when the speaker is closer to the microphone. Hence, certain production based features are used for foreground speech segmentation. However, it is necessary to enhance the foreground speech regions to improve the speech quality. The proposed work exploits the speech production features like Glottal Closure Instants (GCIs) in time domain and vocal tract information in the spectral domain to enhance the desired speaker's speech. Further, the foreground speech is perceptually enhanced using the auditory perception feature in the Mel frequency domain using Mel cepstral coefficients (MCC) and its inversion using Mel log spectrum approximation (MLSA) filter. The focus is on enhancing the production and perceptual features of foreground speech rather than relying on modeling the interfering sources. The speech data is collected in natural environments from different speakers in order to evaluate the proposed method. The enhanced speech signals derived at three different stages of the proposed method are evaluated with state of the art methods in terms of subjective and objective measures. The proposed method provides improved performance compared to the considered state of the art methods.

5.1 Motivation for Foreground Speech Enhancement

In most of the cases, the speech signal recorded in the natural environment is degraded by other interfering acoustic sources. The natural environment typically refers to an office or laboratory environment with relatively high acoustic background noise like call centers. The degradation can be of different levels and consists of one or more interfering sources. Speech enhancement is still a challenging task when the signal is recorded using a single sensor in natural environment [62]. The task is complicated if the interfering sources are from other background speakers. The perceptual quality improvement in terms of intelligibility and reduction in the background noise to enable comfortable speech communication is of paramount importance [123]. In such cases, it is of interest to segment the desired speaker's speech from

rest of the interfering background noise first and then enhance the desired speaker's speech.

The general approach of most methods is to model the additive noise only components from the noisy speech signal and suppress them to obtain the enhanced speech [2, 4]. The noise modeling depends on efficient segmentation of additive background noise from rest of the speech regions. The suppression of noise usually takes place in the spectral domain through subtraction. However, such spectral subtraction methods introduce distortions in enhanced speech because of overestimating the noise spectrum in the form of undesired musical tones. Many methods were proposed in order to overcome this problem [124,125]. The human auditory perceptual cues were considered to improvise the spectral subtraction methods [37, 126]. The objective is not to completely suppress the additive noise. Rather, the purpose is to utilize the masking properties of the human auditory system and live with a certain amount of residual noise that is below the masking threshold. The advantage of such algorithms over conventional methods is that they do not introduce spectral distortions in the form of audible musical noise due to over subtraction. However, utilizing complex auditory phenomena like psychoacoustic models for the sake of speech enhancement can make such methods to be more complex.

There is some evidence that humans perceive speech by capturing the features from a high signal to noise ratio (SNR) regions and extrapolate those features to low SNR regions, both in temporal and spectral domains [9]. The idea is to enhance the high SNR regions further relative to low SNR regions for enhancing the noisy speech signal. This is equivalent to the phenomenon of Lombard effect, where, speaker emphasizes the production apparatus to increase SNR of the produced speech when background noise in feedback path increases [127]. There are some methods proposed in the literature that basically utilize the high SNR regions like instants of significant excitation of the speech signal in the temporal domain to enhance speech [8]. The idea is to emphasize instants of significant excitation which are predominantly glottal closure instants (GCIs) relative to other regions of LP residual signal. The GCIs are located using Hilbert envelope of LP residual (HELP), using which, a temporal weighting function is derived so to enhance GCI positions relative to other regions. The synthesized speech signal from such temporal enhancement may not completely eliminate the background noise, nevertheless, the

distortion caused by such enhancement methods are minimal. The recently proposed method in [43] utilizes temporal enhancement as the preliminary stage of enhancement and subsequently uses spectral enhancement to eliminate the remaining residual noise. The advantage of this method is that it causes lesser spectral distortion.

There is evidence that when the speech signal is recorded in the natural environment, the speech production characteristics tend to vary depending on the levels of interfering sources [128]. Therefore, speech enhancement techniques may have to focus on exploiting such unique characteristics of the speech signal. The present work focuses on the practical scenario when the speech signal is recorded in different natural environments. In a typical recording scenario (head mounted microphone or mobile phones held close to ears), the desired speaker is closer to microphone relative to other interfering sources. In this work, it is assumed that the proximity of the desired speaker is closest to microphone compared to the distance between other sound sources and the microphone. Due to the close proximity of speaker to microphone as explained in Section 4.3 and also the modified speech production characteristics due to acoustic feedback, there are significant differences in the nature of signal for foreground speech and the background noise [32, 129]. The core idea being that not all foreground speech regions are affected equally by interfering background noise. In particular, the instants of significant excitation in temporal domain and formant locations in the spectral domain remain robust to background noise. Hence, such regions from high SNR regions of the speech signal can be utilized for speech enhancement. Furthermore, the proposed method utilizes the subtle aspects of human auditory feature to enhance the foreground speech. The merits of the proposed method lie in exploiting the speech production features such as excitation source and vocal tract information along with auditory perception feature and study the benefits of each in terms of speech enhancement.

The rest of the chapter is organized as follows: Section 5.2 illustrates the overall block diagram of the proposed method and explains different modules in brief. Section 5.2.1 describes the details of excitation source based enhancement, while, Section 5.2.2 explains the details of formant based enhancement using a block diagram. The Section 5.2.3 explains the details of analysis and synthesis submodules using MCCs and MLSA filter, respectively. Section 5.3

describes the details of the evaluation procedure and results obtained in terms of subjective and objective evaluation scores. The summary of the present work is mentioned in Section 5.4.

5.2 Foreground Speech Enhancement

The proposed work mainly consists of foreground speech segmentation and multistage foreground speech enhancement modules. If $s(n)$ is the speech signal recorded in foreground scenario of a natural environment, then there can be other interfering sources along with foreground speech signal. The objective of the current work is to first temporally segment the foreground speech regions from rest of the background noise and then enhance the foreground speech regions. The overall block diagram of the proposed work is shown in Fig. 5.1. The details of the foreground speech segmentation module are explained in Section 4.4.

The foreground speech enhancement is carried using multiple stages. The proposed enhancement scheme is explained using the block diagram shown in Figure 5.1, it is classified into four major modules *viz*, foreground speech segmentation, excitation source based enhancement, formant based enhancement and perceptual based enhancement modules. The foreground speech segmentation is discussed in Section 4.4. It can be observed that all the modules are connected sequentially. The enhanced speech signal obtained from different modules are named as excitation source based enhancement ($s_t(n)$), formant based enhancement ($s_f(n)$), and perceptual based enhancement ($s_p(n)$), respectively.

5.2.1 Excitation Source based Foreground Speech Enhancement

The excitation source information, especially the instants of significant excitation are high SNR regions which can be used as anchor points to enhance the foreground speech regions. A robust method is required to locate the instants of significant excitation even when the speech signal is recorded in noisy environments. It is shown in [117] that ZBF is robust in identifying the locations of GCIs in foreground scenarios and there is no requirement of estimating $F0$. The positive zero crossings of ZBFS derived from Eqn. (4.5) precisely match with the locations of GCIs. In order to modify the excitation source signal using the GCI locations, it is beneficial

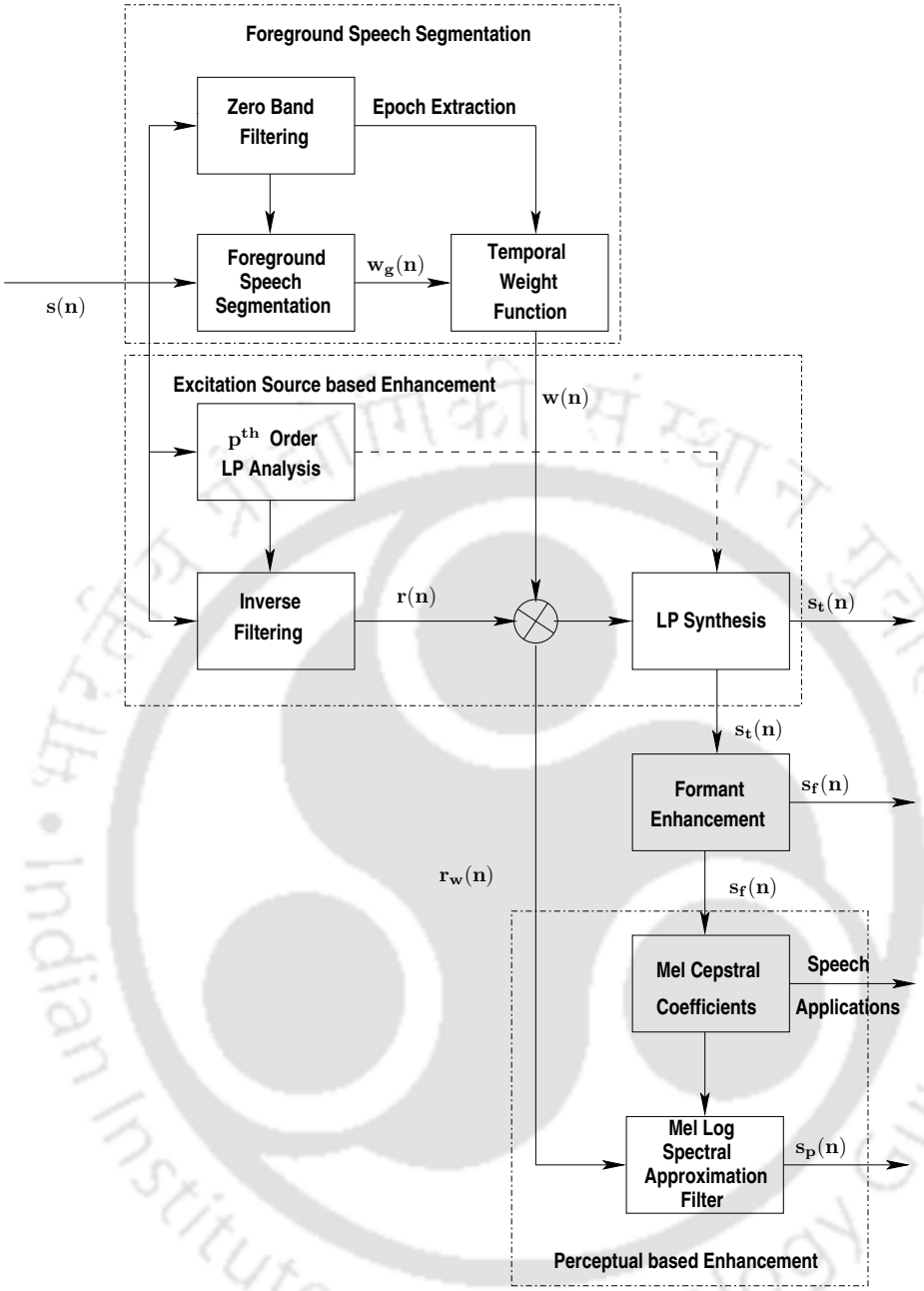

Figure 5.1: The overall block diagram of the proposed foreground speech segmentation and enhancement method, where, $s(n)$ is the input speech signal recorded in foreground scenario, $w_g(n)$ is the gross weight function that mainly segments the foreground speech regions from rest of the background noise, $w(n)$ is the final temporal weight function, $r(n)$ is the LP residual signal, $r_w(n)$ is the temporally weighted LP residual signal, $s_t(n)$ is EBE output, $s_f(n)$ is formant based enhanced output and $s_p(n)$ is perceptually enhanced output

to resolve the speech signal in terms of source and filter components.

In linear prediction (LP) analysis, the vocal tract system can be modeled as a time varying

all-pole filter using frame-based analysis. The LP analysis works on the principle that the current speech sample can be predicted from the past p samples, where, p is called the linear prediction order and is selected as $F_s/1000 + 4$ (F_s is the sampling frequency). In order to compute the LP coefficients, the frame size is chosen to be 20 ms with a frame shift of 10 ms. If $s(n)$ denotes the speech signal recorded in foreground scenario, then the predicted sample at the time instant n is given by

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{5.1}$$

where, a_k is the set of LP coefficients predicted. The residual error is the difference between the actual sample sequence $s(n)$ and the predicted sample sequence $\hat{s}(n)$ and is given by the relationship

$$l(n) = s(n) - \hat{s}(n) \tag{5.2}$$

From Eqns. (5.1) and (5.2), the residual signal $l(n)$ can be written in z domain as

$$L(z) = S(z) + \sum_{k=1}^{p} a_k S(z) z^{-k} \tag{5.3}$$

i.e.,

$$A(z) = \frac{L(z)}{S(z)} = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{5.4}$$

where, the LP residual signal can be obtained by filtering the speech signal $S(z)$ through the filter $A(z)$, which is generally called as inverse filtering. The prediction error is relatively high at GCI locations compared to other regions of the speech signal. Hence, the amplitude levels of LP residual signal is higher at GCI locations compared to other regions of LP residual signal. The LP residual signal is uncorrelated and therefore any modification of such residual signal introduces the least distortion for the later synthesis of the speech signal. It is shown in [8] that enhanced speech can be synthesized by modifying the LP residual signal without much audible distortion. The LP residual signal is derived from the noisy speech signal and modified by retaining 2 ms regions around GCIs. The method proposed in [43] uses the similar approach using GCIs as anchor points using HELP. The identification rate (IDR) and accuracy (IDA) of locating GCIs using HE of LP residual are inferior in the case of noisy speech signal compared

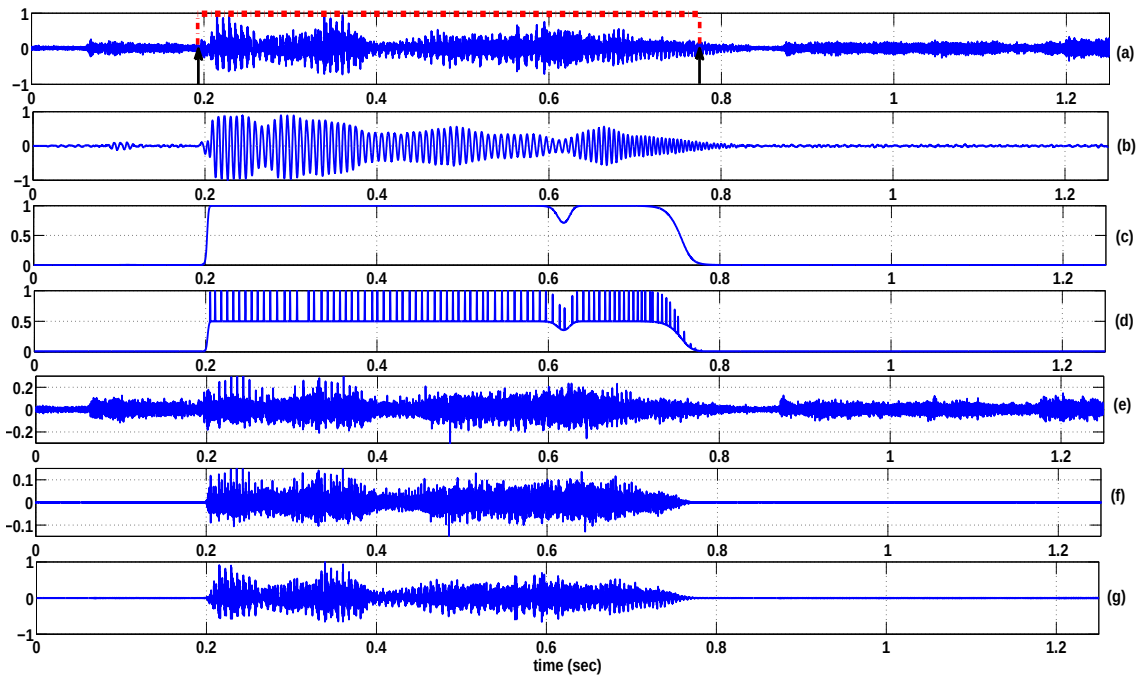

Figure 5.2: Illustration of excitation based enhancement. (a) speech signal recorded in foreground scenario (dotted lines and arrows indicate the foreground region), (b) the positive zero crossings of ZBFS indicate the epoch locations, (c) the foreground weight function $w_g(n)$, (d) the temporal weight function $w(n)$ by combining the evidence of epoch locations with foreground segmentation, (e) LP residual signal derived from speech signal, (f) modified LP residual signal $r_w(n)$ (g) excitation based enhanced speech signal synthesized.

to recently proposed methods [95]. Alternatively, in the proposed method, ZBF is used to locate GCIs directly from speech signal recorded in natural environments.

This is illustrated in Figure 5.2, where, Figure 5.2(a) shows the foreground speech signal recorded when music is playing in the background. It can be noticed that there is a significant amount of background music present along with the foreground speech signal. The foreground regions are shown in Figure 5.2(a) as dotted lines, and the corresponding ZBFS can be obtained by passing the speech signal through Eqns. (4.4) and (4.5) is shown in Figure 5.2(b). The positive zero crossings of ZBFS correspond to GCI locations of the speech signal. The region around GCI locations can be used as anchor points to modify the LP residual signal. The LP residual signal derived from speech signal as shown in Figure 5.2(e). In order to emphasize the regions around GCI locations, a fine weight function is obtained similarly to [43]. The GCI location is convolved with Hamming window function $h_w(n)$ having a temporal duration of

3 ms that closely corresponds to closed phase interval of a glottal cycle. If GCIs are considered as shifted train of impulses, then the fine weight function $w_f(n)$ is given by

$$w_f(n) = \left(\sum_{k=1}^{N_k} \delta(n - i_k) \right) * h_w(n) \tag{5.5}$$

where, N_k is the total number of GCIs located, i_k is the estimated location of GCIs. The minimum value of $w_f(n)$ is set to a threshold value of T in order to keep the distortion low because of overemphasizing the GCI locations in LP residual and the relationship is expressed as

$$w_f(n) = \begin{cases} T, & \text{if } w_f(n) < T \\ w_f(n), & \text{otherwise} \end{cases} \tag{5.6}$$

where, T is set to 0.5 in this work. It can be noted that temporal processing is not sensitive to a range of T values [43]. The final weight function $w(n)$ is obtained by multiplying foreground segmentation $w_g(n)$ as shown in Figure 5.2(c) with fine weight function $w_f(n)$ and is expressed as

$$w(n) = w_g(n) \times w_f(n) \tag{5.7}$$

The normalized final weight function $w(n)$ is multiplied to residual signal $l(n)$ to obtain the weighted LP residual signal (WLPR) $r_w(n)$ as shown in Figure 5.2 (f). The temporally enhanced speech signal $s_t(n)$ can be synthesized by the transfer function given in z domain as

$$S_t(z) = \frac{R_w(z)}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{5.8}$$

where, $S_t(z)$ is temporally enhanced speech signal and $R_w(z)$ is WLPR in z domain, while, a_k are the LP filter coefficients.

