Speech Enhancement in Modulation Domain using Codebook-based Speech and Noise Estimation

Vidhyasagar Mani

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

February 2016

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of M Eng in Electrical & Computer Engg.

2016/02/02

Abstract

Conventional single-channel speech enhancement methods implement the analysis-modification-synthesis (AMS) framework in the acoustic frequency domain. Recently, it has been shown that the extension of this framework to the modulation domain may result in better noise suppression. However, this conclusion has been reached by relying on a minimum statistics approach for the required noise power spectral density (PSD) estimation.

Various noise estimation algorithms have been proposed over the years in the speech and audio processing literature. Among these, the widely used minimum statistics approach is known to introduce a time frame lag in the estimated noise spectrum. This can lead to highly inaccurate PSD estimates when the noise behaviour rapidly changes with time, i.e., non-stationary noise. Speech enhancement methods which employ these inaccurate noise PSD estimates tend to perform poorly in the noise suppression task, and in worst cases, may end up deteriorating the noisy speech signal even further. Noise PSD estimation algorithms using *a priori* information about the noise statistics have been shown to track non-stationary noise better than the conventional algorithms which rely on the minimum statistics approach.

In this thesis, we perform noise suppression in the modulation domain with the noise and speech PSD derived from an estimation scheme which employs the *a priori* information of various speech and noise types. Specifically, codebooks of gain normalized linear prediction coefficients obtained from training on various speech and noise files are used as the *a priori* information while performing the estimation of the desired PSD. The PSD estimates derived from this codebook approach are used to obtain a minimum mean square error (MMSE) estimate of the clean speech modulation magnitude spectrum, which is then combined with the phase spectrum of the noisy speech to recover the enhanced speech signal. The enhanced speech signal is subjected to various objective experiments for evaluation. Results of these evaluations indicate improvement in noise suppression with the proposed codebook-based modulation domain approach over competing approaches, particularly in cases of non-stationary noise.

Sommaire

Les méthodes conventionnelles de rehaussement de la parole à canal unique utilisent une structure d'analyse-modification-synthèse (AMS) dans le domaine fréquentiel. Récemment, il a été démontré que l'utilisation de cette structure dans le domaine de la modulation pourrait offrir une meilleure suppression du bruit. Toutefois, cette conclusion a été obtenue en se basant sur une approche à statistique minimale pour l'estimation de la densité spectrale de puissance (PSD) du bruit ambiant.

Plusieurs algorithmes d'estimation de bruit ont été proposés au fil des ans dans la littérature sur le traitement de la parole. D'ordinaire, les méthodes d'estimation de bruit qui se servent d'une approche à statistique minimale vont créer un décalage temporel dans l'estimation spectrale du bruit. Ce décalage peut engendrer beaucoup d'imprécision dans l'estimation de la PSD en présence de bruit dont les caractéristiques changent rapidement dans le temps, c.-à-d., bruit non-stationnaire. Les méthodes de rehaussement de la parole qui utilisent ces estimateurs ont tendance à mal performer dans la tache de suppression de bruit, et dans les pires cas, peuvent même détériorer encore plus le signal déjà bruyant. À cet égard, les algorithmes d'estimation de PSD utilisant l'information statistique du bruit *a priori* conduisent à une meilleure performance dans l'estimation et le suivi des paramètres de bruit non-stationnaire que les méthodes conventionnelles qui reposent sur une approche à statistique minimale.

Dans ce mémoire, nous avons mis en œuvre un algorithme de suppression de bruit dans le domaine de la modulation en utilisant des PSD pour la parole et le bruit ambiant dont l'estimation repose sur l'utilisation d'un dictionnaire (codebook). Plus précisément, nous utilisons des dictionnaires de coefficients de prédiction linéaire normalisés, créés à partir de plusieurs enregistrements de parole et de bruits, afin d'estimer les PSD requises. Les PSD ainsi estimées à partir de ces dictionnaires sont utilisées pour obtenir une estimation de type "erreur quadratique moyenne minimale" du spectre d'amplitude du signal vocal dans le domaine de la modulation. Le spectre d'amplitude résultant est ensuite utilisé pour recouvrer le signal parole en utilisant le spectre de phase du bruit ambiant. Des signaux vocaux, rehaussés via l'algorithme proposé, ont été soumis a différentes évaluations démontrant une amélioration significative dans la suppression du bruit ambiant, particulièrement en présence de bruit non-stationnaire.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Prof. Benoit Champagne, for his continued support during my thesis work. With no doubt, this thesis would not have been possible without his guidance, advice and lots of helpful ideas. I am grateful for the financial support provided by Prof. Champagne via his research grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada, and Microsemi Canada Ltd without which this thesis would not have been possible. Special mention to Dean(Microsemi), Patrick(Microsemi) and Prof. Zhu (Concordia) for all their constructive suggestions during the progress meetings. To all the members at MMSP Laboratory I have been fortunate enough to get acquainted with over the last two years: Ali, Hanwook, Siamek, Reza, Golnaz, Jiaxin, Ryan, David, Hao and Su - thank you for making the lab a fun-filled place (Special shout out to David for helping me with the translation of my abstract into french). To all my friends and well-wishers in both Canada and India, thank you for all your support and comfort. In the end, I would like to thank my family for their steadfast belief and encouragement.

Contents

List of Figures

List of Tables

List of Acronyms

AMS	Analysis-Modification-Analysis
FFT	Fast Fourier Transform
LBG	Linde-Buzo-Gray algorithm
LP (coefficients)	Linear Predictive (coefficients)
MMSE	Minimum Mean Square Error
MME	Modulation domain based MMSE STSA Estimator
MS	Minimum Statistics
MSS	Modulation domain based Spectral Subtraction
NSS	Non-linear Spectral Subtraction
OLA	OverLap Add
PDF	Probability Density Function
PESQ	Perceptual Evaluation of Speech Quality
PSD	Power Spectral Density
SegSNR	Segmental Signal-to-Noise Ratio
SNR	Signal-to-Noise Ratio
SSA	Signal Subspace Approach
STFT	Short-Time Fourier Transform
STSA	Short-Time Spectral Amplitude Estimator
VAD	Voice Activity Detector
Wβ-SA (Estimator)	Weighted β Spectral Amplitude (Estimator)
WE STSA (Estimator)	Weighted Euclidean Short-Time Spectral Amplitude (Estimator)
FIR	Finite Impulse Response
IIR	Infinite Impulse Response
ITU-T	International Telecommunication Union

ESTI	European Telecommunication Standard Institute
AMR	Adaptive multi Rate (codec)
MCRA	Minima Controlled Recursive Averaging
IMCRA	Improved Minima Controlled Recursive Averaging
STP (parameter)	Short Term Predictor (parameter)
AR (coefficient)	Auto-Regressive (coefficient)
ML	Maximum Likelihood
LLF	Log Likelihood Function

Chapter 1

Introduction

This chapter serves as a general introduction to the thesis. It begins with a concise overview on the problem of noise reduction and quality enhancement for single channel speech signals. This is followed by a literature survey on more specific topics pertinent to the research conducted in this work, including modulation domain processing and codebook assisted speech enhancement. The main objective and technical contributions of the thesis are then summarized. Finally, the organization of the upcoming chapters is outlined.

1.1 Overview of Single Channel Speech Enhancement

Speech enhancement is defined as the improvement in the intelligibility and/or quality of a degraded speech signal through the application of signal processing tools and techniques [1]. It finds its use in many speech communication applications such as mobile phones, voice over IP (VoIP), speech recognition, hearing aids, teleconferencing systems and other areas where the perceptual quality of speech signal is paramount.

The performance of any speech enhancement method is arbitrated by two broad perceptual criteria, namely: quality and intelligibility. Intelligibility of a speech signal is a measure of how comprehensible the speech is, and it is determined by words, phrases and sentences that can be discerned or understood by a listener. Speech quality is determined by the amount of background noise, superimposed cross talk, etc., as well as the distortion present in the processed speech signal. It has been shown that these criteria are rarely satisfied simultaneously, especially when it comes to the amount of distortion and background noise present in a speech signal. An improvement in quality through noise reduction

typically comes at the expense of speech distortion or worse, loss of intelligibility. A good speech enhancement method strives to perform noise reduction while ensuring that the resulting speech distortion is within a tolerable limit.

Most of the single channel enhancement methods in use today work are limited by this trade-off. Increasing the degrees of freedom with multiple microphone channels can help push back this limit. By acquiring spatial information of a target speech source, multi-channel techniques such as delay-and-sum (DS) beamforming and minimum variance distortionless response (MVDR) beamforming [2, 3] can provide enhancement performance superior to single channel methods. In this thesis, however, we will be focusing on single channel enhancement methods, due to convenience in implementation as well as cost and size considerations.