The temporally enhanced speech signal is shown in Figure 5.2(g). Temporally, there is an overall reduction in the background noise after temporal enhancement. This is further illustrated using Figure 5.3, where, Figure 5.3(a) shows the foreground speech signal recorded while music is being played at the background, and Figure 5.3(e) shows the corresponding narrowband spectrogram. There is a significant amount of background noise present throughout the temporal duration of the speech recording. The narrowband spectrogram shown in Figure

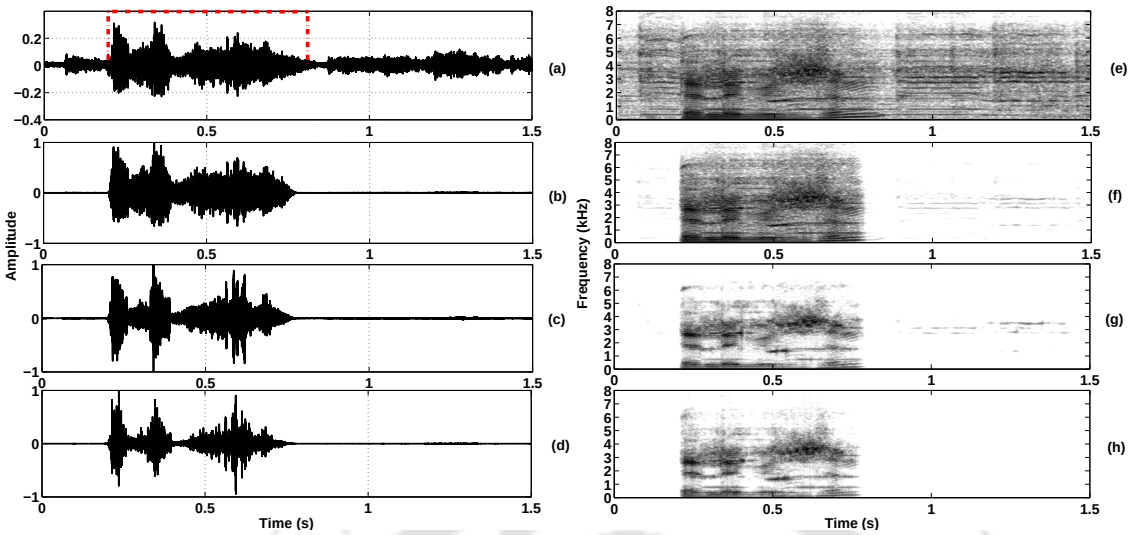

Figure 5.3: Illustration of different enhancement outputs and their narrowband spectrogram plots. (a) speech signal recorded in foreground scenario, (b) excitation based enhanced output, (c) formant based enhanced output, (d) perceptual based enhanced output, (e), (f), (g) and (h) are corresponding narrowband spectrograms.

5.3(f) obtained from temporally enhanced foreground speech signal illustrates the significant reduction of background noise. Although there is a reduction of background noise in the foreground regions, still there is audible background noise present in the foreground regions and it is evident from the spectrogram plots. Hence, excitation source based enhancement alone may not be sufficient to suppress the background noise present in the foreground speech regions.

5.2.2 Formant based Foreground Speech Enhancement

The Sections 4.4 and 5.2.1 helped to temporally segment the foreground speech regions from rest of the background noise and further enhance the foreground speech using excitation source information. However, the foreground speech enhancement $s_t(n)$ using excitation source information is still left with some residual background noise. Considering the fact that, instants of significant excitation are high SNR regions in the temporal domain, similarly, the formant peak locations are high SNR regions in the spectral domain. The vocal tract information is intact in most of the foreground scenarios, and hence, such information can be exploited to enhance the foreground regions further. The formant peak enhancement relative to spectral valleys has

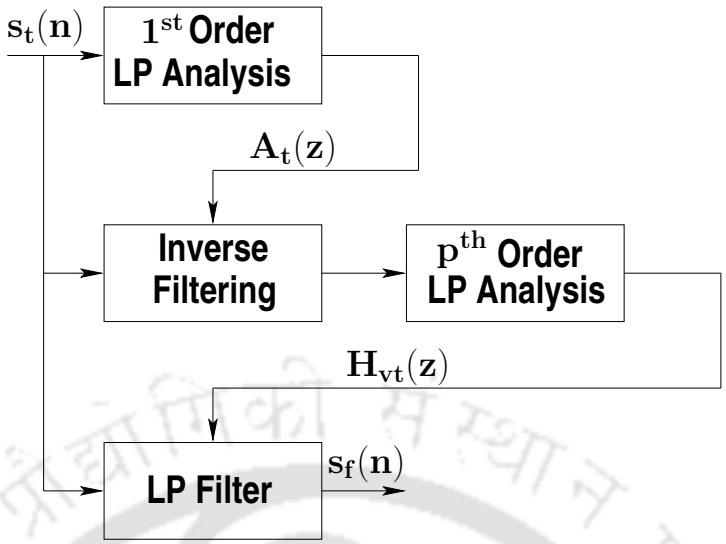

Figure 5.4: The formant enhancement block diagram, where, $s_t(n)$ is excitation based enhanced foreground speech signal, $A_t(z)$ is the 1^{st} order LP filter, $H_{vt}(z)$ is the p^{th} order LP filter and $s_f(n)$ is formant enhanced foreground speech signal.

been exploited in many methods for speech enhancement [130]. In the proposed method the 1^{st} stage of foreground speech enhanced output $s_t(n)$ is further subjected through formant based enhancement (FBE). The formant enhancement is carried on the LP spectrum which helps to enhance the formant locations relative to adjacent valleys. The formant enhancement is shown in the form of the block diagram in Figure 5.4.

$A_t(z)$ is the 1^{st} order LP inverse filter expressed as

$$A_t(z) = 1 + bz^{-1} \tag{5.9}$$

and $H_{vt}(z)$ is the LP filter predicted from p^{th} order LP analysis where p is chosen as $F_s/1000 + 4$ [52]

$$H_{vt}(z) = \frac{1}{1 + \sum_{k=1}^{p} c_k z^{-k}} \tag{5.10}$$

In order to estimate the spectral tilt from $s_t(n)$ a 1^{st} order LP analysis is carried using Eqn. (5.9) as shown in the block diagram. The residual signal obtained from 1^{st} order LP analysis is the foreground enhanced speech signal $s_t(n)$ minus the spectral tilt. Hence, the LP filter estimation using Eqn. (5.10) would model the vocal tract information without the spectral tilt. Therefore, the foreground speech signal $s_t(n)$ subjected through $H_{vt}(z)$ represented by

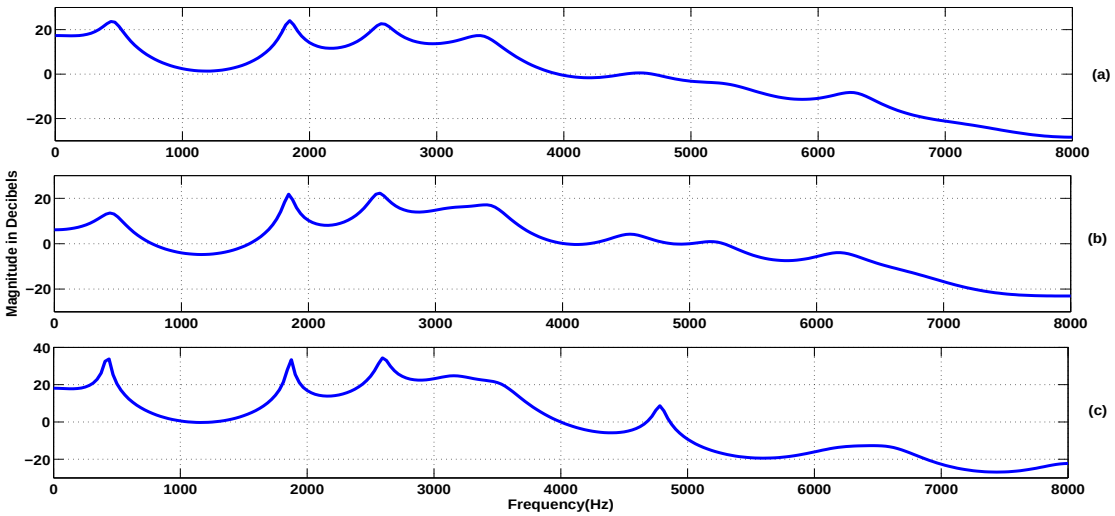

Figure 5.5: Illustration of formant enhancement. (a) LP magnitude spectrum obtained from a 20 *ms* voiced segment before formant enhancement (b) LP magnitude spectrum of the intermediate stage $H_{vt}(z)$ and (c) LP magnitude spectrum after formant enhancement.

Eqn. (5.10). This will further enhance the formant peaks relative to adjacent valleys while maintaining the spectral tilt. The output speech signal $s_f(n)$ is formant enhanced foreground speech signal. The formant enhancement is illustrated using LP magnitude spectrum plots obtained from voiced frame before and after formant enhancement as shown in Figure 5.5. The Figure 5.5(a) shows the LP magnitude spectrum of coefficients derived from the speech signal $s_t(n)$. The Figure 5.5(b) shows the LP magnitude spectrum $H_{vt}(z)$. It can be noticed that the LP magnitude spectrum is similar to Figure 5.5(b) except that the magnitude response is relatively flat. It can be noticed that the LP spectrum derived from speech signal $s_f(n)$ as shown in Figure 5.5(c) has sharper formant peaks relative to Figure 5.5(a). Also, the peak to adjacent valley ratio is increased.

This is further illustrated using Figure 5.3(c) and (g), which shows the foreground speech signal $s_f(n)$ after formant enhancement and its corresponding narrow band spectrogram, respectively. It can be noticed that there is a reduction of background noise and formant tracks are enhanced in the foreground regions. The effect of passing foreground speech signal $s_t(n)$ through LP filter makes formant peaks much sharper than the original signal by moving poles of the filter closer to the unit circle. Hence, Sections 5.2.1 and 5.2.2 use the production aspects to

enhance the foreground speech signal. The enhancement of foreground speech signal is achieved without exclusive modeling of noise, and such enhancement does not introduce unwanted musical noise to the enhanced speech signal. Since the poles of the all-pole LP filter moves close to unit circle in case of such formant enhancement, this makes speech sound unnatural [131]. Consequently, further processing is necessary to perceptually make the speech more natural and enhance the foreground speech from any left over residual noise.

5.2.3 Perceptual based Foreground Speech Enhancement

The formant enhanced foreground speech signal obtained as explained in Section 5.2.2 is further subjected to enhancement using cepstral analysis and synthesis on the Mel frequency scale. However, the cepstral analysis involves deriving MCCs using which the enhanced speech signal is synthesized through MLSA filter [132]. The detailed description of MCC is given in Appendix-A. The advantage of using MCCs is two fold, where, the coefficients can be directly used in speech and speaker recognition applications apart from foreground enhancement. The MCCs $C_\alpha(m)$ are the Fourier cosine coefficients of the spectral envelope derived from Mel log spectrum. The spectrum represented by MCCs closely resembles the human auditory spectral resolution having higher resolution at lower frequencies and lower resolution at higher frequencies [133]. The MLSA filter is applied to get the vocal tract response from the MCCs approximately using the adaptive algorithm. The true spectrum of MLSA filters for m^{th} order MCCs $c(m)$ is given by

$$H^\alpha(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{5.11}$$

is an all-pass function, which represents the mel-warped frequency characteristics and α is a coefficient corresponding to the mel-scale (for example $\alpha = 0.35$ for 10kHz sampling rate)

$$\beta_\alpha(\Omega) = tan^{-1}\frac{1 - \alpha^2 sin(\Omega)}{(1 + \alpha^2)cos(\Omega) - 2\alpha} \tag{5.12}$$

where, α depends on sampling frequency and $\beta_\alpha(\Omega)$ is the phase of all-pass function, the smooth spectral envelope $G_\alpha(\tilde{\Omega})$ of the mel log spectrum is expressed as polynomial function of order

M given by

$$G_\alpha(\tilde{\Omega}) = \sum_{m=0}^{M} C_\alpha(m) cos(m\tilde{\Omega}) \tag{5.13}$$

where, $\tilde{\Omega}$ is mel frequency scale given by $\beta_\alpha(\Omega)$ and $C_\alpha(m)$ are the cepstral coefficients of order M.

In order to compute a 34 dimensional MCC, a Hamming windowed frame of size 20 ms with a frame shift of 10 ms is considered. The smoothed spectral envelope is computed from the MCCs using MLSA filter. The MLSA filter provides the best mean square approximation of log spectrum envelope on the linear frequency scale and further used to directly synthesize the best quality speech signal. The MLSA filter needs excitation signal along with MCCs in order to synthesize the speech signal. The excitation signal is synthesized using F_0 information along with the voiced/unvoiced decision. Alternatively, in the current work, WLPR $r_w(n)$ is used as excitation signal along with MCCs to synthesize the perceptually enhanced foreground speech signal $s_p(n)$. This is illustrated in Figure 5.6, where, Figure 5.6(a) shows the log magnitude spectrum of 20 ms voiced frame taken from the original recording. Figure 5.6(b) shows the log magnitude spectrum from excitation based enhancement (EBE) method, while Figure 5.6(c) illustrates the log magnitude spectrum of FBE method. Figure 5.6(d) shows the smoothed log magnitude spectrum derived from MLSA filter. It can be observed that the smoothed log magnitude spectrum forms the envelope of the spectrum shown in Figure 5.6(c). Consequently, the log magnitude spectrum of speech signal synthesized using MLSA filter using WLPR signal $r_w(n)$ and MCCs is shown in Figure 5.6(e). It can be noticed that the sharpness of the formant peaks is relatively reduced. This helps to get rid of the unnaturalness that was introduced in Section 5.2.2 due to formant enhancement. Also, the current block further attenuates the remaining background noise. The Figures 5.3(d) and (h) shows the perceptually enhanced speech signal and its narrowband spectrogram, respectively. It can be observed that the sharpness of formant locations are reduced and further there is attenuation of background noise.

The various steps involved in foreground speech segmentation and enhancement is summarized in Table I.

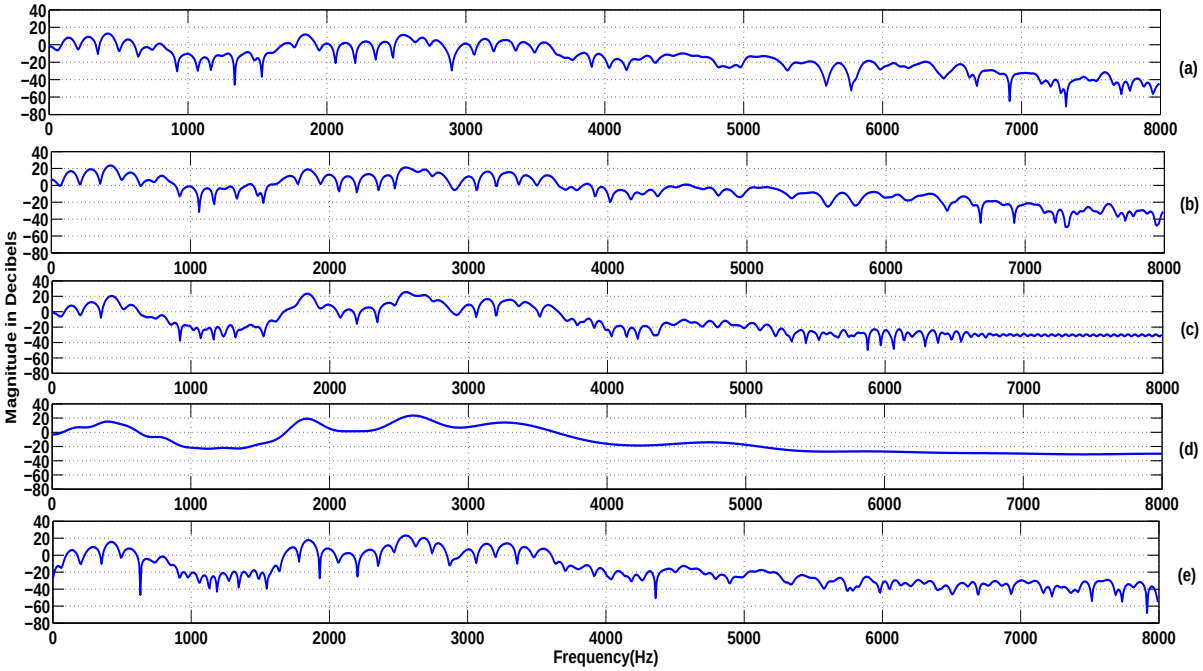

Figure 5.6: Illustration of different outputs obtained from 20 *ms* voiced frame using log magnitude spectrum of (a) original speech recording in foreground scenario, (b) excitation based enhanced output, (c) formant based enhanced foreground speech, (d) smoothed envelope obtained from MCCs using MLSA filter and (e) perceptually enhanced foreground speech signal using MLSA filter.

5.3 Experimental Results and Discussions

The proposed scheme of foreground speech enhancement is evaluated with some of the well-known speech quality measures along with 3 states of the art methods considered. In order to evaluate the proposed work, both subjective and objective measures are used. The mean opinion scores (MoS) and preference test scores are used for subjective analysis which are discussed in Sections 5.3.1.1 and 5.3.1.2 in detail. The clean speech signal is seldom available while recording in noisy environments. Therefore, two objective measures *viz.,* foreground-to-background-ratio (FBR) similar to *a posteriori* SNR and epoch-to-non-epochal-ratio (ENR) measurements are introduced in Section 5.3.1.2. Also, all methods considered for evaluation are compared using FBR and ENR measurements. The data set consisting of natural recordings in noisy environments are used to evaluate different methods using MoS, preference test, FBR, and ENR measures. Subjective analysis of enhancement methods is time-consuming task and mostly

Table 5.1: Foreground Speech Segmentation and Enhancement

.

- **Foreground Speech Segmentation:**
  - Pass the speech signal $s(n)$ through *Zero Band Filtering* to obtain *ZBFS*.
  - Using *ZBFS*, compute *NACC* and *ZBFS* energy.
  - Modulation spectrum is computed using analytic signal $\tilde{s}(n)$.
  - Final evidence $E(n)$ is additive combination of all three features.
  - The gross level feature $w_g(n)$ is obtained by passing $E(n)$ through sigmoidal function having slope parameter $\lambda = 20$ and T value derived from mean of $E(n)$.
- **Excitation Source based Enhancement:**
  - The epoch locations are obtained from *ZBFS* and fine weight $w_f(n)$ are derived by convolving Hamming window function having the temporal width of 3 ms.
  - Final weight function $w(n)$ is obtained using $w_g(n)$ and $w_f(n)$.
  - Compute LP residual signal of foreground segmented speech.
  - Multiply the weight function $w(n)$ with LP residual to obtain WLPR signal.
  - Excitation source based enhanced signal $s_t(n)$ is synthesized using WLPR and LP coefficients.
- **Formant based Enhancement:**
  - Perform 1^{st} order LP analysis on $s_t(n)$.
  - Inverse filtering of foreground speech signal to obtain the LP residual signal minus the spectral slope.
  - p^{th} order LP analysis on 1^{st} order LP residual signal to calculate the LP coefficients.
  - The foreground speech signal $s_t(n)$ is passed through LP filter to obtain the formant enhanced foreground speech signal $s_f(n)$.
- **Perceptual based Enhancement:**
  - Compute 34^{th} dimensional *MCCs*.
  - Smoothed magnitude spectral envelope is estimated using efficient *MLSA* filter.
  - The perceptually enhanced foreground speech signal $s_p(n)$ is synthesized by passing *WLPR* as modified excitation source and the smoothed magnitude spectral envelope estimated from *MCCs* through *MLSA* filter.

it is often difficult to get the right subjects for subjective evaluations. Therefore, perceptual evaluation of speech quality (PESQ) is one of the important objective measures that closely resembles subjective analysis [134]. However, PESQ measurement requires reference clean speech signal for evaluations. Hence, TIMIT speech database is used to benchmark the proposed work [122].

5.3.1 Performance Evaluation of Foreground Speech Enhancement using Natural Recordings

In order to evaluate the proposed method, the speech files recorded in natural environment from ten different speakers are considered which includes 3 female and 7 male speakers. Typically each speaker has spoken two to three different sentences in each such recordings. The spoken sentences are chosen from English radio broadcast and it has the composition of 74.10% of voiced sounds and 25.90% unvoiced sounds. All speakers were native Indians and were well versed with English as their second language. The speech is recorded in 7 different natural environments that includes busy city road traffic, office room when mosaic machine is operating in the background, air condition machine room, home environment when television is ON, room environment when background music is playing along with vocals, crowded hostel mess (babble noise) and building construction site when concrete mixing machine is ON. All speech files are recorded in foreground scenario, where, the foreground speaker is closer to microphone sensor relative to other interfering sources. The recording setup includes headphone along with microphone connected to laptops. The speech files are recorded using WAVESURFER [116] tool at the sampling rate of 16 kHz. Three different headphone sets of different make and prices are used for the recordings in order to maintain the variability of sensors. It can be noted that the headphones used for recording had no special front end pre-processing circuits to enhance the speech signal. In all such recordings, the speaker is wearing the headphone and the typical microphone is closer to the mouth of the speaker (within 1 to 2 inches). The background acoustic sources are far away from microphone sensor compared to foreground speaker.

The proposed method is compared with three other methods from the existing literature.

The spectral domain subtraction based method is one of the earliest methods for enhancing the speech signal distorted by additive noise. In spectral subtraction based methods, the noise components are modeled by average magnitude spectrum using several frames of noise-only regions. The noise spectrum is subtracted from the signal spectrum to obtain the enhanced output [2]. The method proposed in [4] assumes that the Fourier coefficients of speech and noise can be independently modeled as zero mean Gaussian random variables. The method aims to minimize the mean square error between the clean speech and enhanced speech signal and hence called as minimum mean square error (MMSE). There are several modifications suggested to these two basic methods in literature to improve the quality of enhancement by reducing the musical noise. The MATLAB implementations available in VOICEBOX [135] are used to compare with the proposed method in this chapter. Recently, a method is proposed that makes use of temporal and spectral processing to enhance the degraded speech signal [43]. The temporal processing relies on HELP signal to extract the epoch locations and enhance such locations using LP residual signal similar to excitation based enhancement module in this work. The signal is enhanced in spectral domain by a comb-like function that emphasizes the fundamental component and its harmonics. Both temporal and temporal-spectral enhancement are compared with the proposed method.

5.3.1.1 Subjective Evaluation using MoS Score

The subjective evaluation is carried using three parameters as foreground speech (FGS), background noise (BKG) and the overall quality (OVL) of enhanced speech signal. The original and enhanced speech files were provided to 24 different subjects in random order for subjective evaluation scores similar to [136]. The subjects included 16 male and 8 female adult listeners having an average age of 26 ± 5 years. All listeners are well versed with English language and English was their medium of instruction in their academic studies. The listeners are mainly working in the area of speech processing, speech synthesis, speaker verification and speech recognition topics and they typically have 1 to 3 years of experience in this domain. All listeners had normal hearing abilities without any kind of hearing impairment. The files were

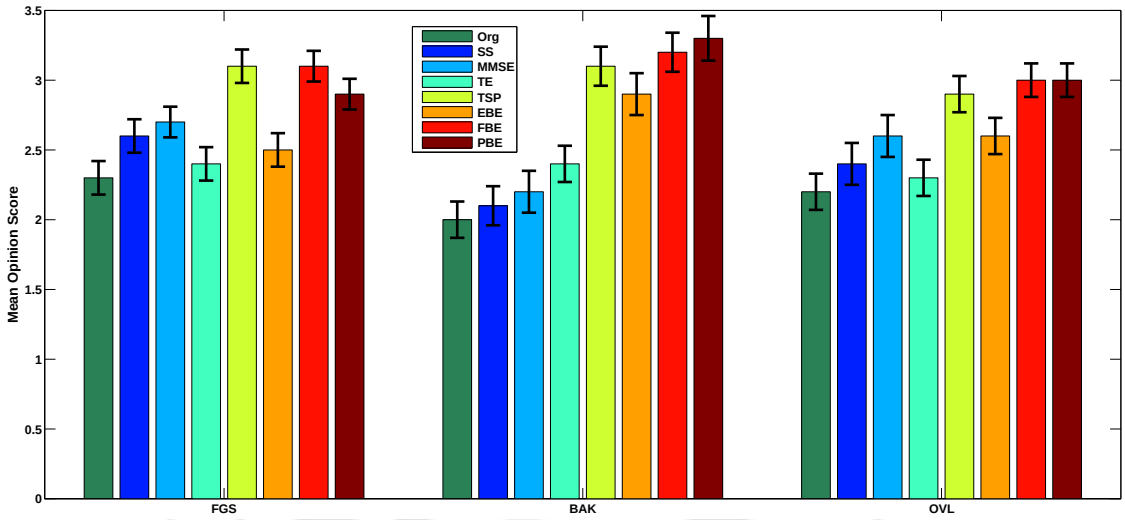

Figure 5.7: Bar graph representing the mean opinion scores obtained from different subjects, where, FGS is foreground speech, BKG is background noise and OVL is overall ratings. The graph also depicts the error computed using 95% confidence interval from the subjective scores. Org - Original, SS - spectral subtraction, MMSE - minimum mean square error, TE - temporal enhancement, TSP - temporal spectral processing, EBE - excitation based enahancement, FBE - formant based enhancement and PBE - perceptual based enhancement

listened to in a typical lab environment when AC is switched ON. The files listened without visualizing the waveform using similar types of headphones connected to PC platform. The subjects were provided with the following instructions for each of such parameters.

- Focus listening on FGS regions alone in terms of reduced background noise, intelligibility and lesser distortion with scales suggesting [1 - Very Unnatural, 2 - Fairly Unnatural, 3 - Somewhat Natural, 4 - Fairly Natural, 5 - Very Natural].

- Focus listening on BKG regions alone in terms of reduced background noise, and lesser distortion with scales suggesting [1 - Very Intrusive, 2 - Somewhat Intrusive, 3 - Noticeable but not Intrusive, 4 - Somewhat Noticeable, 5 - Not Noticeable].

- Focus listening on the OVL in both foreground and background regions in terms of reduced background noise, intelligibility and lesser distortion with scales suggesting [1 - Bad, 2 - Poor, 3 - Fair, 4 - Good, 5 - Excellent].

The subjects were provided with three sets of same files arranged in random order for evaluation. The Figure 5.7 shows the results of subjective evaluation in terms of MoS using the original and enhanced speech files from different methods. The Figure 5.7 shows the bar graph plots along with a margin of error with 95% confidence interval. It can be observed from Figure 5.7 that all three scores i.e., FGS, BKG, and OVL is lower in the case of the original signal as expected and this forms the reference score to evaluate other methods. The subjective scores obtained for MMSE is better compared to SS in all three parameters. Also, MMSE based method is better than temporal enhancement in terms of FGS and OVL, while TE remains better in terms of BKG compared to SS and MMSE. Though conceptually TE and EBE are similar in approach, it can be noticed that EBE performs better than TE in all three parameters. The reason of EBE performing better than TE is because of better identification rate (IDR), identification accuracy (IDA), lower miss rate (MR) to extract epoch locations using ZBF compared to HELP [117]. Also, ZBF remains robust to noisy conditions and hence leads to lower distortion while temporally processing the LP residual signal. This helps to set the reasonably lower threshold value to attenuate the background noise in case of EBE. Consequently, further processing modules can benefit from such higher attenuation of background noise in the proposed work.

It can be noticed that TSP, FBE, and PBE methods perform better than SS, MMSE, TE and EBE methods. The TSP and FBE based methods outperform other methods in terms of FGS, while FBE and PBE based methods are better in terms of BKG and OVL. It can be noted that spectral processing in the case of TSP is useful in attenuating the valleys of spectral magnitude and formant peaks remain unmodified. In the case of FBE, the formant peaks are boosted further relative to spectral valleys. The formant locations are high SNR regions in the spectral domain and hence relative enhancement of formant peak locations, helps to reduce the background noise and enhance the foreground speech regions. This is evident from subjective evaluation using FGS and OVL parameters. Any residual background noise left over can be reduced further by PBE method and this can be observed from Figure 5.7 by the increased BKG and OVL. Also, it can be observed that MoS of TSP, FBE and PBE methods in terms

of FGS, BKG and OVL are better than TE and EBE methods.

5.3.1.2 Subjective Evaluation using Preference Test

Preference test is one of the simpler tests to assess the speech quality. A subset of speech files that were used for MoS evaluations in Section 5.3 are used in the preference test. The speech files recorded from 5 different speakers in 5 different environments are used. The speakers included 2 female and 3 male speakers and each have spoken an English sentence chosen from English Broadcast. Totally 25 different speech files are assessed by 10 different subjects that included 5 female and 5 male listeners. All listeners were adults and not having any hearing impairments and the average age of listeners is 28 ± 5 years. The listening environment was in a typical computer laboratory with relatively less background noise. The listeners used high-quality headphones of similar make to listen to the speech files. The files are enhanced using SS, MMSE, TE, TSP, EBE, FBE and PBE methods. A pairwise listening test was conducted, where, one of the speech files were either EBE, FBE or PBE while the other is SS, MMSE, TE or TSP. The listeners were asked to listen to reference file before listening to the pair of files. Here, the reference file is the original recording in foreground scenario. The 5 different noisy environments included mosaic machine noise, building construction noise, traffic noise, background music with vocals and background speech. The subjects were given with following instructions to assess the file

- Listen to reference file before listening to individual pair of files.
- Asses the speech files in terms of reduced background noise, lesser distortion and speech intelligibility by giving equal weightage for all 3 features.

Table 5.2 shows the preference test scores averaged for 5 different types of noisy files assessed by 10 different subjects. The percentage score refers to which subjects have preferred EBE, FBE and PBE over other methods. It can be observed that the subjects have preferred EBE over SS and TE, while MMSE and TSP score better compared to EBE. The EBE does not cause distortion in the form of musical noise. Also, due to better epoch extraction method, the

Table 5.2: Subjective evaluation of different methods using preference test score in terms of percentage (indicates the preference of proposed compared to other methods), where, SS - spectral subtraction, MMSE - minimum mean square error approximation, TE - temporal enhancement, TSP - temporal and spectral processing, EBE - excitation based enhancement, FBE - formant based enhancement, PBE - perceptual based enhancement

Method	SS	MMSE	TE	TSP
EBE	51	44	53	39
FBE	58	54	61	49
PBE	61	58	63	54

background noise is suppressed compared to TE method. However, the temporal enhancement alone is not sufficient to reduce the background noise and hence MMSE and TSP score better than EBE. In the case of FBE, it is preferred relatively higher compared to SS, MMSE, and TE. The formant enhancement helps to increase the spectral peaks which are essentially high SNR regions in spectral domain compared to spectral valleys. Hence, there is a reduction of background noise in the spectral domain and this may be the reason for preference over SS, MMSE, and TE. The sharpening of formant peaks leads to audible distortion and this can be the reason for low preference score compared to TSP. However, the spectral envelope is smoothened by PBE module to reduce the distortion and further eliminates any residual noise left in previous stages. Hence, overall PBE performs better in terms of preference score. In the case of preference test, it is difficult to assess the specific reason for the choice made by subject as the preference depends on all 3 factors. However, comparing preference test score with MoS test reveals that the overall trend of preference test score is similar to OVL of MoS from Section 5.3.

5.3.1.3 Foreground-to-Background-Ratio (FBR)

When speech signals are recorded in natural environments seldom speech and noise signals are available separately to *a priori* estimation of speech and noise signal powers, respectively.

Since the speech signals are recorded in natural environments and there can be many types of interfering noises including background speakers. As explained in Section 4.4 foreground speech segmentation is one of the reliable ways to temporally segment foreground speech from rest of the background regions. Hence, foreground segmentation $w_g(n)$ obtained from Eqn. (4.15) is used to segment the foreground and background regions using the following relation

$$f(n) = \begin{cases} 1, & \text{if } w_g(n) > \mu_{w_g(n)} \\ 0, & \text{otherwise} \end{cases} \tag{5.14}$$

where, $f(n)$ is the binary signal indicating foreground and background regions and $\mu_{w_g(n)}$ is the mean value derived from $w_g(n)$. Hence, the SNR computed using foreground segmentation is called foreground-to-background-ratio (FBR). The FBR can be computed using the following relationships

$$\hat{\sigma}_f^2 = \frac{1}{L_f} \sum_{n=0}^{l-1} s^2(n).f(n) \tag{5.15}$$

$$\hat{\sigma}_b^2 = \frac{1}{L_b} \sum_{n=0}^{l-1} s^2(n).(1 - f(n)) \tag{5.16}$$

$$FBR = 10\ log_{10} \frac{\hat{\sigma}_f^2}{\hat{\sigma}_b^2} \tag{5.17}$$

where, $\hat{\sigma}_f^2$ is foreground speech power estimation, L_f is the number of samples in foreground speech region, $s(n)$ is the naturally recorded speech signal, $f(n)$ is foreground background binary signal, $\hat{\sigma}_b^2$ is the background power estimation, L_b is number of samples in background region and FBR is the estimated Foreground-to-Background-Ratio from $s(n)$.

In order to study the characteristics of FBR, 10 different speech files from TIMIT database consisting 5 female and 5 male speakers are considered. The clean speech files are added with 4 different types of noise at different levels as shown in Figure5.8. The x-axis represents the added noise levels to speech files at 20, 15, 10, 5 and 0dB, whereas, the y-axis represents the FBR calculated using the relationship given in Eqn. (5.17) expressed in decibels. It can be observed that there is linear relationship between the noise added to clean speech files and the FBR estimated in all 4 noisy cases *viz.,* machine noise, babble noise, background music with vocals

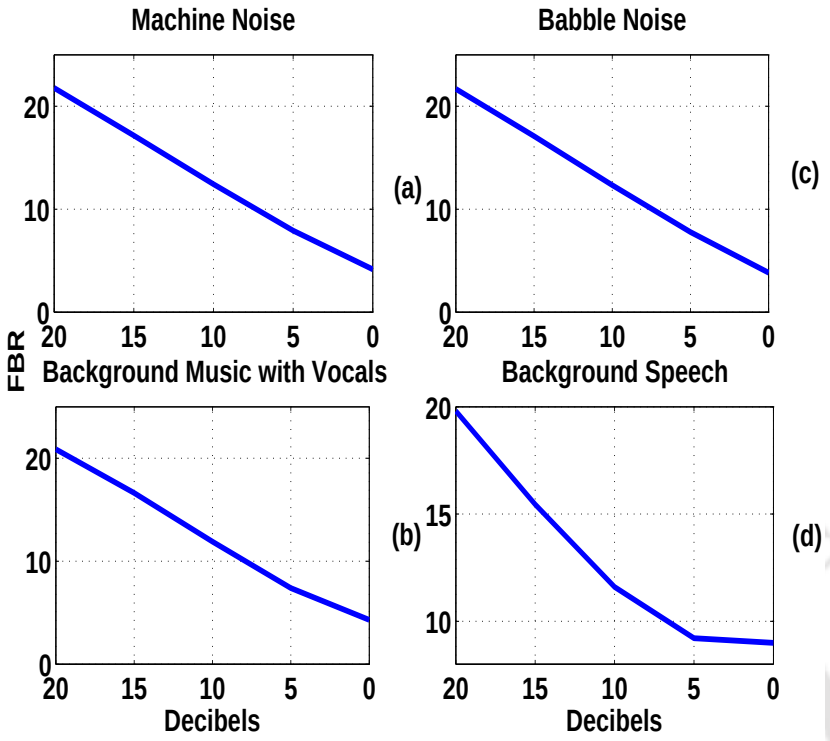

Figure 5.8: Average FBR obtained from 10 different TIMIT speech files by adding noise at different levels of 20, 15, 10, 5 and 0dB, where, (a) Mosaic Machine Noise (b) Hostel Mess Noise (Babble Noise) (c) Background Music with Vocals and (d) Background Speech.

and background speech. However, it can be noticed that there is degradation in estimating FBR in the case of background speech case at 0dB noise level. Overall the FBR estimate is reliable and robust to different types of additive noise at different levels. Hence, FBR can be used to measure *a posteriori* SNR that can help indicate the suppression of background noise by different methods.