Spectral subtraction is one of the earliest techniques used for single channel speech enhancement. Pioneering work on this topic was conducted by scientists at Bell Labs who succeeded in implementing a spectral subtraction algorithm in the analog domain [4]. Digital domain implementations of spectral subtraction were later developed by Boll [5], Berouti *et al.* [6] and Sondhi *et al.* [7]. In this approach, the noisy speech is transformed to the frequency domain, modified by subtracting from it an estimate of the noise spectrum, and then converted back to the time domain. More advanced and sophisticated refinements of the basic spectral subtraction method have been proposed over the years, e.g., [8]. Spectral subtraction is widely used for speech enhancement because of its simple implementation and low computational cost. However, this approach suffers from some resultant artefacts known as musical noise, which can be an annoying distraction for the listener in some cases.

To alleviate the problem of musical noise, Minimum Mean Square Error (MMSE) based spectral amplitude estimators are also widely used for speech enhancement. Ephraim and Malah were first to propose the MMSE Short-Time Spectral Amplitude (STSA) estimator [9]. This estimation method is developed by minimizing the expected value of a cost function which serves as a measure of the error between the estimated and clean speech spectra. A closed form solution for this optimization problem is obtained based on the assumptions that the speech and noise signals are additive in the time domain, and that their short-time spectral components can be modelled as zero-mean Gaussian random variables which are statistically independent and identically distributed.

A Log-MMSE based method was later proposed by Ephraim and Malah in [10]. The cost function for this method is the mean square error between the log spectra of the clean

and estimated speech, rather than the error between the spectra themselves. The motivation behind using log spectra is based on the properties of the human auditory system, which performs a logarithmic compression of the spectral amplitudes of incoming speech signals. Subsequent STSA methods with cost functions based on the internal mechanisms of the human auditory systems have been proposed over the years, including the Weighted-Euclidean STSA (WE STSA) estimator [11], the β-SA estimator [12] and Weightedβ-SA (Wβ-SA) estimator [13]. These techniques are generally free of musical noise and achieve better noise suppression than the basic MMSE STSA and Log-MMSE STSA methods.

Besides the spectral subtraction and MMSE-STSA estimation, other methods have been proposed as well. For instance, linear spectral estimators based on the Weiner filtering approach [14, 15] have been widely studied and also shown to be free of musical noise whilst performing reasonable noise suppression. Time-domain methods based on Kalman filtering have also received considerable attention over the years, as in [16–18]. Kalman filter based enhancement methods generally result in lesser speech distortion when compared to other methods, and the residual noise is also mostly free of annoying artefacts. However, Kalman filter based methods have limited noise suppression capability when compared to other methods.

Departing from the more traditional approaches, Ephraim and Van Trees proposed a speech enhancement method based on signal subspace decomposition in [19]. It aims to estimate the clean speech signal after decomposing the noisy speech signal into so called speech signal and noise subspaces. More sophisticated extensions of the subspace approach have been proposed over the subsequent years [20–23].

Nowadays, the most commonly used speech enhancement methods, such as spectral subtraction, MMSE-STSA estimation and Wiener filtering, implement the required processing of the noisy speech signal in the frequency domain following the application of the Fourier transform. Furthermore, they make little use of *a priori information* that may be available about the target speech and noise background. In this thesis, our interest lies in a new signal processing framework, known as the modulation domain, for the enhancement of speech signal. Furthermore, we will seek to incorporate *a priori* information about the speech and noise signals to boost the enhancement performance.

1.2 Literature Review

In this section, a literature review on specific topics related to the thesis research, especially modulation domain processing and codebook assisted parametric speech enhancement, is presented. A more detailed technical description of these topics along with their mathematical formulations will be found in later chapters.

1.2.1 Modulation Domain Processing

Most of the conventional speech enhancement methods discussed in Section 1.1 typically involve implementation of a three-stage processing framework known as analysis-modification-synthesis (AMS) [24–26]:

Analysis stage : In this stage, the short-time Fourier transform (STFT) is applied on successive (windowed) frames of the noisy speech signal;

Modification stage : The spectrum of the noisy speech is modified by means of mathematical operations for achieving noise suppression;

Synthesis stage : The enhanced speech is recovered by performing inverse STFT followed by overlap-add synthesis (OLA) [25] in the time domain.

Recent research has shown that extension of this framework into the modulation domain may result in improved noise suppression and better speech quality [27–29]. Modulation frequency domain refers to another layer of frequency representation obtained by applying the STFT a second time (but with longer frames and frame advance) on the spectral amplitudes of a speech signal. These frequency components, known as modulation frequencies, are meant to model the "slow" temporal variations of individual frequency components of a spectrum. Research in auditory physiology has shown that cells in the auditory cortex are best driven by sounds that combine both spectral and temporal modulations. Based on these evidences, Kowalski *et al.* [30–32] have even postulated that that the auditory system performs a spectro-temporal analysis which re-encodes the acoustic spectrum of sound waves in terms of its spectral and temporal modulations. Acoustic experiments have shown convincing evidence that the auditory system has channels which are tuned for the detection of modulation frequencies, different from the better known channels (such as critical bands or auditory filters) tuned for the detection of spectral frequency in [33]. Emerging evidences from studies on speech perception tend to demonstrate that the most important perceptual information lies at specific modulation frequencies below 16Hz [34].

In particular, faithful representation of these modulations is critical for the perception of speech.

These physio-auditory, psycho-acoustic and auditory-perception evidences underline the significance of a processing framework which works on modulation frequencies as well as spectral frequencies. Modulation domain based speech enhancement strives to improve the quality of corrupted speech by reproducing the modulation of the spectral amplitudes of speech signal better than conventional AMS based methods. In the case of spectral subtraction, experimental evidence suggests that musical noise distortion is lesser when the subtraction is performed in the modulation domain than in the conventional frequency domain [27]. Other approaches such as Kalman filtering [29] and MMSE STSA [28] have also shown positive results when extended to the modulation domain.

1.2.2 Codebook Assisted Speech Enhancement

Most speech enhancement algorithms, including those operating in the modulation domain, require an estimate of the background noise power spectral density (PSD). Noise estimation methods are broadly classified into two groups, namely hard decision and soft decision methods. In hard decision methods, the noise statistics are tracked only during silence or noise-only periods, i.e., when the signal lacks any speech utterances. A voice activity detector (VAD) can be used to identify these silence periods in a noisy speech signals [35–37]. This approach for noise PSD estimation works reasonably well when the background noise remains stationary through the alternation of speech activity and silence periods. However, the noise estimates tend to be highly inaccurate when the background noise exhibits a non-stationary behaviour. Problems also arise while estimating the background noise PSD for low signal-to-noise ratio (SNR) speech signals, where the VAD often makes an incorrect decision.

Soft decision methods track the noise PSD even during speech activity, and hence, in these methods, the noise estimate is updated more frequently than in hard decision methods. Noise estimation methods based on the minimum statistics approach generally fall under the soft decision category [38–42]. Martin [38] proposed a soft decision method which involves tracking the minima of the noisy speech STFT over a finite time window. Cohen *et al.* [39] proposed a new method known as the minima controlled recursive averaging (MCRA), in which the noise estimate is updated by tracking noise only regions of the noisy

speech spectrum over time. This tracking of noise only periods is done by calculating the speech presence probability in each frequency bin. After identifying a silence period, the noise estimate is updated for the current frame by averaging it recursively with the noise PSD of a previous frame. Cohen later proposed an updated iteration of MCRA called as the improved minima controlled recursive averaging (IMCRA) [40].

Minimum statistics and its offshoots [38–42] assume that the background noise exhibits a slowly varying behaviour while performing the PSD estimation. This may not be the case in environments with rapidly changing background noise such as, e.g., a street intersection with passing vehicles or a busy airport terminal or a train station with intercom announcements. Whilst these soft decision methods are more robust than the hard decision estimation methods, they do produce inaccurate noise estimates due to the inherent lag created by the minima search operation particularly when the background noise shows rapidly changing behaviour. This update delay results in inaccurate noise estimation, especially if the noise characteristics change suddenly.

Codebook based approaches [43–47] try to overcome this limitation by estimating the noise parameters based on *a priori* knowledge about different speech and noise types. In these approaches, joint estimation of the speech and noise PSD is performed on a frame-by-frame basis by exploiting *a priori* information stored in the form of trained codebooks of short term prediction parameter vectors. These codebooks are generated during a preliminary training stage that employs pre-selected speech and noise signal samples for the intended application. Examples of the short term parameters used in these codebooks are the gain normalized linear predictive (LP) coefficients [43–46] and the cepstral coefficients [47, 48]. Specifically, the codebooks are trained by windowing a set of representative speech and noise signal samples into frames, obtaining the corresponding short term prediction parameter vector for each frame, and finally performing vector quantization on these vectors to obtain the final set of representative codebook vectors [49].