The same set of files used to evaluate subjective measures as described in Section 5.3 are used to evaluate the performance of different methods using FBR. Totally 27 different speech files collected naturally in different noise environments are considered for FBR measurement. The FBR is computed for all the enhanced outputs using different methods. The objective scores are illustrated using bar graph in Figure 5.9 in terms of average FBR expressed in decibels (dB). Also, the plot indicates the margin of error with 95 % confidence interval. The set of files used to evaluate objective scores is same as used in subjective evaluation. It can be noticed that

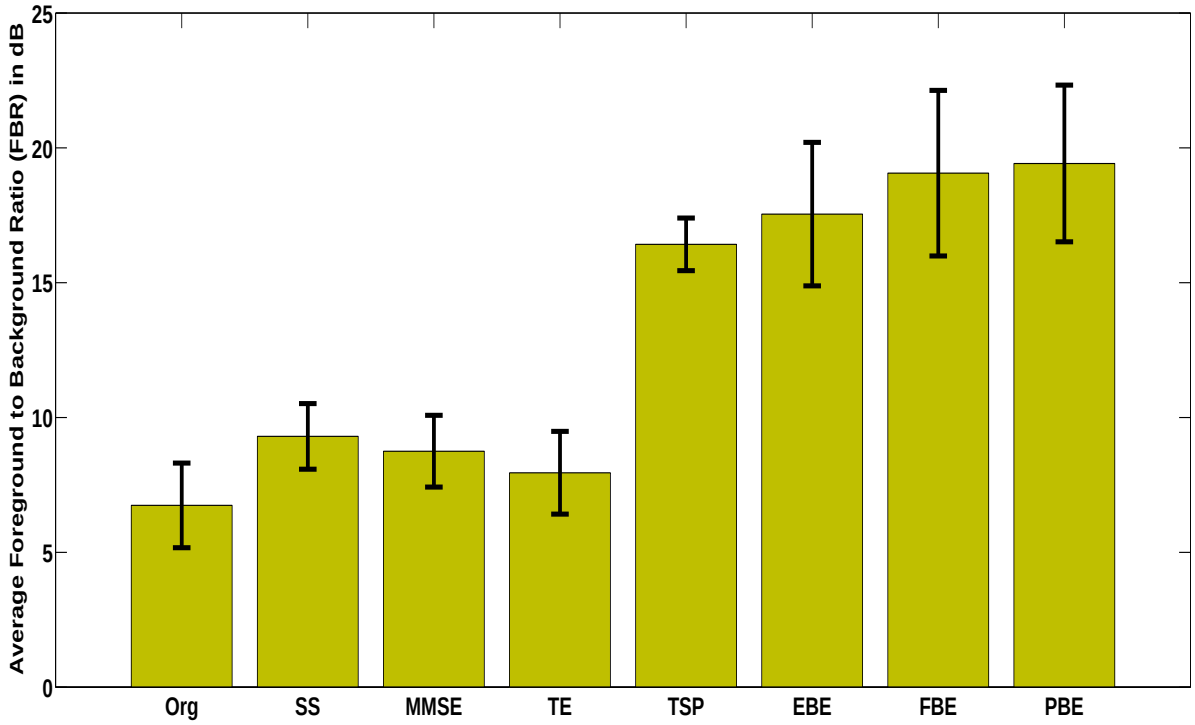

Figure 5.9: Bar graph representing objective scores in terms of average FBR obtained from original and processed outputs. The graph also depicts the error computed using 95% confidence interval from the objective scores.

the performance of TSP, EBE, FBE and PBE methods are superior compared to SS, MMSE and TE methods. The objective scores obtained further corroborates the subjective analysis as discussed above. However, the performance of FBE and PBE remains best in terms of FBR. The difference between the average scores of the original signal and the enhanced outputs from FBE and PBE shows that there is an improvement of 12 dB. This shows that the proposed method performs best in attenuating the background noise compared to other methods and still maintains the overall quality of enhanced foreground speech. The similar trend can be observed from subjective analysis using BKG parameter as shown in Section 5.3.

5.3.1.4 Epoch-to-Non-Epochal-Ratio (ENR)

The effect of interfering sources is not uniform throughout the foreground regions. There are certain regions of foreground speech that are relatively more robust to interfering sources, especially the regions around instants of significant excitation are high SNR regions compared

to other regions within a glottal cycle. Hence, the ratio between energy around epochal region to non-epochal region within a glottal cycle of foreground speech can be an important objective measure to evaluate different methods. In order to compute such a ratio, Hilbert Envelope of LP residual (HELP) signal is considered similar to [137], where, Hilbert Envelope of LP residual $e(n)$ is given by the following relationship

$$h(n) = \sqrt{e^2(n) + {e_h}^2(n)} \tag{5.18}$$

where, $e_h(n)$ is the HELP signal $e(n)$. The Hilbert transform is obtained by interchanging the real and imaginary parts of 1024 point DFT of $e(n)$ and then taking inverse DFT [101]. This measurement, when compared between the original recording and enhanced output will help to assess the performance of different methods in terms of enhancing high SNR regions further relative to other regions. Since, such a quantity essentially measures the energy ratio between epochal to non-epochal regions, hence, such a measurement is called as *Epoch-to-Non-Epochal-Ratio* (ENR).

The epochal energy is calculated by considering the summation of normalized energy around 3 ms of epoch locations, where, 3 ms closely corresponds to glottal closure interval. The non-epochal energy is computed by the summation of normalized energy excluding the 3 ms region around the epoch locations with reference to each glottal cycle. The ENR can be computed using the following relationships

$$\hat{E} = \sum_{k=1}^{N_k} \frac{\frac{1}{2M+1}\sum_{p=i_k-M}^{i_k+M} h_e^2(p)}{\frac{1}{L_1}\sum_{q=i_{k-1}+M+1}^{i_k-M-1} h_e^2(q) + \frac{1}{L_2}\sum_{r=i_{k+1}-M-1}^{i_k+M+1} h_e^2(s)} \tag{5.19}$$

and

$$\hat{O} = \sum_{k=1}^{N_k} \frac{\frac{1}{2M+1}\sum_{p=i_k-M}^{i_k+M} h_o^2(p)}{\frac{1}{L_1}\sum_{q=i_{k-1}+M+1}^{i_k-M-1} h_o^2(q) + \frac{1}{L_2}\sum_{r=i_{k+1}-M-1}^{i_k+M+1} h_o^2(s)} \tag{5.20}$$

where, $h_e(p)$ is HELP derived from enhanced foreground speech, $\hat{E}$ is the estimation of ENR using enhanced foreground speech, N_k is the total number of epochs in the foreground speech regions, M corresponds to samples of 1.5 ms, i_k is the epoch location at k^{th} epoch, i_{k-1} is the previous epoch location to i_k, i_{k+1} is the epoch location after i_k, $L_1 = i_k - i_{k-1} - 2M - 2$,

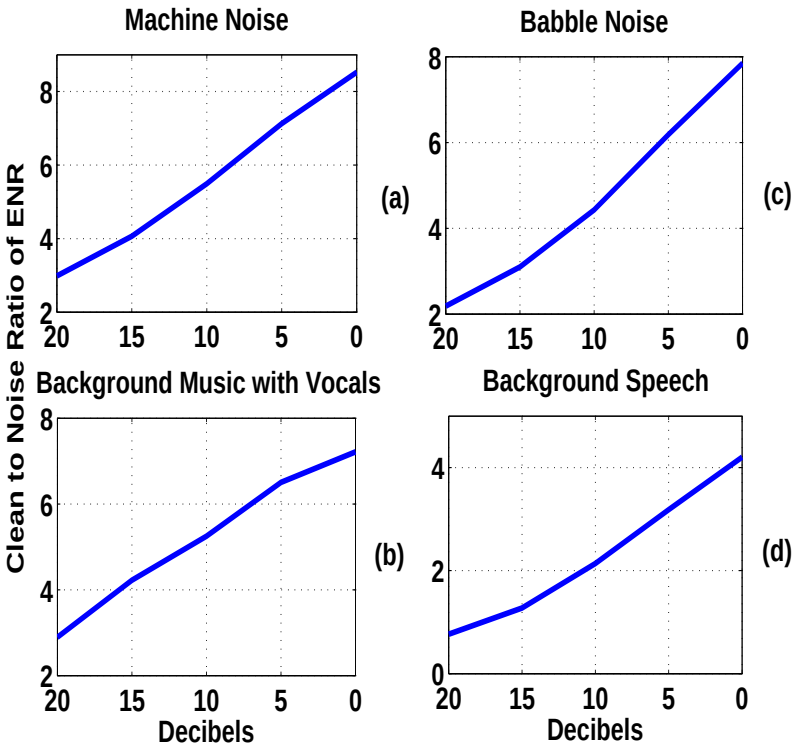

Figure 5.10: Average ENR obtained from 10 different TIMIT speech files by adding noise at different levels of 20, 15, 10, 5 and 0dB, where, (a) Mosaic Machine Noise (b) Hostel Mess Noise (Babble Noise) (c) Background Music with Vocals and (d) Background Speech.

$L_2 = i_{k+1} - i_k + 2M + 2$, $\hat{O}$ estimation of ENR using HELP of original recording $h_o(p)$. The ratio between $10log_{10}\frac{\hat{E}}{\hat{O}}$ represents the improvement achieved in terms of enhancing high SNR regions further relative to low SNR regions. In order to characterize ENR a similar approach that is followed in Section 5.3.1 for FBR is followed. The 10 speech files from different speakers from TIMIT database is additively corrupted using 4 different types of noise at different levels. The Figure 5.10 shows the ENR evaluation for clean and additive noise cases. It can be observed that there is a linear relationship between additive noise and ENR for all 4 different noise types. It can be noticed that ENR increases linearly as additive noise increases.

The ENR is computed for all the enhanced outputs using different methods. The objective scores are as shown in Table 5.3 that is computed by averaging the scores from all files. It can be noted that the ratio is computed between enhanced output to original recording while computing ENR. It can be observed that EBE and PBE methods are best amongst all other

Table 5.3: Objective evaluation of different methods using Epoch-to-Non-Epochal-Ratio (ENR), that is computed as a ratio between ENR of enhanced foreground speech to original recordings. The table represents the average ratio expressed in decibels (dB) computed across all the enhanced speech files obtained from different methods.

SS	MMSE	TE	TSP	EBE	FBE	PBE
1.28	0.83	3.68	2.31	4.23	1.26	5.46

methods in terms of ENR. The improvement achieved in the case of EBE is due to better identification of epochs using ZBF, as a result of which the noise components are suppressed in LP residual to obtain WLPR. The speech signal synthesized using WLPR results in enhanced high SNR regions relative to low SNR regions. However, in the case of PBE, the spectral envelope is smoothened and this leads to a reduction of background noise, mainly in low SNR regions. Hence, EBE and PBE methods have high ENR compared to other methods. In the case of FBE, due to LP filter enhancement, the formant peaks are sharpened and thereby increasing the gain of the filter transfer function at formant locations. The enhanced LP filter convolves with excitation signal resulting in relatively higher amplitude at non-epochal regions and hence the ENR is lower in the case of FBE. The enhancement of foreground speech signal using FBE is achieved mainly because of enhancement of high SNR regions in the spectral domain.

5.3.2 Perceptual Evaluation of Speech Quality (PESQ)

The PESQ is carried using 10 TIMIT speech files taken from 5 female and 5 male speakers. The clean speech files are corrupted by adding 5 different types of noise at 3 different levels. The Table 5.4 shows the average PESQ scores obtained from 10 different speech files. It can be observed that in the case of mosaic machine noise (MMN) the spectral subtraction method performs well. Since MMN is nearly stationary noise and hence SS is able to better model the background noise in such cases. However, in case of hostel mess noise (babble noise) and traffic noise which are relatively non-stationary in nature, TSP, and PBE methods performs better than other enhancements. The background music with vocals and background speech cases are

Table 5.4: Perceptual Evaluation of Speech Quality Scores, where, MMN - Mosaic Machine Noise, MN - Hostel Mess Noise, TN - Traffic Noise, BM - Background Music with Vocals, BS - Background Speech SS - spectral subtraction, MMSE - minimum mean square error approximation, TE - temporal enhancement, TSP - temporal and spectral processing, EBE - excitation based enhancement, FBE - formant based enhancement, PBE - perceptual based enhancement

Noise	Decibels	Original	SS	MMSE	TE	TSP	EBE	FBE	PBE
	15	2.5	**3.2**	3.1	2.9	3.0	2.9	3.0	3.1
MMN	**10**	2.3	**3.1**	3.0	2.8	2.9	2.8	2.9	3.0
	5	1.8	**2.7**	2.6	2.4	2.5	2.5	2.5	2.6
	15	2.6	3.0	2.9	2.8	**3.1**	2.8	2.9	3.0
MN	**10**	2.2	2.7	2.6	2.7	**3.0**	2.7	2.8	2.9
	5	1.9	2.3	2.2	2.5	**2.6**	2.4	2.5	**2.6**
	15	2.4	2.6	2.5	2.8	**2.9**	2.7	2.8	**2.9**
TN	**10**	2.2	2.3	2.2	2.7	**2.8**	2.6	2.7	**2.8**
	5	1.8	2.0	2.0	2.2	**2.3**	2.0	2.1	2.2
	15	2.4	2.8	2.7	2.7	**3.0**	2.8	2.9	**3.0**
BM	**10**	2.2	2.4	2.3	2.5	2.7	2.6	2.7	**2.8**
	5	1.8	2.1	2.0	2.2	**2.4**	2.2	2.3	**2.4**
	15	2.8	2.8	2.7	2.8	3.0	2.8	3.0	**3.1**
BS	**10**	2.7	2.7	2.6	2.7	2.9	2.8	2.9	**3.0**
	5	1.9	1.8	1.8	2.2	2.3	2.3	2.3	**2.4**

the most challenging, as the background noise characteristics are similar to foreground speech regions. Due to the robustness of foreground speech segmentation and better estimation of GCI locations using ZBF the performance of the proposed work is relatively better compared to other methods. Overall the performance of TSP and PBE are better and comparable. Though, it is difficult to correlate all the parameters of MoS with PESQ scores, the trend remains consistent with subjective analysis.

5.4 Summary

In this chapter, a new way to approach the problem of speech enhancement is suggested, where, the distance between the foreground speaker to microphone and rest of the background sources is utilized. The proposed work relies on known production and perceptual features to enhance the foreground speech. The advantage of proposed method is that the distortion is significantly lower and does not introduce unwanted musical noise like other methods as spectral subtraction and MMSE. The method exploits reliable ZBF for foreground segmentation and extraction of glottal closure instants due to which higher attenuation of background noise is possible with least distortion. The advantage of using ZBF is that there is no necessity of finding F_0 of foreground speaker. The performance of proposed work is compared with 3 other existing state of the art methods in terms of subjective and objective evaluations. It is found that the proposed method can significantly attenuate the background noise and still maintain the better quality of enhanced foreground speech signal. As illustrated in Figure 5.1 the enhanced speech signal can be used in different speech-based applications and that may include a practical spoken query system.

6

Robust Spoken Query System using Foreground Speech Segmentation and Enhancement

Contents

Objective

A mobile based spoken query system can be accessed from an uncontrolled environment and typically the desired speaker poses a query in foreground scenario. However, the degradation of speech signal due to the presence of other interfering sources can impact spoken query system adversely. Hence, using foreground speech segmentation and enhancement modules proposed in previous chapters as front end pre-processing blocks can help improve the robustness of a practical spoken query system. The work exploits the speech production aspects like robust components of excitation source in time domain and vocal tract information in the spectral domain to temporally segment the desired speaker's speech and further enhance it. It is shown that using such front end segmentation and enhancement modules for training acoustic models and handling the spoken query can be an effective solution in a practical scenario. The advantage of foreground speech segmentation is demonstrated using the improved performance of automatic speech recognition (ASR). Further, the robustness of foreground speech enhancement is shown by evaluating the performance of ASR under additive noise conditions. The performance of the proposed work is compared with other considered state of the art techniques. The ASR module in spoken query system is based on the hidden Markov model (HMM). Three different techniques for modeling the observation probabilities are explored in this work *viz.,* Gaussian mixture model (GMM), subspace GMM (SGMM) and deep neural network (DNN).

6.1 Spoken Query System in Natural Environment

Speech is one of the natural forms of human to human communication. Hence, it is ubiquitous to extend this mode of communication for human-computer interaction (HCI) and automatic speech recognition (ASR) is an integral part of such interaction. Spoken query system is an interactive system which is specific to the application domain and the complexity of the system varies based on its usage [138]. Based on the complexity and restriction on humans natural flow of language, spoken query systems can be broadly categorized as command and control, directed dialog, and natural language [139]. The command and control system is a

heavily constrained system to the user and it is limited to yes/ no or digits in voice to replace DTMF inputs. The command and control system suffers least from word error rate (WER) as vocabulary is highly restricted. On the contrary, natural language systems often pose no restrictions on the speaker's vocabulary and mostly find their applications in dictation based systems. Alternatively, directed dialog system lies in between in terms of complexity and vocabulary. The directed spoken query system guides the user through speech based prompts to the desired goals and mostly they are handled in a remote fashion.

In modern day scenario, due to higher penetration of mobile phones amongst people, many novel applications are possible. The spoken query-based systems for farmers can be a solution to obtain the price of commodities in particular districts [140,141]. This system can be categorized as directed dialog system based on the complexity and vocabulary size. There are many aspects of such a practical system that can act adversely to its functioning. More commonly there will be no control over the environment in which the users can access such system. The desired spoken query gets recorded along with other interfering sources at the background [62]. The presence of such background noise can have an adverse effect on recognition performance and in turn, impacts the delivery of commodity price to users. The task is complicated if the interfering sources are from other background speakers or from the background sources having speech-like characteristics.

One way to handle such an issue is to model different background noise [142]. However, it requires a large amount of labeled noisy files to model different kinds of noise. It is not always practically feasible to capture all types of background noise sufficiently to build models for each type of background noise. Modeling different types of noise can impact the systems performance for unseen types of background noise. Alternatively, it is necessary to have a robust front end pre-processing module that can temporally segment the desired speaker's speech from rest of the background noise and further enhance it [32, 33]. The objective of such segmentation and enhancement should be such that the recognition performance of spoken query system must be improved under degraded conditions.

Speech enhancement is one of the active areas of research and several methods have been

proposed in the literature. The general approach of most methods is to model the additive noise only components from the noisy speech signal and suppress them to obtain the enhanced speech [2,4]. The noise modeling depends on efficient segmentation of additive background noise from rest of the speech regions. The suppression of noise usually takes place in the spectral domain through subtraction. However, such spectral subtraction methods introduce distortions in enhanced speech because of overestimating the noise spectrum in the form of undesired musical tones. Many methods were proposed in order to overcome this problem [75, 124]. The human auditory perceptual cues were considered to improvise the spectral subtraction methods [38, 126]. The objective is not to completely suppress the additive noise. Rather, the purpose is to utilize the masking properties of the human auditory system and live with a certain amount of residual noise that is below the masking threshold. The goal of such enhancement methods available in the literature is to reduce the background noise and perceptually improve the quality of speech signal. However, the enhancement of speech signal perceptually may not be necessarily suitable for recognition purpose and at times can degrade the performance of spoken query system.

There is a requirement for speech enhancement method which helps to improve the perceptual quality of speech signal and still can improve the performance of spoken query system in noisy environments. In spoken query system it is safe to assume that the desired speaker is closest to mobile sensor compared to other interfering sources in normal speaking mode. Typically the desired speaker's mouth is closer to the microphone within 1 to 3 in, whereas the interfering sources are far away from the sensor. The speech signal so recorded from the desired speaker is termed as *foreground speech* and rest of the interfering sources are termed as *background noise* as explained in Section 4.3. There is evidence that when the speech signal is recorded in the natural environment, the speech production characteristics tend to vary depending on the levels of interfering sources [128]. Due to the close proximity of speaker to the microphone and also due to modified speech production characteristics because of acoustic feedback, there are significant differences in the nature of signal for foreground speech and the background noise as shown in Section 4.4. Hence, it is desirable to design a method that can effectively use such

properties of the signal to segment spoken query from rest of the background noise and further enhance it [33].

The interfering background sources do not affect all regions of foreground speech regions equally [8, 43]. In particular, the instants of significant excitation which consists of predominantly glottal closure instants (GCI) in time domain and formant locations in spectral domain are a relatively high signal to noise ratio (SNR) regions in foreground speech. Hence, such high SNR regions are least affected by background noise and they remain robust. In this work, the relevant foreground query is segmented from rest of the background noise using high SNR regions and further used to enhance the foreground speech. In order to enhance the foreground speech, the high SNR regions are further boosted relative to low SNR regions. The advantage of such segmentation and enhancement method is that it causes minimal distortion for synthesized speech signal, unlike other methods. The exaggeration of high SNR regions should help to discriminate the acoustic-phonetic units better and thereby help improve the performance of spoken query system. This chapter demonstrates the usefulness of foreground speech segmentation and enhancement on Assamese spoken query system by improving the overall recognition performance significantly.

The effectiveness of foreground speech segmentation and enhancement is studied in terms of improved speech quality in Sections 4.4 and 5.2. It can be noted that the study was conducted using wideband headphone recordings. However, the merits of the proposed work in this chapter is to study the effect of using foreground speech segmentation and enhancement as a front end pre-processing module in a spoken query system using narrowband telephone recordings. Three different acoustic modeling techniques *viz.,* Gaussian mixture model-hidden Markov model (GMM-HMM), subspace Gaussian mixture model-hidden Markov model (SGMM-HMM), and deep neural network-hidden Markov model (DNN-HMM) are used to study the impact of such modules on ASR performance. Though a preliminary study using a front end foreground speech segmentation for its effectiveness on improving ASR performance is reported in [143], this work extends the study to use foreground speech segmentation in tandem with enhancement as a pre-processing module.

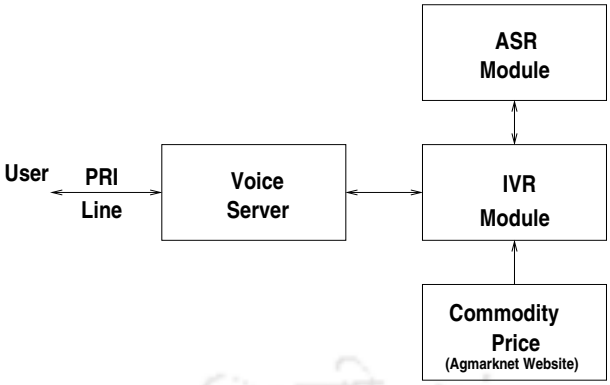

Figure 6.1: The overall block diagram of spoken query system.

The rest of the chapter is organized as follows: Section 6.2 briefly describes the spoken query system used to demonstrate the effectiveness of foreground speech segmentation and enhancement. The Section 6.3 explains briefly different modules involved to generate the multistage outputs and the details of foreground speech segmentation module. Sections 6.4.1 describes the details of excitation source based enhancement, while, Section 6.4.2 explains the details of formant based enhancement. Section 6.5 briefly describes different acoustic modeling techniques used to realize ASR and further demonstrates the effectiveness of using foreground speech segmentation and enhancement as front end processing through improved performance of ASR. The summary of the present work is described in Section 6.6.

6.2 Assamese Spoken Query System

The overall block diagram of the spoken query system realized is as shown in Fig. 6.1. The spoken query system is developed to deliver the price of a particular agricultural commodity belonging to a district in Assam state. Primarily the spoken query system consists of voice server, ASR module, interactive voice response (IVR) module, and a web-crawler to access the commodity price. The users can pose the query to a voice server using their mobile phones through primary rate interface (PRI) line. The query itself is limited to an isolated word of commodity and district names in The Assamese language. However, the call flow incorporates yes or no confirmation from users at decision nodes when the confidence of recognition is

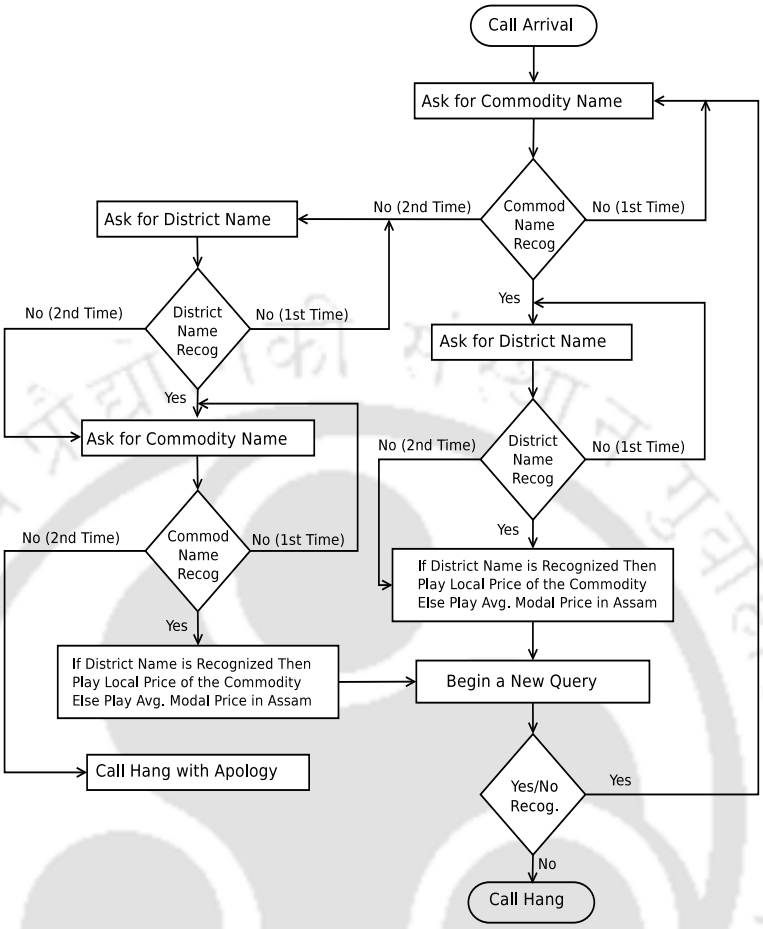

Figure 6.2: The IVR call flow of the spoken query system.

low [141]. The IVR mainly handles the user spoken query through a call flow. The developed system allows user interaction following a call flow as shown in Figure 6.2.

There are two branches in the call flow through which a query may get processed. On making a phone call to the system, initially, the user is prompted to speak the commodity name for which the price is desired. A maximum of two trials is given for its successful recognition. On successful recognition, the user is next prompted to speak the district name in which the price is sought. In case the district name is recognized correctly, in a maximum of two attempts, then the price of the desired commodity in the mentioned district is played back to the caller. On the other hand, if the system fails to recognize the district name correctly, the average modal price across all the districts in Assam is played back instead of a local price. If the

Figure 6.3: Conference call setup for data collection.

commodity name recognition itself fails in the first branch even after two attempts, the user query is processed through the second branch in the call flow. In that branch, the order of query is reversed, asking the district name first followed by the commodity name. Price information disseminated may be district specific or average modal price across all the districts in Assam depending on whether the district name is recognized or not. This feature provides the system with a maximum of four attempts to recognize a particular commodity name and also helps in the conditioning of the user without monotony. In case the commodity recognition fails even after the fourth attempt, the call terminates with a sorry message. At each stage in the call flow, the caller is asked for a confirmation to validate the hypothesis generated by the ASR system. A number of confidence-building and latency reduction measures are incorporated and are discussed in the following subsections. The call flow is tuned based on user feedback to arrive at the structure shown in Figure 6.2.

The details of all the necessary modules to realize the spoken query system are explained in the Sections to follow.

6.2.1 Speech Data Collection

The practical goal of the spoken query system realized is to help farmers to update the price of commodity price. Hence, there is a control on the surroundings from which the farmers can pose the query. Therefore it is necessary to collect the spoken query data in natural

environments. The collection of task-specific data is done following a call flow similar to the one shown in Figure 6.2. During database collection, all the decision-making processes were taken care of by a wizard-of-Oz. The wizard-of-Oz sets up a conference call among the farmer, the asterisk server and an additional mobile as shown in Figure 6.3. Due to the absence of the ASR system, appropriate codes in dual tone multi-frequency signaling (DTMF) are keyed in by the wizard-of-Oz at the decision-making stages for controlling the call flow based on the user's response to IVR. As the mobile which sets up the conference call cannot send any DTMF input, an additional mobile is used to key in the appropriate DTMF codes. The speech corpus is collected from more than 1000 farmers across different regions of Assam. It covers the vocabulary of 143 words that cater to the commodity list in AGMARKNET that includes commodity and District names in Assam. Also, additional words of *Hoy/ Nohoy* (yes/ no) is collected for nodal confirmation purpose in the call flow.

The following points were taken into account while collecting the speech data:

(i) Assamese language has four major dialect groups:

- *Eastern group* is spoken in Sibsagar, Dibrugarh, Jorhat, Golaghat, Dhemaji, Lakhimpur and Tinsukia districts.
- *Central group* which is spoken in Nagaon, Sonitpur and Marigaon districts.
- *Kamrupi group* spoken in Nalbari, Barpeta, Bongaigaon, Darrang and Kokrajhar districts.
- *Goalporia group* which is spoken in Goalpara and Dhubri districts.

(ii) The speech data is so collected that it covers all the four major dialects of Assamese.

(iii) Multiple handsets and mobile services providers are used while collecting the data to capture possible variabilities

(iv) Data is collected from both male and female speakers in a ratio of 3:1.

(v) As the agricultural commodity names did not cover all the possible phonetic contexts, an additional three hours of speech data from 25 speakers (17 males and 8 females)

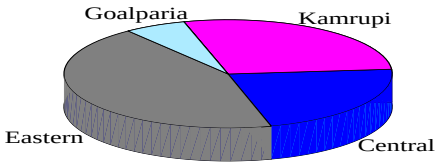

Figure 6.4: The distribution of the collected speech data across the different dialect groups of Assamese. *Eastern group* contains almost all of the districts of upper Assam. Consequently, it is spoken by a very large number of people and thus has the largest share in the speech data. The *Kamrupi group* contains some of the districts in lower Assam and hence has the second highest number of speakers. The *Central group* consists of three districts in middle Assam and so the number of speakers in this group is lesser than the other two. The *Goalporia group* is spoken mainly in Goalpara district and very little in other parts of Assam and hence, it has the least number of speakers.

speaking 28 phonetically balanced sentences is added to boost the context modeling by the triphones.

The Figure 6.4 shows the share of each of the dialect groups of Assamese in the collected speech data. For the system development and evaluation, the collected speech corpus is divided into non-overlapping training and testing sets balanced in terms of the dialect groups. The training set consists of 885 speakers (669 males and 216 females) while the testing set comprises of 275 speakers (206 males and 69 females).

6.2.2 Automatic Speech Recognition

The ASR is one of the important modules in the spoken query system. The performance of spoken query system is tightly coupled to the ASR accuracy. The ASR which is developed to recognize the names in The Assamese language is developed using 36 phonemes. The ASR in earlier work in [141] was realized using HTK [144]. However, in this work, the ASR is realized using Kaldi toolkits [145] to build all 3 acoustic models. Since the speech data is collected over telephone channel the sampling rate of both training and testing set is 8 kHz. The MFCC are used as features to parametrize the speech signal. In order to extract MFCC features, a Hamming windowed frame size of 20 ms with a frame shift of 10 ms is employed. The features

are cepstral mean normalized and include delta and delta-delta coefficients. The training phase involves building statistical HMM using iterative Baum-Welch re-estimation algorithm. Each monophone is modeled by 3 states left to right HMM model, while each state is being modeled by 16 GMM. However, the silence modeling is done using 32 GMMs. The triphone modeling is done by state tying and triphone models. The Viterbi decoder is used to decode the information of the test signal in the form of commodity and District names. Individual equally likely wordnets are used to decode commodity and District names and nearly 10% of collected data is used for evaluating the performance of the system. Also, the ASR is developed using SGMM-HMM and DNN-HMM-based acoustic models, the details are provided in Section 6.5.1.

6.2.3 Price Information Database

Once the commodity and district names are supposedly recognized by ASR, the next task is to disseminate the corresponding commodity price. The AGMARKNET website enlists the name and price of the commodities. The updated price of commodities in each of those districts are crawled every day from AGMARKNET website [146] and the latest prices are updated using MySQL database. A Crontab based web-crawler is used to update the price database regularly with the latest prices of the commodities listed on the AGRMARKNET website. The farmer's query is used to retrieve the price information stored in the database and the pre-recorded messages are used for generating the required voiced responses for information dissemination.

6.3 Foreground Speech Segmentation

The overall block diagram of proposed foreground speech segmentation and enhancement is as shown in the Fig. 6.5. Broadly the proposed work consists of 3 major modules to process the speech signal *viz., foreground speech segmentation (FGS)*, *excitation based enhancement (EBE)*, and *formant based enhancement (FBE)*. Both training and testing speech signals are subjected through same processing block as shown in Fig. 6.5. The multistage outputs derived from FGS, EBE, and FBE modules are denoted by $s_w(n)$, $s_t(n)$, and $s_f(n)$, respectively. The output of each such module is used to extract MFCC feature. In order to study the effect

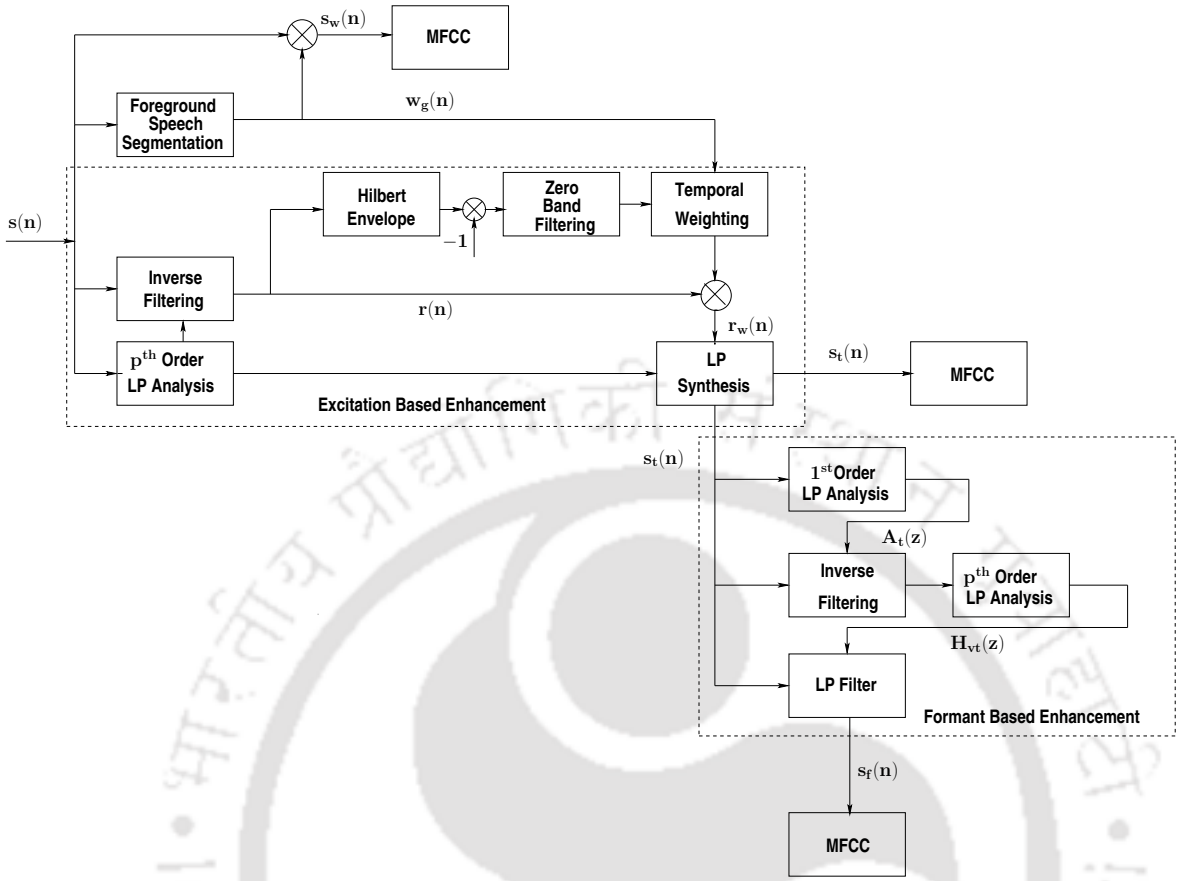

Figure 6.5: The overall block diagram of the proposed foreground speech segmentation and enhancement method, where, $s(n)$ is the input speech signal recorded in foreground scenario, $w_g(n)$ is the gross weight function that mainly segments the foreground speech regions from rest of the background noise, $w(n)$ is the final temporal weight function, $l(n)$ is the LP residual signal, $r_w(n)$ is the temporally weighted LP residual signal, $s_t(n)$ is excitation based enhanced output, and $s_f(n)$ is formant based enhanced output.

of individual outputs on spoken query system, separate acoustic models are built and the test results are evaluated. The objective of foreground speech segmentation is to temporally segment relevant foreground speech from rest of the background interference. The foreground speech segmented output $s_w(n)$ is derived by multiplying speech signal $s(n)$ with the gross weight function $w_g(n)$. The details of foreground speech segmentation module shown in Fig. 6.5 is explained in Section 4.4.

6.4 Foreground Speech Enhancement

The foreground speech enhancement mainly consists of excitation source and formant based enhancement modules and they are explained in detail in the following Sections.