These codebooks are then used during the enhancement stage to jointly estimate the speech and noise PSDs, where the joint estimation is done on a frame-by-frame basis. This estimation amounts to finding the best combination of gain variance and short term parameter vectors from the codebook for both speech and noise spectra, based on the observed noisy speech spectrum. Several approaches, including maximum likelihood, maximum a posteriori (MAP), MMSE and hidden Markov models (HMM) are available to tackle this issue, as proposed in, e.g., [43–47]. Once the speech and noise PSD have been properly

estimated, they can be used to compute a gain function for the purpose of speech enhancement, as per previously described enhancement methods such as MMSE-STSA or Wiener filtering.

1.3 Thesis Objective and Contribution

The use of these codebook methods in the acoustic AMS framework has shown promising results in the enhancement of speech corrupted by non-stationary noise [43–47]. However, to the best of our knowledge, they have not been applied yet to the modulation domain framework, where they could also bring similar benefits. The main objective of this thesis is therefore to combine the concepts of codebook driven speech and noise estimation with speech enhancement in the modulation domain and determine if by proceeding in this way, additional performance gains can be obtained.

In this thesis, research towards this objective has contributed to the proposal of a new speech enhancement method [1]. This enhancement procedure employs a Bayesian STSA spectral estimator in the modulation domain, which incorporates codebook assisted noise and speech PSD estimates obtained based on the work of Srinivasan *et al.* [43, 45, 46]. The proposed method consists of two stages, that is, codebook training and enhancement processing. During the training stage, we use codebooks of gain normalized linear prediction coefficients obtained using the Linde-Buzo-Gray (LBG) algorithm [49], as representative *a priori* information of the speech and noise spectra. In the enhancement stage, the speech and noise PSD estimates derived from the codebook are transformed into the modulation domain, where they are used to estimate the *a priori* and *a posteriori* signal-to-noise ratio (SNR) of the noisy speech. These SNRs are used to develop a gain function based on the MMSE-STSA criterion [9]. This gain function is applied to the magnitude spectrum of the noisy speech in the modulation domain in order to suppress noise. The modified magnitude spectrum is combined with the modulation phase spectrum of the noisy signal. The resulting spectrum is transformed into acoustic frequency domain through a procedure similar to the inverse FFT and OLA synthesis used in traditional AMS. Finally, once in the frequency domain, the inverse FFT and OLA synthesis are applied again to obtain the enhanced speech signal in time domain.

[1] Part of this thesis will be presented at the 3rd IEEE Global Conference on Signal and Information Processing, to be held in Orlando, USA, in Dec. 2015

The performance of this newly proposed speech enhancement method is compared with standard enhancement methods such as MMSE-STSA [9], modulation domain based MMSE-STSA (MME) [28]and codebook-based Wiener filtering [45]. The enhanced speech signals are subjected to a series of objective evaluations such as the ITU standardised Perceptual Evaluation of Speech Quality (PESQ) measure [116] and Segmental SNR (SegSNR). Results of objective evaluations indicate improvement in noise suppression with the proposed codebook-based speech enhancement method over the other benchmark methods, particularly in cases of non-stationary noise.

1.4 Thesis Organization

In Chapter 2, the concept of short-time modulation domain processing is covered in detail. Selected speech enhancement methods based on modulation domain processing, such as spectral subtraction [27] and MMSE spectral estimator [28], are reviewed.

Chapter 3 covers conventional approaches used for noise estimation. The primary focus is on the minimum statistics approach [38] and some of its offshoots, such as Cohen's Minima controlled recursive averaging [39, 40]. The accuracy of some of these estimation procedures is compared for various noise types. Their limitation with regards to noise estimation in non-stationary conditions is discussed as well.

Chapter 4 deals with the estimation of speech and noise PSDs with the help of trained codebooks of short term prediction parameters. Topics such as the training of speech and noise codebooks and joint estimation of speech and noise PSDs are covered in some detail.

In Chapter 5, speech enhancement methods which employ these codebook based estimates are presented. This includes the proposed speech enhancement method, which combines MMSE-based modulation domain processing along with codebook based speech and noise PSD estimation, known as CB-MME. A soft decision Weiner filter is also implemented as suggested from previous work in this area.

Chapter 6 presents the experimental set up used to investigate the performance of the proposed speech enhancement methods. Details regarding test speech utterances, noise types and benchmark methods are listed. The results of the experiments are presented and discussed.

Some concluding remarks regarding the thesis research are provided in Chapter 7.

Chapter 2

Modulation Domain Framework

In this chapter, modulation domain processing is covered in detail. The chapter starts with a brief overview of the conventional acoustic domain processing in Section 2.1, which covers the AMS framework and some chosen speech enhancement methods in the acoustic domain. In Section 2.2, the discussion shifts towards modulation domain for which a concise background review is provided, including various evidences and findings which buttress the significance of modulation domain in the human auditory system. Following that, some recently proposed speech enhancement methods in the modulation domain are discussed in some detail.

2.1 Acoustic Domain Processing

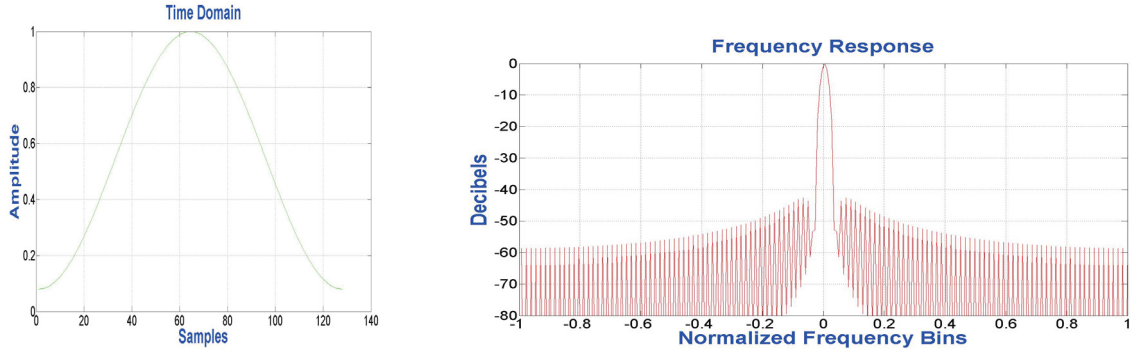
2.1.1 The AMS framework

Conventional speech enhancement methods implement the AMS framework in the acoustic frequency domain, where the acoustic frequency spectrum of a speech signal is defined by its STFT. To this end, an additive noise model is assumed, i.e.,

$$x[n] = s[n] + d[n], \tag{2.1}$$

where $x[n]$, $s[n]$ and $d[n]$ refer to the noisy speech, clean speech and noise signals respectively, while $n \in \mathbb{Z}$ is the discrete-time index. STFT analysis of (2.1) results in,

$$X(\nu, k) = S(\nu, k) + D(\nu, k) \tag{2.2}$$



(a) Time samples (N =128) (b) Frequency response

Fig. 2.1 Time domain representation and frequency response of Hamming window

where $X(\nu,k)$, $S(\nu,k)$ and $D(\nu,k)$ refer to the STFTs of the noisy speech, clean speech and noise signals, respectively, while ν and k are the time frame and discrete spectral frequency bin indices, respectively. The STFT $X(\nu,k)$ is obtained from,

$$X(\nu,k) = \sum_{n=-\infty}^{\infty} x(n)w(n-\nu F)e^{-2jkn\pi/N} \qquad k \in \{0,1,2....,N-1\} \tag{2.3}$$

where $w(n)$ is a windowing function of duration N samples, and F is the frame advance. We note that N also refers to the number of frequency bins in the STFT $X(\nu,k)$. The windowing function is chosen based on various considerations such as main lobe width, side lobe peak level, spectral leakage and equi-ripple effect. In this work, the Hamming window is used for this purpose in the analysis stage. The mathematical definition for the Hamming window is given by (2.4),

$$w(l) = \begin{cases} \alpha - \beta\cos(\frac{2\pi l}{N}), & \text{if } 0 \le l \le N \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

where $\alpha \approx 0.54$ and $\beta = 1-\alpha \approx 0.46$. The time and frequency domain representations of the Hamming window for $N = 128$ are shown in Figure 2.1.

The STFT of a signal is represented by its acoustic magnitude and phase spectra as,

$$X(\nu,k) = |X(\nu,k)|e^{j\angle X(\nu,k)} \tag{2.5}$$

Most single channel speech enhancement methods, such as spectral subtraction [5, 6, 8], Bayesian spectral estimation methods [9–13], etc., implement the modification part of the AMS framework by modifying the noisy magnitude spectrum whilst retaining the phase spectrum[1]. A more detailed discussion of these enhancement methods will be presented in Subsection 2.1.2. Without loss of generality, the modification stage of an AMS based speech enhancement method can be expressed as follows,

$$\hat{S}(\nu,k) = f(X(\nu,k), |D(\nu,k)|) \tag{2.6}$$

where $\hat{S}(\nu,k)$ is the STFT of enhanced speech signal, $f(\cdot)$ is a function that represents the spectral modification made to the noisy speech spectrum ($X(\nu,k)$) which depends on the enhancement method being used, and $|D(\nu,k)|$ is an estimate of the short term magnitude spectrum of the background noise.