6.4.1 Excitation Source based Enhancement

The epoch locations are instants of significant excitation which are high SNR regions and they are robust to interfering sources. Hence, such locations can be used as anchor points to enhance foreground speech region [43]. It is shown in [8] that modified linear prediction residual (LPR) by boosting GCI locations introduces the least distortion to the enhanced speech signal. Hence, it becomes imperative to accurately identify the epoch locations from foreground speech before temporal enhancement. It is shown in [117] that, the epoch locations can be extracted robustly using ZBF. However, ZBF works well for wideband speech signal as shown in Section 5.2, whereas the objective of the current work is to enhance spoken query over the telephone channel. The ZBF relies on frequency components closer to 0 Hz, while the components below 300 Hz are missing in telephone channel signal. Therefore, it is difficult to estimate epoch locations directly using ZBF from the speech signal. Alternatively, the epoch locations can be extracted using Hilbert Envelope of LP Residual (HELP) signal [58]. The linear prediction analysis is carried using the frame size of 20 ms and a frame shift of 10 ms. The linear prediction residual (LPR) derived through inverse filtering is random polarity signal and hence it is difficult to extract epoch locations [101]. However, HELP is a unipolar signal and also, this emphasizes epoch locations relative to other regions of LPR. The HELP signal is the magnitude of its complex time function (CTF) [147]. The LP residual signal acts as the real part of CTF and Hilbert transform of LP residual signal acts as the imaginary part. Since HE is a magnitude function and hence it is unipolar in nature. The Hilbert envelope $h(n)$ of LP residual sequence $l(n)$ is given by

$$h(n) = |l_a(n)|\,. \tag{6.1}$$

where, $l_a(n)$ is a complex time function and can be computed as follows,

$$l_a(n) = l(n) + jl_h(n) \tag{6.2}$$

where $l_h(n)$ is Hilbert transform of LP residual sequence $l(n)$. The Hilbert transform is computed as

$$l_h(n) = IDFT(L_H(\omega)), \tag{6.3}$$

where

$$L_H(\omega) = \begin{cases} +jL(\omega), -\pi \leq \omega < 0 \\ -jL(\omega), 0 \leq \omega \leq \pi \end{cases} \tag{6.4}$$

and $L(\omega)$ is DFT of $l(n)$. DFT refers to discrete Fourier transform and IDFT refers to inverse of DFT. Therefore the magnitude of complex time function $l_a(n)$ i.e., HE of LP residual is given by,

$$h(n) = \sqrt{l^2(n) + l_h^2(n)}. \tag{6.5}$$

The HELP is passed through ZBF given by Eqns. (4.4) and (4.5) to obtain zero band filtered signal (ZBFS). The positive zero crossings of ZBFS exactly estimates epoch locations of foreground speech signal. It should be noted that the polarity of HELP has to be reversed as shown in Fig. 6.5 before passing it through ZBF as the phase of ZBFS is important to extract epoch locations [148]. The Fig. 6.6(a) shows the signal recorded in foreground scenario and can be noticed that there is a significant amount of background noise present. The Fig. 6.6(c) illustrates HELP computed using Eq. (6.5). The GCI locations are extracted and they are used to derive weighted linear prediction residual (WLPR) as explained in Section 5.2. The WLPR computed from the signal is shown in Fig. 6.6(g), it can be noticed that background regions from LPR are attenuated and the epoch locations are further emphasized in foreground regions. The excitation based enhanced speech $s_t(n)$ is synthesized using WLPR as excitation source while retaining the original linear prediction coefficients (LPCs). This is illustrated in Fig. 6.6(h), and it can be observed that background regions are attenuated compared to original recording shown in Fig. 6.6(a). This is further illustrated using Fig. 6.7, where Figs. 6.7(a)

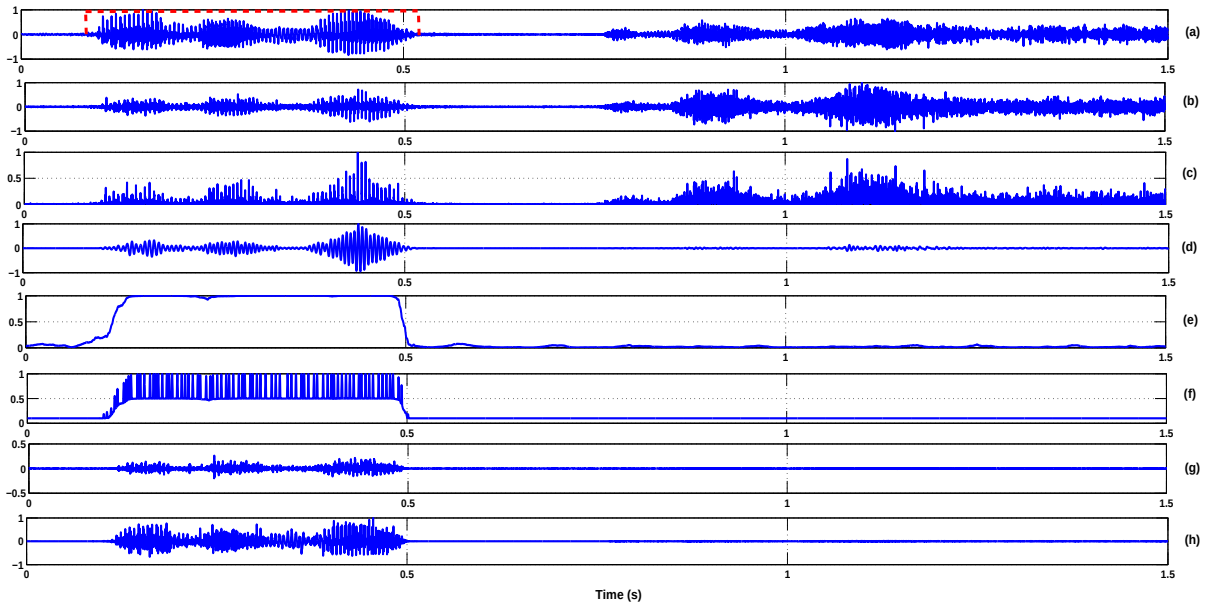

Figure 6.6: Illustration of excitation based enhancement. (a) speech signal recorded in foreground scenario (dotted lines and arrows indicate the foreground region), (b) LP residual signal derived from speech signal, (c) Hilbert envelope of LP residual signal, (d) ZBFS derived from passing HELP through ZBF, (e) gross weight function $w_g(n)$, (f) the positive zero crossings of ZBFS indicate the epoch locations, the temporal weight function $w(n)$ by combining the evidence of epoch locations with foreground segmentation, (g) modified LP residual signal $r_w(n)$, and (h) excitation based enhanced speech signal synthesized $s_t(n)$.

and (d) shows a spoken query recording and its corresponding spectrogram, respectively. It can be observed that there is a significant amount of background noise present in the recording. Further, Figs. 6.7(b) and (e) illustrates excitation based enhanced signal and its corresponding spectrogram, respectively. It can be noticed that there is a significant reduction of background noise, however, still there is some residual noise left over post EBE. The EBE signal is used to extract MFCCs in order to build different acoustic models.

6.4.2 Formant based Enhancement

The Sections 6.3 and 6.4.1 helped to temporally suppress background noise and enhance foreground speech signal. However, the temporal enhancement alone may not be sufficient to eliminate background noise in the spectral domain. The epoch locations are high SNR regions in the temporal domain, similarly formant peaks form high SNR regions in the spectral domain and they are preserved in foreground scenario. Further enhancement of formant peaks relative

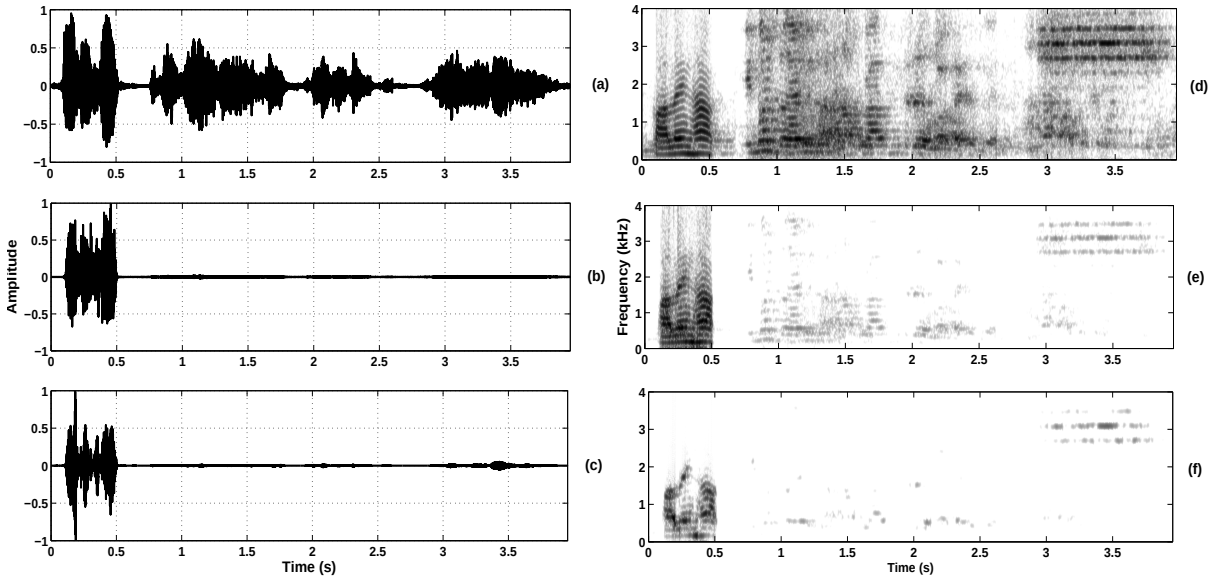

Figure 6.7: Illustration of enhancement outputs obtained at different stages and their narrowband spectrogram plots. (a) spoken query recorded in foreground scenario, (b) excitation based enhanced output, (c) formant based enhanced output, ((d) - (f)) are corresponding narrowband spectrograms.

to adjacent valleys should relatively enhance speech signal in the spectral domain. One of the reliable ways to estimate formant peaks is through linear prediction analysis as explained in Section 5.2.2. The formant enhancement is shown in the form of the block diagram in Fig. 6.5. This is illustrated in Fig. 6.7, where Figs. 6.7(c) and (f) shows formant enhanced foreground speech signal and its corresponding spectrogram. It can be noticed from the spectrogram that formants are enhanced relative to spectral valleys and further there is attenuation of background noise. The exaggeration of formant peaks should help in better discrimination of different acoustic-phonetic units and thereby help improve the recognition accuracy of ASR system.

6.5 Experimental Results and Discussions

In order to build the spoken query system, the speech data is collected from 1160 different farmers in the natural environment. The natural environment includes several other interfering sources along with desired foreground speech. Consequently, the degradation of the

speech signal by interfering sources will adversely affect the performance of spoken query system. However, the desired speaker's mouth is closer to mobile microphone sensor compared to other acoustic sources. Hence, the foreground speech is temporally segmented from rest of the background noise and further enhanced utilizing the unique signal characteristics of foreground speech. The Sections 6.3, 6.4.1, and 6.4.2 discussed different stages involved in foreground segmentation and enhancement. It is of interest to study the impact of foreground segmentation, excitation based enhancement, and formant based enhancement on the performance of spoken query system. In this work the effectiveness of foreground speech segmentation and enhancement is studied in the context of the commodity, as WER reported in case commodity is higher compared to District names [143]. Since both training and testing speech files are recorded in the natural environment both sets are subjected to foreground segmentation and enhancement. However, it is shown recently in several studies that SGMM-HMM and DNN-HMM-based models are superior to GMM-HMM-based acoustic modeling. Hence, it is of relevance to study the impact of foreground segmentation and enhancement in aforementioned acoustic modeling techniques. The details of different acoustic models used to realize ASR system is discussed in brief.

6.5.1 Acoustic Modeling of Spoken Query System

The interactive voice response system (IVRS) proposed in [141] responds to spoken query over a mobile phone through a telephone channel. The spoken query is limited to isolated word of 108 different commodity names. Though the spoken query includes district names, in this work only commodity names are considered for experimentation. The performance of IVRS depends on the ability to recognize commodity name spoken by the user. Since the spoken query is recorded in the natural environment, due to the presence of other interfering sources the performance of ASR suffers. Therefore, it is necessary to have a front end pre-processing module that can robustly segment the relevant foreground speech from rest of the background noise and enhance it to improve the performance of ASR. In order to build acoustic models for ASR, the MFCCs are extracted from the speech signal. The enhanced outputs $s_w(n)$, $s_t(n)$,

and $s_f(n)$ obtained from 3 different stages of the proposed work as explained in Sections 6.3, 6.4.1, and 6.4.2, respectively, are used to extract MFCC feature vectors.

The MFCC feature is computed using a Hamming window of length 20 ms with a frame rate of 100 Hz and a pre-emphasis factor of 0.97 is employed for pre-processing. The 13 dimensional MFCC base features ($C_0 - C_{12}$) are computed by employing 21 channel filter bank. The 13-dimensional base MFCC features are spliced over 4 frames to the left and to the right. The dimensionality of resulting 117-dimensional feature is then reduced to 40 using linear discriminant analysis (LDA). Further decorrelation is performed using maximum likelihood linear transform (MLLT). This is followed by speaker normalization through feature-space maximum-likelihood regression (fMLLR). The time-spliced LDA+MLLT+fMLLR transformed features (40-dimensional) are used as the input for system training.

For the system development and evaluation, the Assamese speech corpus [141] is divided into non-overlapping training and testing sets balanced in terms of the dialect groups. The training set consists of 30 hours of speech data collected from 885 native Assamese speakers (669 males and 216 females). The training set consists of more than 33,000 isolated word utterances constituting the commodity names, district names, and confirmation (yes/ no). The performance of the system is evaluated on a test set comprising of 2552 isolated word utterances from 275 speakers. It can be noted that the speakers in training and test sets are non-overlapping. Different acoustic modeling techniques are used in the current work in order to study the effect of foreground speech segmentation and enhancement on spoken query system. All these models are build using Kaldi toolkit [145] in this work. The details of GMM-HMM, SGMM-HMM, and DNN-HMM modeling techniques are briefly described in the following Sections.

6.5.1.1 GMM-HMM

The crossword triphone acoustic model training along with decision tree based state tying is employed. A 3 state HMM with 16 diagonal covariance Gaussian components per state is used for modeling each triphone. Silence and a short pause are modeled using a 3 state HMM with

32 Gaussian components per state. Decoding is performed using an equally likely wordnet and a dictionary of 256 words (including alternate pronunciations). A wordnet of 108 commodity names is employed for decoding the test set. The word error rate (WER) metric is used as a measure of recognition performance.

6.5.1.2 SGMM-HMM

In this section, we provide a brief summary of the parametrization of the SGMM for ASR. The SGMM based acoustic modeling is also realized using the Kaldi toolkit on the same training data. The number of Gaussians used for training the universal background model (UBM) is selected as 400. A number of leaves and Gaussians in the SGMM is selected to be 9000 and 7000, respectively. In our experimental setup, we have chosen the subspace dimension (S) as well as the feature dimension (D) to be equal, i.e., $S = D = 40$. The time-spliced LDA+MLLT+fMLLR transformed features are used for SGMM-HMM system training as well.

6.5.1.3 DNN-HMM

The DNN based acoustic modeling is implemented using the Kaldi toolkit. The input features are further spliced in time considering a context size of 9. This implies that the input is spliced over 4 frames to the left and right of the central frame, or 9 frames in total. The number of hidden layers is varied from 2 to 5 and fixed at 5 finally. An initial learning rate of 0.015 is selected which is reduced to 0.002 in 20 epochs. It should be noted that these epochs are different from instants of significant excitation. Here, epochs are the number of passes through the training data in DNN context. Extra 10 epochs are employed after reducing the learning rate to 0.002. Kaldi employs a preconditioned form of stochastic gradient descent (SGD). In this approach, instead of using a scalar learning rate, a matrix-valued the learning rate is employed. This is motivated by the basic idea to reduce the learning rate in dimensions where the derivatives have a high variance. This approach, in turn, is to control instability and stop the parameters moving too fast in any one direction. The minibatch size for neural net training was selected as 512. The total number of parameters trained happens to be 1.5 million. The same wordnets and dictionary are employed in the decoding using the SGMM and

Table 6.1: Performance evaluation of ASR in terms of WER (in %) using GMM-HMM, SGMM-HMM, and DNN-HMM acoustic modeling techniques, where, ORG - original recordings, FGS - foreground segmentation, EBE - excitation based enhancement, FBE - formant based enhancement, SS - spectral subtraction, and MMSE - minimum mean square error approximation.

Modeling Technique	ORG	FGS	EBE	FBE	SS	MMSE
GMM-HMM	16.00	14.50	15.52	14.73	22.73	22.53
SGMM-HMM	13.36	10.85	12.07	11.52	31.27	26.53
DNN-HMM	14.69	11.95	14.15	12.23	31.15	27.82

the DNN based systems.

6.5.2 Performance Evaluation

The user comfort level which essentially measures the ease of using IVRS by a user directly depends on ASR performance. It is shown in our earlier work that the improvement in ASR performance directly impacts the user comfort level positively [143]. Hence, performance evaluation of ASR is necessary to measure the effectiveness of IVRS. The proposed method is compared with 2 other methods chosen from the literature. The spectral domain subtraction based method is one of the earliest methods for enhancing speech signal distorted by additive noise. In spectral subtraction based methods, the noise components are modeled by average magnitude spectrum using several frames of noise-only regions. The noise spectrum is subtracted from signal spectrum to obtain the enhanced output [2]. However, the method proposed in [4] assumes that the Fourier coefficients of speech and noise can be independently modeled as zero mean Gaussian random variables. The method aims to minimize the mean square error between the clean speech and enhanced speech signal and hence it is called as minimum mean square error (MMSE). There are several modifications suggested to these two basic methods in literature to improve the quality of enhancement by reducing the musical noise. The Matlab implementations available in VOICEBOX [135] are used to compare with the proposed method in this paper.

The performance of recognizing right commodity names are tabulated in terms of word error rate (WER). Table 6.1 shows the results obtained from ASR performance evaluation. It can be noticed that the WER for the original signal is substantially higher and this can be attributed to the fact that there is a significant amount of background noise along with foreground speech. The foreground speech segmentation is effective in attenuating the background noise and this is reflected by the improved recognition performance compared to baseline. The SGMM-HMM-based system is the best performer amongst other modeling techniques. The SGMM-HMM-based system performs well under limited training data conditions [149]. It can be observed that the FGS, EBE, and FBE results are comparable for all 3 different acoustic modeling techniques. However, there is a significant decrease in the performance of spectral subtraction and MMSE based enhancement methods. It can be noticed that the performance of ASR is lower than baseline after the enhancement using these two methods. The robustness of foreground speech enhancement can be further tested by adding different types of noise at varying levels to test speech files.