Synthesis of the enhanced signal is performed by applying inverse STFT on the modified spectrum followed by performing Overlap Add synthesis (OLA) [25] on the resulting frames. The inverse STFT transforms the frequency elements back into time domain, according to

$$\hat{s}_\nu(n) = \begin{cases} \frac{1}{N}\Big(\sum_{k=0}^{N-1} \hat{S}(\nu,k)e^{j2\pi kn/N}\Big)w_s(n), & \text{if } n \in \{0,1,2....,N-1\} \\ 0, & \text{otherwise} \end{cases} \tag{2.7}$$

where $w_s(n)$ is the synthesis window. The enhanced signal is reconstructed by overlapping and adding the time shifted frames of $\hat{s}_\nu(n)$ as in,

$$\hat{s}(n) = \sum_{\nu=-\infty}^{\infty} \hat{s}_\nu(n-\nu F) \tag{2.8}$$

To ensure that the windowing procedures in analysis and synthesis windowing processes do not introduce unwanted modifications to the speech signal overall, the analysis and synthesis

[1]There are a few enhancement methods which modify both the magnitude and the phase spectra of a noisy speech signal. Wiener filter based speech enhancement [14] is an example of such methods.

windows are required to satisfy the following condition for perfect signal reconstruction,

$$\sum_{\nu=-\infty}^{\infty} w_s(n-\nu F)w(n-\nu F) = 1 \qquad \forall n. \tag{2.9}$$

A pictorial representation of the AMS procedure is presented in Fig. 2.2.

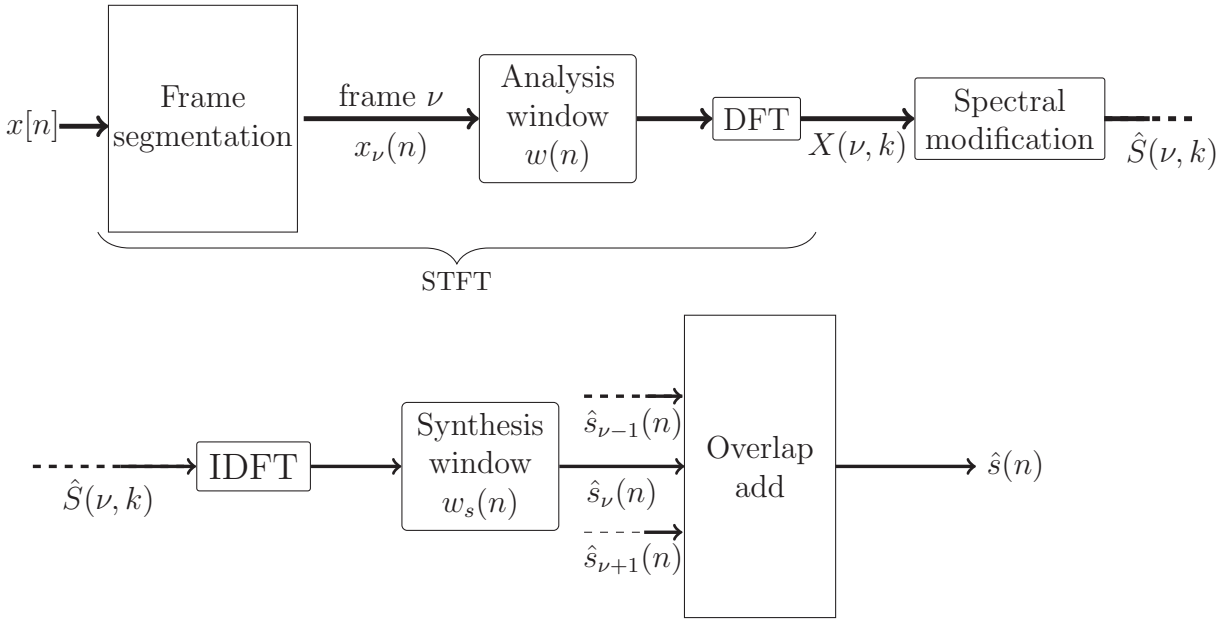


Fig. 2.2 AMS STFT framework for single channel speech enhancement

2.1.2 Acoustic Domain based Speech Enhancement Methods

Spectral subtraction is one of the earliest and widely used speech enhancement methods. It has proven to be effective for noise removal, while its simplicity of implementation makes it particularly attractive for real-time applications. Spectral subtraction methods [5, 6, 8] try to retrieve the magnitude spectrum of the clean speech by subtracting an estimate of the noise spectrum from the noisy speech spectrum. As mentioned in Chapter 1, Boll and Berouti conducted pioneering research work on developing spectral subtraction methods in digital domain [5, 6]. The mathematical description for a basic implementation of spectral subtraction is given by,

$$|\hat{S}(\nu,k)| = |X(\nu,k)| - |\hat{D}(\nu,k)| \tag{2.10}$$

where $|X(\nu,k)|$ is the magnitude spectrum of the noisy speech, $|\hat{D}(\nu,k)|$ is an estimate of magnitude spectrum of the background noise, and $|\hat{S}(\nu,k)|$ is the estimated clean speech spectrum following spectral subtraction. A spectral floor is usually introduced to ensure that the subtraction process does not result in negative spectral values. Whilst being simple in implementation, spectral subtraction results in the introduction of residual noise artefacts. These artefacts are mostly composed of tones at random frequencies and time, resulting in a perceptually annoying form of noise known as musical noise. This residual noise typically results from the mismatch between the noise spectral estimate and actual noise present in the noisy speech.

More sophisticated non-linear subtraction methods have been suggested for dealing with the musical noise. Virag suggested a non-linear spectral subtraction method which takes the properties of human auditory system into account [8]. The main equation for this non-linear subtraction method is given by,

$$|\hat{S}(\nu,k)| = \begin{cases} (|X(\nu,k)|^{\gamma} - \alpha|\hat{D}(\nu,k)|^{\gamma})^{\frac{1}{\gamma}}, & \text{if } (|X(\nu,k)|^{\gamma} - \alpha|\hat{D}(\nu,k)|^{\gamma})^{\frac{1}{\gamma}} > \beta|\hat{D}(\nu,k)| \\ \beta|\hat{D}(\nu,k)|, & \text{otherwise} \end{cases} \tag{2.11}$$

where the over-subtraction factor $\alpha \geq 1$ generally has a linear dependence on the SNR, the spectral floor parameter $0 < \beta < 1$ ensures that the enhanced speech magnitude spectrum does not fall below a certain value, and γ sets the domain in which the spectral subtraction takes place. Specifically, with $\gamma = 1$ the subtraction is realized in the magnitude spectral domain, while with $\gamma = 2$ it is realized in the power spectral domain.

MMSE-STSA estimation [9] is another popular enhancement method which implements AMS in the acoustic frequency domain. It models the clean speech and noise magnitude spectra in terms of Gaussian distributions and employs a Bayesian approach to estimate the amplitude spectrum of the clean speech. An estimate of the clean speech spectrum is obtained by minimizing the mean square error $\mathcal{E}$ between the clean and estimated speech spectra, defined as

$$\mathcal{E} = \mathrm{E}[(|S(\nu,k)| - |\hat{S}(\nu,k)|)^2] \tag{2.12}$$

where $|S(\nu,k)|$ is the short term magnitude spectrum of clean speech signal, $|\hat{S}(\nu,k)|$ is the desired estimate and $\mathrm{E}[\cdot]$ denotes statistical expectation. The solution to (2.12) is given by

the conditional expectation of $|S(\nu,k)|$), that is,

$$|\hat{S}(\nu,k)| = \mathrm{E}\big[|S(\nu,k)| \;\big|\; X(\nu,k)\big] \tag{2.13}$$

where $X(\nu,k)$ is the STFT coefficient of the observed noisy speech signal. In the sequel, to simply the presentation, the time frame index ν is dropped from (2.13), which we now express as[2],

$$|\hat{S}_k| = \mathrm{E}\big[|S_k| \;\big|\; X_k\big] = \iint |S_k| f_{S_k|X_k}(S_k|X_k)\, dS_k \tag{2.14}$$

where $f_{S_k|X_k}(.)$ is the conditional probability density function (PDF) of the clean speech spectrum S_k given the observed noisy speech spectrum X_k. Through Bayes rule, (2.14) can be reinterpreted as,

$$|\hat{S}_k| = \frac{\iint |S_k| f_{X_k|S_k}(X_k|S_k) f_{S_k}(S_k)\, dS_k}{\iint f_{X_k|S_k}(X_k|S_k) f_{S_k}(S_k)\, dS_k} \tag{2.15}$$

Since the background noise is additive and independent of the clean speech signal (i.e. $X_k = S_k + D_k$), the conditional PDF $f_{X_k|S_k}(X_k|S_k)$ can be expressed as a shifted version of the background noise PDF, i.e.,

$$f_{X_k|S_k}(X_k|S_k) = f_{D_k}(X_k - S_k) \tag{2.16}$$

Applying (2.16) in (2.15) we obtain,

$$|\hat{S}_k| = \frac{\iint |S_k| f_{D_k}(X_k - S_k) f_{S_k}(S_k)\, dS_k}{\iint f_{D_k}(X_k - S_k) f_{S_k}(S_k)\, dS_k} \tag{2.17}$$

For solving (2.17), the STFT coefficients of clean speech and noise are modelled as statistically independent zero-mean, circular Gaussian random variables,

$$f_{S_k}(S_k) = \frac{1}{\pi\sigma^2_{S_k}} e^{-|S_k|^2/\sigma^2_{S_k}} \tag{2.18}$$

$$f_{D_k}(D_k) = \frac{1}{\pi\sigma^2_{D_k}} e^{-|D_k|^2/\sigma^2_{D_k}} \tag{2.19}$$

[2]Note that the STFT is complex-valued, hence the need of a double integral in (2.14)

where $\sigma^2_{S_k}$ and $\sigma^2_{D_k}$ are the variances of the speech and noise spectra respectively. Substituting (2.18) and (2.19) in (2.17) yields the final form of the MMSE STSA estimator,

$$|\hat{S}_k| = G_k|X_k| \tag{2.20}$$

where

$$G_k = \frac{\sqrt{\pi\zeta_k}}{2\gamma_k} \exp\left(\frac{-\zeta_k}{2}\right)\left[(1+\zeta_k)I_0\left(\frac{-\zeta_k}{2}\right) + \zeta_k I_1\left(\frac{-\zeta_k}{2}\right)\right] \tag{2.21}$$

G_k is the gain applied to the spectral amplitudes of the noisy speech signal. In (2.21), $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel functions of zeroth and first order respectively [50], while ζ_k is a SNR parameter given by,

$$\zeta_k = \frac{\xi_k}{1+\xi_k}\gamma_k \qquad \gamma_k = \frac{|X_k|^2}{\sigma^2_{D_k}} \qquad \xi_k = \frac{\sigma^2_{S_k}}{\sigma^2_{D_k}} \tag{2.22}$$

In the literature, the parameters ξ_k and γ_k are referred as the *a priori* and *a posteriori* SNRs of the noisy speech. The *a posteriori* SNR is the ratio of the noisy speech power spectrum, $|X_k|^2$, and the variance of the background noise, $\sigma^2_{D_k}$. The *a priori* SNR is the ratio of the variances of clean speech and background noise spectra. From (2.21), we can note that the calculation of the spectral gain function G_k requires estimates of the *a priori* and *a posteriori* SNRs.

In practice, $|X_k|^2$ is known by observation while $\sigma^2_{D_k}$ can be obtained by using a noise PSD estimation method as will be explained in Chapter 3. The estimated *a posteriori* SNR can be written as,

$$\hat{\gamma}_k = \frac{|X_k|^2}{|\hat{D}_k|^2} \tag{2.23}$$

where $|\hat{D}_k|^2$ is the noise PSD estimate. The a *priori* SNR can be obtained as the ratio of clean speech and background noise power spectral estimates,

$$\hat{\xi}_k = \frac{|\hat{S}_k|^2}{|\hat{D}_k|^2} \tag{2.24}$$

It can also be expressed as the expected value of $\gamma_k - 1$ [51],

$$\hat{\xi}_k = \mathrm{E}(\gamma_k - 1) \tag{2.25}$$

Ephraim and Malah [9] proposed a *decision directed* approach which combined both (2.24) and (2.25) for estimating ξ_k over consecutive frames,

$$\hat{\xi}(\nu,k) = \tau\frac{|\hat{S}(\nu-1,k)|^2}{|\hat{D}(\nu-1,k)|^2} + (1-\tau)\max[\gamma(\nu,k)-1,0] \tag{2.26}$$

where $|\hat{S}(\nu,k)|$ denotes the MMSE STSA estimate of the clean speech at frequency bin k from a previous frame $\nu-1$, $|\hat{D}(\nu-1,k)|^2$ represents the estimated PSD of the background noise at frequency bin k from a previous frame $\nu-1$, $\gamma(\nu,k)$ is the *a posteriori* SNR for the current frame ν and τ is a recursion averaging parameter with a typical value in the range $0.95 \leq \tau < 1$. This recursive smoothing procedure eliminates large variations across successive frames and therefore reduces the musical noise in the resultant enhanced speech [52]. However, it will respond slowly to abrupt changes in the instantaneous SNR.

Other commonly used single channel acoustic domain based speech enhancement methods include the Log MMSE STSA estimator [10], Weighted-Euclidean STSA (WE STSA) estimator [11], the β-SA estimator [12] and Weightedβ-SA (Wβ-SA) estimator [13], Wiener filtering [14], and signal subspace methods [19–23]. We refer the interested readers to the related publications to get more acquainted with these methods.

2.2 Background on Modulation Domain

The spectral envelope of the speech signal has been traditionally considered as the principal carrier of information. Therefore, much of the works done in various areas of speech processing such as Automatic Speech Recognition (ASR), speech coding, and speech enhancement have been focused towards processing of the spectral envelope. Speech enhancement in particular, is mostly based on short-time processing of the speech signal within the AMS framework, as seen in Section 2.1. In past years, the temporal modulations of the spectral envelope have been receiving considerable attention with regards to research in speech applications. Increasing evidences from research over the past few decades suggest that low frequency modulations of audio and speech signals might be serving as important carriers of information in speech rather than the spectral envelope itself as previously perceived.

In this section, we will review the psychophysical, physiological, and other sources of evidence which support the role of temporal modulation frequencies in speech processing

as originally presented in [53].

2.2.1 Physiological, Psycho-Acoustical and Perceptual Evidences in Literature

H. Dudley, in 1939, became one of the first researchers to observe and conclude that speech signals behaved like low bandwidth processes that modulate on higher frequency carriers [54]. In 1971, Muller observed that the auditory system of mammals show a specialized sensitivity towards the amplitude modulations of narrowband signals [55]. He made this observation by studying the auditory response of rats to amplitude modulated (AM) and frequency modulated (FM) tonal signals. Subsequent research in auditory physiology of various mammals such as bats, rats etc., by Suga [56], Schreiner and Urbas [57] and others showed that this modulated amplitude information is stored in the form of neuron discharges which are preserved through all the levels of the auditory system, starting from the cochlear frequency channels up to higher levels like the auditory cortex.

Research by Kowalski *et al.* [30–32] has shown that cells in the auditory cortex, the highest processing stage in primary auditory pathway, best correspond to sound that combine both spectral and temporal modulations. Based on their experimental results, they postulated that the mammalian auditory system performs a multiscale spectro-temporal analysis on the acoustic spectrum of incoming sound waves and re-encodes this spectrum in terms of its spectral and temporal modulations.

Psychoacoustics generally deals with perception of sounds and responses associated with them. Studies on human perception of amplitude modulation [58] and frequency masking related to the perception of these modulations [59] have led to the postulation that the auditory system has separate channels which are tuned for the detection of modulation frequencies. These modulation frequency channels are different from the better known channels (such as critical bands or auditory filters) tuned for the detection of acoustic spectral frequencies [33]. Sheft and Yost [60] showed that human perception of temporal dynamics corresponds to the perceptual filtering of these modulation frequency channels.

Research on speech perception has shown the significance of temporal modulations of speech in terms of perceived speech quality and intelligibility. Investigation into the role of temporal modulation of spectral envelopes in determining the intelligibility of the speech signal was conducted by Drullman *et al.* [34, 61] and Arai *et al.* [62]. Their experiments involved application of various high-pass, low-pass and band-pass filters on the temporal

envelopes of the acoustic frequency bands. The effects of these filtering operations on the intelligibility of the speech signals were studied. These studies resulted in the conclusion that the most important perceptual information is contained in modulation frequencies below 16 Hz, 4-5 Hz being the most significant. Based on these research findings, low frequency temporal modulations of sound spectra have been suggested to be fundamental carriers of information in speech [53]. Preservation of the temporal modulations has been found to be necessary for maintaining the quality and intelligibility of speech signals. This fact has been used to develop objective measurement indices like the speech transmission index (STI) and its variants such as spectro-temporal modulation index (STMI), for assessing the quality and intelligibility of speech utterances, which have found widespread use in areas like automatic speech recognition [63–65].

From the findings reported in this section, it becomes clear that the properties of amplitude modulations can play an important role in speech processing. This has prompted researchers to focus on concocting new approaches to speech processing which take the characteristic features of temporal modulations of speech spectra into consideration.