The test speech files are added with white and babble noise at 20 and 10 dB levels in order to test the robustness of proposed method. Table 6.2 shows the performance evaluation of ASR using different acoustic modeling techniques. The test speech files are added with white and babble noise at 20 and 10 dB and enhanced using different enhancement methods. The noise files are taken from NOISEX [110] database and passed through telephone channel [150] in order to limit the bandwidth of noise signal before adding them to test speech signal. It can be noticed that there is a significant reduction in baseline performance due to degradation. However, it can be observed that there is a big improvement in the ASR performance post foreground segmentation and enhancement. It is observed there is a notable improvement in the performance of all 3 acoustic modeling techniques. The performance of SGMM-HMM system is the best amongst all other modeling techniques and this is consistent with the earlier study.

It can be noticed that in the case of EBE there is no significant improvement in results compared to FGS. Though there is a reduction of noise after EBE, however, excitation source

Table 6.2: Performance evaluation of ASR by adding white and babble noise at 20 and 10 dB levels in terms of WER (in %) using GMM-HMM, SGMM-HMM, and DNN-HMM acoustic modeling techniques, where, WN - white noise, BN - babble noise, ORG - original recording, FGS - foreground segmentation, EBE - excitation basegd enhancement, FBE - formant based enhancement, SS - spectral subtraction, and MMSE - minimum mean square error approximation.

Noise	Decibels	Modeling Technique	ORG	FGS	EBE	FBE	SS	MMSE
WN	20	GMM-HMM	39.30	21.90	21.24	**18.30**	38.05	35.07
		SGMM-HMM	43.30	15.99	15.52	**14.11**	39.58	38.71
		DNN-HMM	37.26	17.01	17.16	**15.48**	43.50	40.75
	10	GMM-HMM	67.59	47.37	48.51	**35.97**	63.21	35.07
		SGMM-HMM	61.40	29.86	30.09	**23.86**	67.01	38.71
		DNN-HMM	58.27	29.90	31.62	**26.61**	72.30	40.75
BN	20	GMM-HMM	46.39	18.94	19.95	**17.55**	51.96	47.69
		SGMM-HMM	45.65	12.75	12.97	**12.70**	52.19	54.15
		DNN-HMM	46.08	15.25	14.34	**13.32**	50.12	51.33
	10	GMM-HMM	75.35	38.99	43.53	**34.76**	81.50	75.63
		SGMM-HMM	63.05	23.59	22.92	**20.49**	72.57	79.98
		DNN-HMM	62.54	28.17	26.18	**24.10**	71.59	79.31

of foreground speech offers less discriminatory information for ASR. On the contrary, it can be observed that there is a significant improvement in the performance after formant enhancement. The reason for such improvement using formant enhancement can be observed from Figure 5.5, where, formant peaks are enhanced with increased dynamic range and this helps to improve the discriminatory information in the enhanced foreground speech. This enhancement aids all 3 acoustic modeling techniques especially when there is a significant amount of background noise is present. The performance of spectral subtraction and MMSE based enhancement methods are inferior compared to the baseline system. One of the reasons for such poor performance by these enhancement methods can be ascribed to the fact that over subtraction leads to distortion in the enhanced speech signal. However, foreground segmentation, EBE, and FBE introduces the least distortion to the enhanced speech signal and therefore helps to improve the ASR performance.

6.6 Summary

In this chapter, the effectiveness of using foreground speech segmentation and enhancement as front-end pre-processor is studied in the context of a practical spoken query system. The foreground speech segmentation and enhancement helps to attenuate background noise with least distortion compared to other state-of-the-art methods. The ASR performance is evaluated using proposed work and compared with 2 other state-of-the-art methods and it is shown that the proposed work helps to significantly improve the ASR accuracy. The idea is to enhance the known production features rather than modeling background noise to enhance foreground speech signal. It is shown that such enhancement methods are suitable to help improve ASR performance significantly under degraded conditions. Hence, such enhancement methods can be used as front-end processing module in real time deployment of ASR typically in a spoken query system. This further helps to improve user comfort level in handling spoken queries in adverse environments. Though this work demonstrates the usefulness of foreground speech segmentation and enhancement in the context of spoken query system, the same scheme can be extended for other speech-based applications as well.

7

Conclusions

Contents

The main aim of the work was to demonstrate the significance of closeness of speaker to the sensor to perform speech segmentation and enhancement. The speech collected from a speaker close to the sensor is termed as foreground speech. The signal due to rest of acoustic sources is termed as background noise. The goal set was to develop methods for segmentation and enhancement. The segmentation deals with temporal separation of foreground and background. The enhancement focuses on enhancing required speech components. Both the tasks were viewed by exploiting the knowledge of excitation source, vocal tract and suprasegmental information.

7.1 Conclusions from the Work

A method for epoch extraction based on the stable realization of zero frequency filter called zero band filter is proposed. The motivation is drawn from the drawback of marginally stable realization of zero frequency filter. Due to which the output of the filter is exponentially growing or decaying function of time. In order to extract epochs from such a signal, a trend removal function that relies on 1 - 2 pitch periods is needed, hence F_0 estimation is necessary. The output of stable zero band filter is not exponentially growing or decaying type and hence there is no necessity of F_0 estimation. The proposed work also has no issues with lengthy speech files. The performance is compared with the majority of existing state-of-the-art methods. It is found that the performance of zero band filter is comparable to best epoch extraction methods and robust to additive noise conditions. The method is evaluated in foreground scenario. Further in case of varying pitch conditions, the performance is comparable to zero frequency filter.

The close proximity of speaker to microphone and Lombard effect of desired speaker's speech offers some unique characteristics of signal compared to other background interfering sources which are far away from microphone. The desired speaker's speech spoken closer to the microphone is termed as *foreground speech* and rest of the interfering sources are categorized as *background noise*. Rather than analyzing the speech signal using all frequency components, it is easier to analyze the foreground speech using single frequency component closer to 0 Hz. The zero band filter output is used for signal analysis and establishes the definition for foreground

speech signal.

Not all regions of foreground speech signal are prone to interfering background noise. Especially, the regions around instants of significant excitation are high SNR regions. Due to such high SNR regions, the zero band filtering output has unique characteristics for foreground speech compared to other temporal regions. Hence, certain temporal features are derived from zero band filtered signal that offers discriminatory information between foreground speech and rest of the background regions. The features are combined with vocal tract articulatory gesture-based features to distinguish foreground speech regions from rest of the background and hence called as foreground speech segmentation.

The interfering background noise can be from several acoustic sources and may not be feasible to model all types of background sources. It is difficult to handle unseen interfering background source. A new approach to speech enhancement that utilizes close proximity of foreground speaker to the microphone and emulates Lombard speech effect is proposed. The high SNR regions of foreground speech *viz.,* instants of significant excitation, predominantly glottal closure instants are further boosted in excitation based enhancement. Similarly, the formant peak locations which form high SNR regions in spectral domain are boosted further in formant based enhancement using linear prediction coefficients.

Due to linear prediction filter, poles move closer to the unit circle and the formant based enhancement introduces some unnaturalness to the enhanced signal. To overcome this problem and further attenuate remaining residual noise, perceptual based enhancement using Mel cepstral coefficients and synthesis using Mel log spectrum approximation filter is used. The performance of all 3 different stages of foreground speech in terms of speech quality is evaluated using subjective and objective measures along with other considered state of art methods. Since there is no reference clean speech signal for an objective measure in the naturally recorded speech signal, two new objective measures are proposed in this work.

A practical mobile based spoken query system is prone to interfering background noise, due to which the performance of ASR accuracy reduces. Typically when speaking over the mobile phone, the speaker's mouth is closer to microphone relative to other interfering sources and

hence can be considered as foreground speech. However, the epochs are extracted using Hilbert envelope of linear prediction residual passed through zero band filtering, since the speech signal is recorded through a telephone channel. The ASR accuracy is studied in terms of word error rate for naturally collected spoken queries and later for additive noise cases.

7.2 Major Contributions of the Work

The important contribution of this thesis is to see the problem of single-channel speech enhancement in a new perspective. In this work, the closer distance between the desired speaker compared to other background sources that may include background speakers is utilized for speech segmentation and enhancement. The concepts of Lombard speech effect in the natural environment is used for speech segmentation and enhancement. The major contributions of the thesis is in developing methods for the following:

- Extraction of epoch locations without estimating F_0.
- Definition of *foreground speech* based on experimental analysis.
- Foreground speech segmentation using excitation and vocal tract information based features.
- Using glottal closure instant locations estimated by zero band filtering as anchor points to temporally enhance foreground speech.
- Enhancement of formant peaks by preserving spectral tilt is proposed for foreground speech enhancement.
- Perceptual enhancement using mel cepstral coefficients and mel log spectrum approximation filter.
- Two new objective measures for enhancement when speech signal is recorded in natural environments.

7.3 Scope for Future Work

- The foreground speech analysis reveals that there is a reduction in identification rate of epochs as the distance between microphone and speaker increases. Hence, there is a necessity for a new epoch extraction method that is robust to the distant speech signal.

- Estimation of glottal opening instants can help us in closed phase analysis and this can lead to new excitation based enhancement approach. Also, instantaneous fundamental periods for representing glottis pulse signal can be another production based feature, tried for robust foreground speech segmentation and enhancement.

- Perception based features as strength, power, harmonicity, spectral tilt, loudness, pitch, and timbre can be used for foreground speech enhancement purpose.

- The definition of *foreground speech* is established through analysis. However, the channel of communication between mouth and microphone can be theoretically modeled and such a modeling should help to lead the way to new areas of research. For example, in a single microphone recording scenario, it will be beneficial to find the distance between speaker and microphone using such a model. Next possibility is an inversion of such model can help to compensate for distant speech analysis using single microphone recording.

- The benefits of foreground speech segmentation and enhancement can be tried in other speech-based applications like speaker recognition and continuous speech recognition in natural recording scenarios.

- The natural extension of foreground speech segmentation and enhancement studies can be applied in multi-microphone recordings in a meeting room to segregate each speaker's speech from rest of other speakers voice at the background.

- The current set of studies is based on the assumption that the distance between speaker to microphone distance is limited from 1 to 2 in. However, methods can be explored to extend the range of foreground speech.

- The current study focused on utilizing foreground speech segmentation to enhance foreground speech regions. However, if the interfering background source is from another speaker then methods can be explored to enhance background speech content.

- The studies conducted in this work were limited to recordings done through a close-talking microphone, which is typically the case of headphone and mobile. Future studies can focus to extend the studies using the recordings done through the omnidirectional microphone.

A

Mel-Cepstral Coefficients

Contents

A.1 Mel-Cepstral Co-efficients

Linear prediction is one of the commonly used methods for obtaining the all-pole representation of speech signal. However, not all acoustic units can be efficiently modeled using all-pole model. The spectral zeros are also equally important to represent the speech signal. Alternatively, cepstral modeling is another important means for modeling the speech signal, this gives equal importance for both pole and zero representation of speech signal. One of the drawbacks of using cepstral representation for speech signal is that this overestimates the formant bandwidth. This can be be overcome by generalized cepstral analysis technique commonly called as mel generalized cepstral co-efficients or mel-cepstral coefficients (MCCs). The MCC analysis of speech signal can be viewed as a unified approach to the cepstral and linear prediction based methods. The spectrum estimation varies continuously from cepstral based approach to linear prediction based method [132].

In order to derive the mel-cepstral co-efficients an adaptive algorithm is used. The adaptive algorithm estimates the MCCs similar to least mean square (LMS) algorithm. The adaptive analysis system is implemented with an infinite impulse response (IIR) adaptive filter and such a method ensures the stability of the estimated filter [151]. The spectrum is represented using M^{th} order MCCs $\tilde{c}(m)$ using the following relationship

$$H(z) = exp \sum_{m=0}^{M} \tilde{c}(m)\tilde{z}^{-m} \tag{A.1}$$

where

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, |\alpha| < 1 \tag{A.2}$$

The phase characteristic of the all-pass transfer function $\tilde{z}^{-1} = e^{-j\tilde{\omega}}$ is given by

$$\tilde{\omega} = tan^{-1} \frac{(1 - \alpha^2)sin\omega}{(1 + \alpha^2)cos\omega - 2\alpha} \tag{A.3}$$

The α varies according to the sampling frequency of the speech signal, for example if the sampling frequency is 10 kHz, $\tilde{\omega}$ is a good approximation to the mel scale based on subjective pitch evaluations when $\alpha = 0.35$ [132]. In order to estimate the spectrum the error is minimized

with respect to $\{\tilde{c}(m)\}_{m=0}^{M}$

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi}\{expR(\omega) - R(\omega) - 1\}\, d\omega \tag{A.4}$$

$$R(\omega) = logI_N(\omega) - log|H(e^{j\omega}|^2 \tag{A.5}$$

where $I_N(\omega)$ is the modified periodogram of a weak sense stationary process $x(n)$ with a time window of length N. The Eqn. (A.1) can be re-written as

$$H(z) = exp\sum_{m=0}^{M} b(m)\Phi_m(z) = K.D(z) \tag{A.6}$$

where

$$K = exp\ b(0) \tag{A.7}$$

where

$$D(z) = exp\sum_{m=1}^{M} b(m)\Phi_m(z) \tag{A.8}$$

and

$$\tilde{c}(m) = \begin{cases} b(m), & \text{if } m = M \\ b(m) + \alpha b(m+1), & \text{otherwise} \end{cases} \tag{A.9}$$

$$\tilde{\Phi}_m(z) = \begin{cases} 1, & \text{if } m = 0 \\ \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}\tilde{z}^{-(m-1)}, & \text{otherwise} \end{cases} \tag{A.10}$$

Since $H(z)$ is a minimum phase system, we can show that the minimization of E with respect to $\{\tilde{c}(m)\}_{m=0}^{M}$ is equivalent to

$$\epsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{I_N(\omega)}{|D(e^{j\omega}|^2}d\omega \tag{A.11}$$

with respect to

$$b = [b(1), b(2), ..., b(m)]^T \tag{A.12}$$

The gain factor K that minimizes E is obtained by setting $\frac{\partial E}{\partial K} = 0$.

The MCCs computed are used to re-synthesize the speech signal. In order to estimate the spectrum, the efficient mel-log spectrum approximation filter (MLSA) is used for re-synthesis.

The transfer function $1/D(z)$ is not a rational function, the MLSA filter approximates $1/D(z)$ with sufficient accuracy [152].

Bibliography

[1] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 15, no. 1, pp. 34–43, January 2007.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, April 1979.

[3] B. Chen and P. Loizou, "Speech enhancement using a MMSE short time spectral amplitude estimator with Laplacian speech modeling," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 1097–1100.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[5] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, April 1985.

[6] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Orlando, USA, May 2002.

[7] B. Yegnanarayana and R. Smits, "A robust method for determining instants of major excitations in voiced speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 1995.

[8] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, pp. 25–42, May 1999.

[9] F. S. Cooper, "Acoustics in human communication: Evolving ideas about the nature of speech," *Journal of the Acoustical Society of America*, vol. 68, pp. 18–21, 1980.

[10] P. Sathyanarayana, "Short segment analysis of speech for enhancement," Ph.D. dissertation, Indian Institute of Technology Madras, 1999.

[11] S. R. M. Prasanna, "Event-based analysis of speech," Ph.D. dissertation, Indian Institute of Technology Madras, March 2004.

[12] P. Krishnamoorthy, "Combined temporal and spectral processing methods for speech enhancement," Ph.D. dissertation, Department of EEE, IIT Guwahati, 2009.

[13] H. Meinedo and I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features," in *Interspeech*, 2010.

[14] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Tran. Audio, Speech and Signal Processing*, vol. 14, pp. 1557–1565, 2006.

[15] P. C. Khoa, "Noise robust voice activity detection," Master's thesis, Nanyang Technological University, 2012.

[16] P. S. H. Krishnan, R. Padmanabhan, and H. A. Murthy, "Voice activity detection using group delay processing on buffered short-term energy," in *13th National Conf. Commun*, 2007.

[17] G. Evangelopoulos and P. Maragos, "Speech event detection using multiband modulation energy," in *Interspeech*, 2005.

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

[19] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Tran. Speech and Audio Processing*, vol. 9(5), pp. 504–512, 2001.

[20] J. Ramirez, J. C. Segura, C. Bentez, A. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, 2004.

[21] A. Benyassine, E. Shlomot, and H. Y. Su, "ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application," *IEEE Commun. Mag*, vol. 35, pp. 64–73, 1997.

[22] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 798–807, 2010.

[23] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Tran. Acoustics, Speech, and Signal Processing*, vol. 24, pp. 201–212, 1976.

[24] M. Farsinejad and M. Analoui, "A new robust voice activity detection method based on genetic algorithm," in *Telecommunication Networks and Applications Conference, ATNAC*, 2008.

[25] J. Ramirez, P. Yelamos, J. Gorriz, and J. Segura, "Svm-based speech endpoint detection using contextual speech features," *Electronics letters*, vol. 42, pp. 426–428, 2006.

[26] J. Ramirez, P. Yelamos, J. Gorriz, J. Segura, and L. Garca, "Speech/non-speech discrimination combining advanced feature extraction and svm learning," in *Ninth International Conference on Spoken Language Processing*, 2006.

[27] J. Wu and X. Zhang, "Maximum margin clustering based statistical vad with multiple observation compound feature," in *IEEE Signal Processing Letters*, 2011.

[28] F. Wang, B. Zhao, and C. Zhang, "Linear time maximum margin clustering," *IEEE Trans. Neural Networks*, vol. 21, pp. 319–332, 2010.

[29] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 23, pp. 705–717, 2015.

[30] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Lett.*, vol. 16, no. 6, pp. 469–472, June 2009.

[31] N. Adiga and S. R. M. Prasanna, "Detection of glottal activity using different attributes of source information," *IEEE Signal Process. Letters*, vol. 22, pp. 2107–2111, 2015.

[32] K. T. Deepak, B. D. Sarma, and S. R. M. Prasanna, "Foreground speech segmentation using zero frequency filtered signal," in *Interspeech*, September 2012.

[33] K. T. Deepak and S. R. M. Prasanna, "Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients," *IEEE Trans. Acoust., Speech and Signal Process.*, p. [Accepted for Publication], 2016.

[34] M. A, V. D. S. J, M. P, and T. P, "A spectral conversion approach to single-channel speech enhancement," *IEEE Tran. Audio, Speech, and Language Processing*, vol. 15, pp. 1180–1193, 2007.

[35] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Tran. Acoustics, Speech, and Signal Processing*, vol. 28, pp. 137–145, 1980.