2.2.2 Applications of Modulation Domain Processing

Modulation domain processing has grown in popularity, finding numerous applications in various areas of speech signal processing over the years. This has been a result of all the findings such as those presented in Subsection 2.2.1. In particular, modulation domain processing has been used to varying degrees of success in the solution of key problems in speech processing, such as,

- Speech coding [67, 68];
- Speech recognition [69–74];
- Speaker recognition [75–78];
- Objective speech intelligibility evaluation [34, 61, 79–83];
- Speech enhancement (Detailed description below in Section 2.3).

2.3 Speech Enhancement in Modulation Domain

A number of modulation domain based methods have been proposed for speech enhancement over the years. Early enhancement methods generally focused on development and application of filter banks on the modulation frequencies. A concise review of some of these filter bank methods is presented in the subsection below with [66] as the source of information. Later, an extension of the AMS framework into the modulation domain is also discussed.

2.3.1 Filter Bank based Methods

Speech enhancement methods based on modulation domain were first developed as a part of ASR systems which operated on modulation domain information [69, 70]. Langhans and Strube [84] reported an improvement in intelligibility when a modulation filterbank, based on the modulation transfer function (MTF), was applied to the logarithmic subband envelopes of the speech spectrum instead of the subband envelopes themselves.

Hermansky *et al.* [69, 85] proposed a noise suppression filter bank method based on RASTA-PLP, a speech recognition technique they had proposed earlier [86]. The RASTA-PLP (expanded as the *relative spectral transform - perceptual linear prediction*) signal analysis technique is one of the first speech recognition techniques to use modulation domain processing. It generally involves application of bandpass filters which suppress the spectral components which change slowly or quickly as compared to the speech spectrum. Their earlier work [69] involved application of a $5th$ order infinite impulse response (IIR) bandpass filter on the temporal modulation of the cubic-root of speech power spectrum with 1 Hz and 15 Hz used as the cut off frequencies. This method led to a reduction in background noise albeit with some resultant distortion. In [85], a finite impulse response (FIR) modulation domain filter which resembled a Wiener filter in its behaviour was developed by training on speech samples with additive noise. This approach resulted in better speech quality with lower phase distortion as compared to their previous work [69]. Avendano *et al.* followed up this approach by incorporating the local SNRs into their FIR filter design [87]. This work was motivated by the observation made in [85] that the characteristics of the filter were related to the local SNRs for each frequency subband. The filtered signal is free of any musical noise, but appears to suffer from resulting fluctuations in levels and residual noise. These issues were reported to be worse when there was a mismatch between the

background noise in training sample used for designing the filter and the actual background noise present in the noisy speech. Mesgarani and Shamma proposed a 2-dimensional (2D) spectro-temporal filtering operation on the 2-dimensional wavelet transform of the noisy speech signal [88]. This filterbank operated on the joint spectro-temporal modulation of the noisy speech in order to suppress noise.

2.3.2 Extension of AMS into Modulation Domain

Modulation domain based filtering methods show some benefits in terms of speech enhancement over the conventional acoustic domain based methods. However, these methods are generally plagued by some serious limitations which render them incapable of widespread use for speech enhancement [66]:

- The filters are usually designed based on the long term properties of the speech signal and are fixed in nature. They are usually applied on the entire signal and hence assume both speech and background noise to show a stationary behaviour. This is usually not the case in real world where we have to deal with mostly non-stationary noises.

- Since most of the filters are designed solely based on the long term properties of the modulation spectrum of the speech signal, the filter performs reasonable noise suppression in silence periods. However, they fail to eliminate background noise present within the speech regions of the modulation spectrum.

Research has been undertaken to overcome above limitations posed by modulation domain filtering. So and Paliwal developed a Kalman filtering method in the modulation domain which does not require stationarity of speech and background noise signals [29]. This method involves the application of Kalman filters on the magnitude spectrum for every frequency bin present in the STFT of the noisy speech signal.

In past few years, the focus has been set on a different approach to modulation domain based enhancement which uses a frame-by-frame analysis similar to the AMS procedure present in conventional enhancement methods. This extended AMS framework in modulation domain requires a multi-dimensional representation of the speech signal which contains the spectral frequencies, obtained from Fourier analysis, along with their temporal modulations. The idea of a bi-frequency system has been around for a long time.

Zadeh [89] was the first to propose a separate dimension for modulation frequency in frequency analysis of signals. His proposed transform had two separate frequency dimensions: the standard spectral frequencies and the transform of the time variation of spectral frequencies. Kailath [90] analysed this joint bi-frequency system function in his work. Atlas and Shamma [53] formulated an analysis-synthesis framework with modulation frequencies which has found applications in many areas of speech processing. This framework has been capable of perfectly reconstructing a signal following its initial analysis.

Paliwal and Wojcicki [27] developed an extended AMS framework for performing speech enhancement in the modulation domain through spectral subtraction. This framework results in a system function consisting of two separate frequencies. For convenience of representation, we will here on differentiate between the conventional and modulation frequencies. The conventional frequency content will be represented by the acoustic spectrum $X(\nu, k)$, which is obtained by performing STFT on the signal $x(n)$ as in (2.3), and where k is the acoustic frequency bin and ν is the frame index in the time domain. In turn, the calculation of the modulation spectrum at any given frequency bin k will involve performing the STFT analysis on the time trajectories of the acoustic magnitude spectrum $|X(\nu, k)|$ over the frame index ν [27]. Specifically, the magnitude spectrum of the noisy speech in each acoustic frequency bin, i.e. $|X(\nu, k)|$, is first windowed and then Fourier transformed again, resulting into the modulation spectrum,

$$Z(t, k, m) = \sum_{\nu=-\infty}^{\infty} |X(\nu, k)| w_M(\nu - tF_M) e^{-2j\nu m\pi/M} \tag{2.27}$$

where $w_M(n)$ is the so-called modulation window of length N_M, $m \in \{0, ..., M-1\}$ is the modulation frequency index, t is the modulation time-frame index, and F_M is the modulation frame advance. The resulting modulation spectrum can be expressed in polar form as,

$$Z(t, k, m) = |Z(t, k, m)| e^{j\angle Z(t,k,m)} \tag{2.28}$$

where $|Z(t, k, m)|$ is the modulation magnitude spectrum and $\angle Z(t, k, m)$ is the modulation phase spectrum.

Speech enhancement in the modulation domain involves modification of the modulation spectrum of the noisy speech signal, as in,

$$\hat{S}(t,k,m) = f(Z(t,k,m), \hat{D}(t,k,m)) \tag{2.29}$$

where $\hat{S}(t,k,m)$ is the modified modulation spectrum from which the enhanced speech signal is obtained, $Z(t,k,m)$ is the modulation spectrum of noisy speech, $f(\cdot)$ is a function that represents the spectral modification made to the noisy speech modulation spectrum $Z(t,k,m)$, which depends on the enhancement method used, and $\hat{D}(t,k,m)$ is the modulation spectrum of the estimated background noise spectrum.

The modified magnitude acoustic spectrum is obtained by performing an inverse STFT operation on the modified modulation spectrum $\hat{S}(t,k,m)$ followed by an OLA synthesis [25]. Specifically, we have

$$|\hat{S}_t(\nu,k)| = \frac{1}{M} \sum_{m=-M/2}^{M/2-1} \hat{S}(t,k,m) e^{j2\pi m\nu/M} \tag{2.30}$$

$$|\hat{S}(\nu,k)| = \sum_t |\hat{S}_t(\nu - tF_M, k)| w_M^s(\nu - tF_M) \qquad \forall\nu \tag{2.31}$$

where $|\hat{S}(\nu,k)|$ is the acoustic magnitude spectrum of recovered speech signal. $w_M^s(\cdot)$ is the window used for synthesis of modified spectrum in acoustic domain. The magnitude spectrum, $|\hat{S}(\nu,k)|$, can be combined with the acoustic phase spectrum, $\angle X(\nu,k)$, and used to synthesize the enhanced speech signal based on the conventional AMS procedure seen in Subection 2.1.1.

As mentioned earlier, Paliwal and Wojcicki [27] implemented a spectral subtraction algorithm in the extended modulation framework. The enhanced speech modulation magnitude spectrum $|\hat{S}(t,k,m)|$ is determined by the following equation which is based on the acoustic domain spectral subtraction [6]:

$$|\hat{S}(t,k,m)| = \begin{cases} (|Z(t,k,m)|^\gamma - \alpha|D(t,k,m)|^\gamma)^{\frac{1}{\gamma}}, & \text{if } |Z(t,k,m)| > \beta|D(t,k,m)| \\ \beta|D(t,k,m)|, & \text{otherwise} \end{cases} \tag{2.32}$$

In (2.32), γ is a factor which determines the domain on which the subtraction takes place. If $\gamma = 1$, the subtraction is performed in the modulation magnitude spectral domain while if $\gamma = 2$, the subtraction is performed in the modulation power spectral domain. α is a subtraction factor which varies linearly with the local SNR, β is a spectral floor parameter which ensures that the subtraction does not result in values going below a certain pre-chosen spectral floor.

Paliwal *et al.* followed up by developing a MMSE based Bayesian estimator for speech enhancement in modulation domain [28], where the calculation of processing gain is based on the conventional MMSE STSA estimator [9]. Specifically,

$$|\hat{S}(t,k,m)| = G(t,k,m)|Z(t,k,m)| \tag{2.33a}$$

$$G(t,k,m) = \frac{\sqrt{\pi\zeta}}{2\gamma} \exp\left(\frac{-\zeta}{2}\right) \left[(1+\zeta) I_0\left(\frac{-\zeta}{2}\right) + \zeta I_1\left(\frac{-\zeta}{2}\right)\right] \tag{2.33b}$$

where $G(t,k,m) > 0$ is the optimal processing gain. $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel functions of zeroth and first order respectively, $\zeta \equiv \zeta(t,k,m)$ is a SNR parameter obtained from the *a priori* and *a posteriori* SNRs in the modulation domain (denoted by $\xi \equiv \xi(t,k,m)$ and $\gamma \equiv \gamma(t,k,m)$ respectively). The estimates of the *a priori* and *a posteriori* SNRs ($\hat{\xi}$ and $\hat{\gamma}$) are obtained in a similar fashion as in the conventional acoustic domain based MMSE method - The *a posteriori* SNR $\hat{\gamma}(t,k,m)$ is estimated as a ratio between the noisy speech power spectra and estimated noise power spectra in modulation domain (2.34),

$$\hat{\gamma} \equiv \hat{\gamma}(t,k,m) = \frac{|Z(t,k,m)|^2}{|\hat{D}(t,k,m)|^2}. \tag{2.34}$$

The *a priori* SNR $\hat{\xi}(t,k,m)$ is estimated using a decision directed method similar to the one used in acoustic domain based MMSE estimator (2.35), that is,

$$\hat{\xi} \equiv \hat{\xi}(t,k,m) = \tau \frac{|\hat{S}(t-1,k,m)|^2}{|\hat{D}(t-1,k,m)|^2} + (1-\tau) \max\left[\hat{\gamma}(t,k,m) - 1, 0\right] \tag{2.35}$$

The SNR parameter $\zeta \equiv \zeta(t,k,m)$ is estimated using the estimates of the *a priori* and *a posteriori* SNRs from (2.34) and (2.35) as follows,

$$\hat{\zeta} = \frac{\hat{\xi}}{1+\hat{\xi}}\hat{\gamma} \tag{2.36}$$

The enhanced magnitude spectrum $|\hat{S}(t,k,m)|$ is combined with the modulation phase spectrum of the noisy speech to get the enhanced spectrum, which then is subjected to inverse STFT. This estimation method is referred to as the MME (Modulation domain based Mmse Estimator) [28].

The above methods perform enhancement modification on the modulation magnitude spectra of noisy speech signals. Zhang and Zhao proposed a modified enhancement approach based on a Real-Imaginary (RI) modulation framework, which performs spectral subtraction on the real and imaginary acoustic spectra of the noisy speech separately in the modulation domain [91]. This RI framework ensures that both the magnitude and phase of the noisy speech spectrum are enhanced. Noticeable improvements in magnitude and phase have been reported especially in the case of low SNR (highly noisy) speech signals for this method. Schwerin and Paliwal proposed a MMSE based spectral estimator in the RI modulation framework [92]. They reported better results than the spectral subtraction based method [91] based on objective experiments and subjective listening tests. They also reported that the perceived theoretical advantage of using a RI modulation framework, i.e. enhancement of noisy phase spectrum in addition to noisy magnitude spectrum, did not seem to impact the overall quality of speech signal as expected. Wang and Brookes developed a signal subspace method that operates in the modulation domain and appears to outperform conventional signal subspace method [19] as well as modulation domain based spectral subtraction [27] for colored noises [93]. In this work, the modulation domain MMSE spectral estimator [28] will be used as a basis for developing the proposed codebook-based CB-MME method.

Chapter 3

Noise Estimation Methods

This chapter delves into the topic of power spectral density (PSD) estimation of background noise present in speech signals. A brief review of the noise PSD estimation methods based on voice activity detection (VAD) is presented first. The focus then shifts towards soft decision methods based on minimum statistics. The minima controlled recursive averaging (MCRA) and its improved version (i.e., IMCRA) are described in detail since they have been reported to be widely in use for noise PSD estimation. Our discussion of these conventional noise estimation methods will include comments on their advantages and entailing drawbacks with regards to their performance under various conditions. Their estimation performance in the case of background noise with non-stationary behaviour will receive particular attention.

3.1 Background

Most speech enhancement methods require an estimate of the background noise statistics to work with. For example, the spectral subtraction method uses an estimate of the noise power spectrum for performing weighted subtraction from the power spectrum of the observed noisy speech [5, 6, 8], the MMSE STSA methods requires a noise spectral estimate for calculating the *a priori* and *a posteriori* SNRs [9], finally, the signal subspace methods require an estimate of the noise covariance matrix in their operations [19]. Accuracy of the noise power estimate plays a critical role in the performance of these enhancement methods. Underestimation of the noise power spectrum results in incomplete suppression of the background noise, and hence high level of residual noise, whereas overestimation can result in spectral distortion when enhancement is performed, which can lead to loss of speech

intelligibility.

In the context of speech enhancement, the noise PSD can be expressed as an expectation of the short term squared magnitude spectrum of the background noise, $|D(\nu, k)|^2$, i.e.,

$$P_d(\nu, k) \triangleq \mathrm{E}[|D(\nu, k)|^2] \tag{3.1}$$

As mentioned earlier in Subsection 1.2.2, conventional noise estimation methods can be categorized into two families, namely hard decision based methods and soft decision based methods. These methods, which attempt to derive an estimate of $P_d(\nu, k)$ in (3.1) from the noisy speech signal, are briefly are briefly described in the following sections.

3.2 VAD based Noise Estimation

Voice activity detection (VAD) has found use in many areas of speech processing such as speech coding, low bit rate transmission, feature extraction processes in speech recognition, and for noise estimation in speech enhancement [94]. The working of a VAD in noise estimation is based on the principle that the PSD statistics of the noisy speech devolves to the PSD of the background noise during silence periods [95].

In most standard VADs such as the ITU-T recommended G.729 Annex B [96] and ESTI standardized VADs (ESTI AMR 1 and 2) [97], an optimal decision rule is used as a classifier to identify silence and speech periods in the noisy speech signal, where a minimum threshold energy level is employed for differentiating between the two. The noise PSD is updated in every silence period frame by recursively averaging the squared magnitude spectrum of the observed signal in this new frame with the background noise PSD estimates from previous frames. This smoothing operation is performed in order to improve the robustness of the estimation algorithm and reduce the variance of the estimated background noise PSD.

Research over the years has resulted in the development of VADs which perform better than the standard ones mentioned above in the context of activity detection and noise estimation. For examples, J. Sohn *et. al* proposed a robust VAD which employs the decision-directed parameter estimation method in the likelihood ratio test [35]. This proposed VAD scheme has been reported to perform significantly better than the ITU standard G.729 B VAD under low SNR conditions. Sangwan *et. al* proposed a contextual VAD scheme which combines both contextual and frame specific information to improve the detection

performance [36]. This scheme takes cues to activity within the current frame as well as its neighbouring frames, unlike other standard VAD methods which assume that the cues to activity lie within the current frame alone. Experimental results reported in [36] show that this context based VAD method performs better than the standard algorithm ETSI AMR VAD-1.

3.2.1 Limitations of VAD based Noise Estimation

VAD based noise PSD estimation performs reasonably well when the background noise shows a stationary behaviour, such as white noise and pink noise etc. However, the various types of background noise encountered in real life rarely exhibit such a stationary behaviour. For example, background noise due to vehicles on a street, or speaker announcements in public areas such as train station, airport etc., are characterized by a rapidly changing behaviour. In such cases, estimating the noise statistics only during silence periods may not be adequate since these descriptors may have changed significantly between contingent speech and silence periods. This usually leads to inaccurate estimates of the background noise PSD which, in turn, drastically affects the performance of any speech enhancement method. It has also been noted that the ability of a VAD to identify silence and speech periods deteriorates as the SNR of the noisy speech signal declines. Based on these findings, there is a need for noise estimation methods which update the noise PSD estimates more frequently, to keep up with rapidly changing noise behaviour, and that are also robust enough to deal with declining SNR.

3.3 Minimum Statistics based Noise Estimation

In 2001, Martin proposed a minimum statistics based method for estimating the noise PSD by observing the noisy speech [38]. Unlike the VAD based hard decision methods, this method does not discriminate between silence and speech activity frames, and updates the estimated noise PSD on both sets of frames. A precursor for this estimation method can be found in Martin's previous work [98].