[36] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, pp. 1586–1604, 1979.

[37] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 7, pp. 126–137, 1999.

[38] T. S. Gunawan, E. Ambikairajah, and J. Epps, "Perceptual speech enhancement exploiting temporal masking properties of human auditory system," *Speech Commun.*, vol. 52, pp. 381–393, 2010.

[39] ——, "Perceptual speech enhancement exploiting temporal masking properties of human auditory system," *Speech Commun.*, vol. 52, pp. 381–393, 2010.

[40] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[41] P. Krishnamoorthy and S. R. M. prasanna, "Application of combined temporal and spectral processing methods for speaker recognition under noisy, reverberant or multi-speaker environments," *Sadhana*, vol. 34, no. 5, pp. 729–754, October 2009.

[42] P. Krishnamoorthy and S. R. M. Prasanna, "Reverberant speech enhancement by temporal and spectral processing," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 17, no. 2, pp. 253–266, February 2009.

[43] ——, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication*, vol. 53, pp. 154–174, February 2011.

[44] G. Pradhan and S. R. M. Prasanna, "Speaker verification under degraded condition: a perceptual study," *Int. Journal of Speech Technology (Springer)*, vol. 14, no. 4, pp. 405–417, Oct. 2011.

[45] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded condition," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 19, no. 8, pp. 2552–2565, May 2011.

[46] D. Freeman, G. Cosier, C. Southcott, and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," in *International Conference on Acoustics, Speech, and Signal Processing*, 1989.

[47] F. Beritelli, S. Casale, and A. Cavallaero, "A robust voice activity detector for wireless communications using soft computing," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1818–1829, 1998.

[48] A. Sangwan, M. Chiranth, H. Jamadagni, R. Sah, and R. V. Prasad, "Vad techniques for real-time speech transmission on the internet," in *5th IEEE International Conference on High Speed Networks and Multimedia Communications*, 2002.

[49] K. Itoh and M. Mizushima, "Environmental noise reduction based on speech/nonspeech identification for hearing aids," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[50] *ETSI, "Universal mobile telecommunication systems mandatory speech codec speech Processing fFunction, AMR SPEECH Codec; voice activity detector (3GPP TS 26.094 version 4.0.0 release 4)," 2001.*

[51] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell system Tech. Journal*, vol. 54, pp. 297–315, February 1975.

[52] L. R. Rabiner and R. W. Scafer, *An Introduction to Digital Speech Processing (Foundations and Trends in Signal Processing).* Now Publishers Inc, 2007.

[53] A. M. Liberman, *Speech: A Special Code*, M. Press, Ed. MA Cambridge, 1996.

[54] D. Poeppel, "The analysis of speech in different temporal integration windows: cerebral lateralization as asymmetric sampling in time," *Speech Commun.*, vol. 41, pp. 245–255, 2003.

[55] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice acitvity detection," *IEEE Journal On Selected Topics in Signal Processing*, vol. 54, pp. 297–315, 1975.

[56] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 600–613, 2011.

[57] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 352–333, September 1995.

[58] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *IEEE Signal Process. Letters*, vol. 14, pp. 762–765, October 2007.

[59] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 7, pp. 609–618, November 1999.

[60] M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 14, no. 2, pp. 456–466, March 2006.

[61] S. Ahmadi and A. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE Tran. on Speech and Audio Processing*, vol. 7, pp. 333–338, 1999.

[62] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Taylor and F. Group, Eds. CRC Press, 2013.

[63] T. Kunnunen, E. Cherneko, M. Tuononen, P. Frnti, and H. Li, "Voice activity detection using mfcc features and support vector machine," in *Int. Conf. on Speech and Computer*, 2007.

[64] T. Fukuda, O. Ichikawa, and M. Nishimura, "Phone-duration-dependent long-term dynamic features for a stochastic model-based voice activity detection," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[65] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in *Proc. IEEE ICASSP*, 2000.

[66] P. L. Cerf and D. V. Compornolle, "A new variable frame rate analysis method for speech recognition," *IEEE Signal Process. Letters*, vol. 1, pp. 185–187, 1994.

[67] K. M. Pointing and S. M. Peeling, "The use of varaible frame rate analysis in speech recognition," *Comput. Speech Lang.*, vol. 5, pp. 169–179, 1991.

[68] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 3, pp. 217–231, March 2001.

[69] K. Li, M. N. S. Swamy, and M. O. Ahmad, "An improved voice activity detection using highre order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 965–974, 2005.

[70] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 16, pp. 1602–1613, November 2008.

[71] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 21, pp. 2471–2480, 2013.

[72] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 30, pp. 679–681, 1982.

[73] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction.* John Wiley and Sons, 2006.

[74] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Washington, USA, April 1979, pp. 208–211.

[75] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 2, pp. 345–349, 1994.

[76] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[77] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2-3, pp. 215–228, June 1992.

[78] H. T. Hu, F. J. Kuo, and H. J. Wang, "Supplementary schemes to spectral subtraction for speech enhancement," *Speech Commun.*, vol. 36, pp. 205–218, 2002.

[79] M. Hasan, S. Salahuddin, and M. Khan, "A modified a priori snr for speech enhancement using spectral subtraction rules," *IEEE Signal Process. Letters*, vol. 11, pp. 270–273, 2004.

[80] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Trans. Audio Speech and Language Processing*, vol. 20, pp. 2080–2094, 2012.

[81] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 1998.

[82] Y. Shao and C. Chang, "A generalized timefrequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system," *IEEE Tran. on Systems, Man, and Cybernetic-Part B: Cybernetics*, vol. 37, pp. 877–889, 2007.

[83] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, April 1992.

[84] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Commun.*, vol. 24, pp. 249–257, 1998.

[85] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 109–118, 2002.

[86] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio Speech and Language Processing*, vol. 20, pp. 92–102, 2012.

[87] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Trans. Audio Speech and Language Processing*, vol. 23, pp. 7–19, 2015.

[88] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, *Progress in Nonlinear Speech Processing*. Springer Berlin Heidelberg, 2007.

[89] K. Hermus, P. Wambacq, and H. V. Hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 1, pp. 195–210, 2007.

[90] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: a regenerative approach," *Speech Commun.*, vol. 10, pp. 45–67, 1991.

[91] Y. Ephraim and H. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Tran. Speech and Audio Processing*, vol. 3, pp. 251–266, 1995.

[92] S. Jensen, P. Hansen, S. Hansen, and J. Sorensen, "Reduction of broad-band noise in speech by truncated qsvd," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 439–448, 1995.

[93] W. Jin and M. S. Scordilis, "Speech enhancement by residual domain constrained optimization," *Speech Commun.*, vol. 48, pp. 1349–1364, 2006.

[94] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *IEEE Conf. on Acoustics, Speech, and Signal Processing*, 2002.

[95] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 20, no. 3, pp. 994–1006, March 2012.

[96] P. Chetana, N. Dhananjaya, and S. V. Gangashetty, "Analysis of acoustic events in speech signals using bessel series expansion," *Springer Circuits SystSignal Process*, vol. 32, pp. 2915–2938, 2013.

[97] B. Yegnanarayana and S. V. Gangashetty, "Epoch-based analysis of speech signals," *Sadhana*, vol. 36, pp. 651–697, October 2011.

[98] B. Yegnanarayana, S. R. M. Prasanna, and S. Guruprasad, "Study of robusness of zero frequency resonator method for extraction of fundamental frequency," in *ICASSP*, 2011.

[99] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modelling of spectrum, pitch and duration in hmm-based speech synthesis," in *Eurospeech*, 1999.

[100] D. Govind, S. R. M. Prasanna, and B. Yegnanarayana, "Neutral to target emotion conversion using source and suprasegmental information," in *Interspeech*, August 2011.

[101] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 309–319, August 1979.

[102] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Am.*, vol. 56, pp. 1625–1629, November 1974.

[103] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Interspeech*, 2009.

[104] M. R. P. Thomas, J. Gudnanson, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 20, pp. 82–91, June 2012.

[105] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing.* Upper Saddle River, NJ: Prentice-Hall, 1999.

[106] K. S. S. Srinivas and K. Prahallad, "An FIR implementation of zero frequency filtering of speech signals," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 20, no. 9, pp. 2613–2617, November 2012.

[107] *"NIST-Speaker Recognition Evaluations." in.* [Online]. Available: http://www.nist.gov/itl/iad/mig/sre12.cfm

[108] *CMU-ARCTIC Speech Synthesis Databases.* [Online]. Available: http://festvox.org/cmu_arctic/index.html

[109] M. R. P. Thomas and P. A. Naylor, "The SIGMA algorithm: A glottal activity detector for electroglottographic signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 1557–1566, 2009.

[110] *Noisex-92*. [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[111] D. Govind and S. R. M. Prasanna, "Epoch extraction in emotional speech," in *SPCOM*, 2012.

[112] *German Emotional Speech Database*. [Online]. Available: http://database.syntheticspeech.de/

[113] S. Guruprasad and B. Yegnanarayana, "Performance of an event-based instantaneous fundamental frequency estimator for distant speech signals," *IEEE Tran. Audio, Speech and Signal Processing*, vol. 19, pp. 1853 –1864, 2011.

[114] D. S. Williamson, Y. Wang, and D. Wang, "Reconstruction techniques for improving the perceptual quality of binary masked speech," *J. Acoust. Soc. Am.*, pp. 892–902, 2014.

[115] S. Guruprasad, "Significance of processing regions of high signal-to-noise ratio in speech signals," Ph.D. dissertation, Ph.D. dissertation, Indian Institute of Technology Madras, Chennai, India, 2011.

[116] K. Sjlander and J. Beskow, "Wavesurfer an open source speech tool," in *Sixth International Conference on Spoken Language Processing*, 2000.

[117] K. T. Deepak and S. R. M. Prasanna, "Epoch extraction using zero band filtering from speech signal," *Circuits, Systems, and Signal Processing*, vol. 34, pp. 2309–2333, December 2014.

[118] C. L. Smith, C. P. Browman, R. S. McGowan, and B. Kaytt, "Extracting dynamic parameters from speech movement data," *Haskins Laboratories Status Report on Speech Research*, vol. SR-105/106, pp. 107–140, 1991.

[119] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *ICASSP*, vol. 3, April 1997, pp. 1647–1650.

[120] T. Irino and R. D. Patterson, "A compressive gammachirp auditory filter for both physiological and psychophysical data," *J. Acoust. Soc. Am*, vol. 109 (5), pp. 2008–2022, May 2001.

[121] M. Unoki, T. Irino, B. Glasberg, B. C. J. Moore, and R. D. Patterson, "Comparison of the roex and gammachirp filters as representations of the auditory filter," *J. Acout. Soc. Am.*, vol. 120(3), pp. 1474–1492, 2006.

[122] TIMIT, *"Timit Acoustic-Phonetic Continuous Speech Corpus"*, NIST Order PB91-505065, National Institute of Standards and Technology, Gaithersburg, MD, USA, 1990, Speech Disc 1-1.1., 1990.

[123] K. Parlak and O. G. Moreno, *Applied Speech Enhancement in Mobile Communication Acoustics: Background Noise Elimination with Filtering Algorithms.* LAP LAMBERT Academic Publishing, 2012.

[124] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 9, pp. 504–512, 2001.

[125] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 20, pp. 1383–1393, 2011.

[126] Y. Hu and P. C. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Trans. Audio, Speech and Lanuage Processing*, vol. 11, pp. 457–465, 2003.

[127] J. C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *Journal of the Acoustical Society of America*, vol. 93, pp. 510–524, 1993.

[128] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *Journal of the Acoustical Society of America*, vol. 84, pp. 917–928, 1988.

[129] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, October 1994.

[130] J. Yang, F. Luo, and A. Nehorai, "Spectral contrast enhancement: Algorithms and comparisons," *Speech Commun.*, vol. 39, pp. 33–46, 2002.

[131] V. Ramamoorthy, N. S. Jayant, R. V. Cox, and M. M. Sondhi, "Enhancement of adpcm speech coding with backward-adaptive algorithms for postfiltering and noise feedback," *IEEE Journal On Selected Areas in Communication*, vol. 6, pp. 364–382, 1988.

[132] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," in *ICSLP*, September 1994, pp. 1043–1046.

[133] G. Fant, *Speech sound and features*. MIT Press, Cambridge, 1973.

[134] A. W. Rix, J. G. Beerends, D. S. Kim, P. Kroon, and O. Ghitza, "Objective assesment of speech and audio quality - technology and applications," *IEEE Tran. Audio, Speech and Signal Processing*, vol. 14, pp. 1890–1901, 2006.

[135] *VOICEBOX: Speech Processing Toolbox for MATLAB*.

[136] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *ICASSP*, May 2006.

[137] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Tran. Audio, Speech and Signal Processing*, vol. 13, pp. 751–761, 2005.

[138] J. Glass, "Challenges for spoken dialogue systems," in *In Proceedings of IEEE ASRU Workshop*, 1999.

[139] S. L. Tomko, "Improving user interaction with spoken dialog systems via shaping," Ph.D. dissertation, Carnegie Mellon University, 2006.

[140] S. Shahnawazuddin, K. T. Deepak, B. D. Sarma, A. Deka, S. R. M. Prasanna, and R. Sinha, "Assamese spoken query system to access the price of agricultural commodities," in *National Conference on Communications*, February 2013.

[141] ——, "Low complexity on-line adaptation techniques in context of assamese spoken query system," *Journal of Signal Processing Systems*, vol. 81, pp. 83–97, 2015.

[142] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Tran. Audio, Speech and Signal Processing*, vol. 11, pp. 568–580, 2003.

[143] S. Shahnawazuddin, K. T. Deepak, S. Imami, S. R. M. Prasanna, and R. Sinha, "Improvements in iitg assamese spoken query system : Background noise suppression and alternate acoustic modeling," *Journal of Signal Processing Systems*, p. [Accepted for Publication], 2016.

[144] *http://htk.eng.cam.ac.uk/*.

[145] *Kaldi Toolkit: http://kaldi.sourceforge.net*.

[146] *http://agmarknet.nic.in/*.

[147] L. Cohen, *Time-Frequency Analysis: Theory and Applications*, S. P. Series, Ed. Ser. Signal Processing Series, Englewood Cliffs: Prentice-Hall, 1995.

[148] K. T. Deepak, K. Ramesh, N. Adiga, and S. R. M. Prasanna, "Speech and egg polarity detection using hilbert envelope," in *TENCON 2015 - 2015 IEEE Region 10 Conference*, 2015.

[149] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model-a structured model for speech recognition," *Computer Speech and Language*, vol. 25, no. 2, pp. 404–439, Apr 2011.

[150] S. King and V. Karaiskos, "The blizzard challenge 2009," in *Blizzard Challenge 2009*, 2009.

[151] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *ICASSP*, 1992.

[152] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (mlsa) filter for speech synthesis," *Trans. IECE*, vol. J66-A, pp. 122–129, 1983.

List of Publications

Journal Publications

- Published and Accepted:

  1. K. T. Deepak and S. R. M. Prasanna, "Epoch Extraction using Zero Band Filtering from Speech Signal," **Circuits, Systems & Signal Processing**, Springer, July 2015, Vol 34, pp. 2309-2333.
  2. S. Shahnawazuddin, K. T. Deepak, B. D. Sarma, A. Deka, S. R. M. Prasanna and Rohit Sinha, "Low Complexity On-Line Adaptation Techniques in Context of Assamese Spoken Query System," **Journal of Signal Processing Systems**, Springer, October 2015, Vol 81, pp. 83-97.
  3. K. T. Deepak and S. R. M. Prasanna, "Foreground Speech Segmentation and Enhancement using Glottal Closure Instants and Mel Cepstral Coefficients," **IEEE/ACM Transactions on Audio, Speech and Language Processing**, April 2016, Vol 24, pp. 1204-1218.
  4. S Shahnawazuddin, K. T. Deepak, Abhishek Dey, Siddika Imani, S R M Prasanna and Rohit Sinha, "Improvements in IITG Assamese Spoken Query System: Background Noise Suppression and Alternate Acoustic Modeling," **Journal of Signal Processing Systems**, [Available Online], Springer.

- Manuscript to be Communicated

  1. K. T. Deepak , S Shahnawazuddin, Abhishek Dey, Rohit Sinha and S R M Prasanna, "Robust Spoken Query System using Foreground Speech Segmentation and Enhancement,".
  2. K. T. Deepak and S. R. M. Prasanna, "Degraded Speech Processing: A review"

Conference and Workshop Publications

- Published:

1. K. T. Deepak, Biswajit Dev Sarma and S. R. M. Prasanna, "Foreground Speech Segmentation using Zero Frequency Filtered Signal," in *Interspeech, Portland, USA*, September 2012
2. S. Shahnawazuddin, K. T. Deepak, B. D. Sarma, A. Deka, S. R. M. Prasanna and Rohit Sinha, "Assamese Spoken Query System to Access the Price of Agricultural Commodities," in *National Conference on Communications, IIT Delhi*, February 2013.
3. K. T. Deepak, and S. R. M. Prasanna, "Analysis of Foreground and Distant Speech," in *TENCON, Macau, China*, November 2015.
4. S Shahnawazuddin, Abhishek Dey, K. T. Deepak, Siddika Imani, S R M Prasanna and R. Sinha, "Enhancements in Assamese Spoken Query System: Enabling Background Noise Suppression and Flexible Queries," in *National Conference on Communications, IIT Guwahati*, February 2016.

- Other Publications

1. K. T. Deepak and S. R. M. Prasanna, "Remote Spoken Document Retrieval using Foreground Speech Segmentation based Isolated Word Recognizer," in *Proc Indicon, IIT Bombay*, December 2013
2. K. T. Deepak, K. Ramesh and S. R. M. Prasanna, "Extraction of Glottal Closure and Opening Instants using Zero Frequency Filtering," in *Indicon, Pune*, 2014 (Obtained Best Oral/Poster Presentation Award).
3. K. T. Deepak and S. R. M. Prasanna, "Reference and Automatic Marking of Glottal Opening Instants Using EGG Signal," in *SPCOM, IISc Bangalore*, July 2014.

4. K. T. Deepak, K. Ramesh, Nagaraj Adiga and S. R. M. Prasanna, "Speech and EGG Polarity Detection using Hilbert Envelope," in *TENCON, Macau, China*, November 2015.

5. Ravi Shankar, Arpit Jain, K. T. Deepak, C. M. Vikram , A. Deka and S. R. M. Prasanna, "Spoken Term Detection from Continuous Speech Using ANN Posteriors and Image Processing Techniques," in *National Conference on Communications, IIT Guwahati*, February 2016.