Minimum statistics methods track the minimum values of the power level in every time frame of the noisy speech PSD hence the name. This is motivated by a working observation that the short term power spectrum of a noisy speech signal often decays to the power level of the background noise as the spectral values of the noisy speech

signal fall [38]. So, tracking the minimum of a noisy speech spectrum in every time frame can help derive the PSD estimate of the background noise. The statistical independence between speech and background noise parameters is another observation which is taken into consideration while developing the formulation of the minimum statistics approach. In Martin's approach [38, 98], a smoothed periodogram of the noisy speech power spectrum is used in the minimum tracking operation. This smoothing of the noisy speech short term spectrum is achieved by applying a first order recursive operation, that is,

$$P_x(\nu, k) = \alpha P_x(\nu - 1, k) + (1 - \alpha)|X(\nu, k)|^2 \tag{3.2}$$

where $|X(\nu, k)|^2$ is the squared magnitude spectrum of the noisy speech and $P_x(\nu, k)$ denotes the recursively smoothed PSD of the noisy speech. The smoothing operation is performed in order to reduce the variance in the estimated background noise PSD obtained from minimum tracking. The value of the recursion parameter α is chosen after taking two issues into consideration. On one hand, a value of α close to one will tend to reduce the variance of the noise PSD estimate. On the other hand, a lower α close to zero means that the smoothed periodogram can track non-stationary behaviour in background noise with ease. Usually, a compromise value is chosen after considering the tradeoff between these two criteria. While in earlier work [98], the recursion parameter α is fixed (i.e. remains constant over time and frequency characteristics), in the case of signals with highly non-stationary behaviour and dynamic range, a constant α cannot satisfy the aforementioned requirements. In [38], this issue gets bypassed with the use of a of time and frequency dependent recursion parameter $\alpha(\nu, k)$, which can adapt itself based on local time frame and frequency conditions, Hence, (3.2) is replaced by the following recursion,

$$P_x(\nu, k) = \alpha(\nu, k) P_x(\nu - 1, k) + (1 - \alpha(\nu, k))|X(\nu, k)|^2 \tag{3.3}$$

An optimal value of $\alpha(\nu, k)$ is obtained by minimizing the mean square error between the smoothed periodogram $P_x(\nu, k)$ and the instantaneous noise spectrum (periodogram) during speech "pause" (silence) periods, formulated as,

$$\mathcal{E} = \mathrm{E}\big[(P_x(\nu, k) - |D(\nu, k))|^2 \;\big|\; P_x(\nu - 1, k)\big] \tag{3.4}$$

Based on (3.4) and (3.3), the optimal recursion value of $\alpha(\nu, k)$ can be written as,

$$\alpha_{opt}(\nu, k) = \frac{1}{1 + (P_x(\nu - 1, k)/|D(\nu, k)|^2 - 1)^2} \tag{3.5}$$

In practical implementation, however, the instantaneous value of the squared magnitude spectrum of the background noise in the ν^{th} frame $|D(\nu, k)|^2$ is not available. This value is substituted with the best available approximation, that is, the estimated noise spectra $|\hat{D}(\nu - 1, k)|^2$ from the previous frame. The framewise minimum tracking operation on $P_x(\nu, k)$ causes an inherent delay in the estimated noise PSD. This lag, along with the approximation being made in (3.5), can lead to significant deviations while estimating the optimal recursion parameter. A monitoring parameter is introduced to contain these deviations which result from the tracking errors. This monitoring parameter compares the frequency averaged short-term PSD estimate of the previous frame, $(1/N)\sum_{k=0}^{N-1} P_x(\nu-1, k)$ to the averaged periodogram of the current frame $(1/N)\sum_{k=0}^{N-1} |X(\nu, k)|^2$, as given by,

$$\tilde{\alpha}_c(\nu) = \frac{1}{1 + (\sum_{k=0}^{N-1} P_x(\nu - 1, k)/\sum_{k=0}^{N-1} |X(\nu, k)|^2 - 1)^2} \tag{3.6}$$

An upper limit $\alpha_{\max} = 0.96$ is also introduced for the optimal smoothing parameter to further improve the performance and reduce the tracking errors, finally leading to,

$$\hat{\alpha}_{opt}(\nu, k) = \frac{\alpha_{\max}\tilde{\alpha}_c(\nu)}{1 + (P_x(\nu - 1, k)/|\hat{D}(\nu - 1, k)|^2 - 1)^2} \tag{3.7}$$

This recursion parameter $\hat{\alpha}_{opt}(\nu, k)$ is suboptimal when compared to the optimal one, i.e., $\alpha_{opt}(\nu, k)$ in (3.5). However, on average, the deviations from the optimal parameter are have been observed to be quite small.

After obtaining $P_x(\nu, k)$ using $\hat{\alpha}_{opt}(\nu, k)$, the minimum tracking operation is performed on this periodogram. The minimum value over a finite temporal window of length L frames, is tracked and used for estimating the PSD of the background noise, that is,

$$P_{min}(\nu, k) = \min\{P_x(\lambda, k) : \nu - L < \lambda \leq \nu\} \tag{3.8}$$

The size of the tracking window L is chosen based on the following criteria [99] :

- The window period has to be longer than the broadest peaks of speech energy and

enough to accommodate at least one silence period in the noisy speech waveform.

- The window period has to be short enough to track abrupt changes in the noise levels especially in the case of non-stationary noise.

The noise estimate obtained from minimum tracking operation, $P_{min}(\nu, k)$, usually exhibit a small bias. This is due to the fact that the minimum of a set of independent random variables with non-trivial PDFs is always smaller than the mean of these variables. A time frame and frequency dependent bias factor, $B_{min}(\nu, k)$, is derived to compensate for this effect. The detailed procedure used for deriving this bias compensation factor can be found in [38]. The final unbiased estimate of background noise PSD can be expressed as,

$$\hat{P}_d(\nu, k) = B_{min}(\nu, k) P_{min}(\nu, k) \tag{3.9}$$

As mentioned earlier, this estimation approach suffers from a time lag due to the minimum tracking operation. The time lag usually lasts over the length of the tracking window i.e. L frames. This lag hinders the estimation method from tracking sudden changes in the noise level robustly, and the resultant PSD estimate will always be delayed by an excess of L frames. Later minimum statistics based approaches such as the MCRA [39] attempt to reduce this time delay in tracking abrupt changes in the noise level.

3.4 Minima Controlled Recursive Averaging(MCRA)

In 2002, Cohen and Berdugo proposed a more sophisticated noise PSD estimation method based on minimum statistics and referred to as minima controlled recursive averaging (MCRA) [39]. The MCRA estimation method is similar to the one proposed by Martin, in that the estimation is performed by tracking the minimum of a smoothed periodogram of the noisy speech obtained by recursive averaging over previous time frames. The smoothing parameter in this recursion is calculated based on a statistical hypothesis testing method which uses the *a priori* speech presence probability whose calculation is further detailed below.

The hypotheses corresponding to the presence and absence of speech in the νth frame and

kth frequency bin are formally defined as,

$$\begin{aligned} H_0(\nu,k) &: X(\nu,k) = D(\nu,k) \\ H_1(\nu,k) &: X(\nu,k) = S(\nu,k) + D(\nu,k) \end{aligned} \tag{3.10}$$

where $X(\nu,k)$, $S(\nu,k)$, and $D(\nu,k)$ denote the STFT coefficients of the noisy speech, clean speech and background noise respectively. Hypothesis $H_0(\nu,k)$ is true when the speech utterance is absent (silence period) while hypothesis $H_1(\nu,k)$ is true when both speech and noise are present in that particular time frame. Similar hypotheses are used for estimating the background noise PSD, $\hat{P}_d(\nu,k)$, that is,

$$\begin{aligned} H_0'(\nu,k) &: \hat{P}_d(\nu+1,k) = \alpha_D \hat{P}_d(\nu,k) + (1-\alpha_D)|X(\nu,k)|^2 \\ H_1'(\nu,k) &: \hat{P}_d(\nu+1,k) = \hat{P}_d(\nu,k) \end{aligned} \tag{3.11}$$

where α_D is a recursion parameter used for smoothing the noise estimate. The overall noise PSD estimation based on these test cases (3.11) can be represented as,

$$\begin{aligned} \hat{P}_d(\nu+1,k) &= \hat{P}_d(\nu,k)p'(\nu,k) + \big[\alpha_D \hat{P}_d(\nu,k) + (1-\alpha_D)|X(\nu,k)|^2\big]\big(1-p'(\nu,k)\big) \\ &= \tilde{\alpha}_D(\nu,k)\hat{P}_d(\nu,k) + \big[1-\tilde{\alpha}_D(\nu,k)\big]|X(\nu,k)|^2 \end{aligned} \tag{3.12}$$

where

$$p'(\nu,k) \triangleq \Pr\big(H_1'(\nu,k) \mid X(\nu,k)\big) \tag{3.13}$$

$$\tilde{\alpha}_D(\nu,k) \triangleq \alpha_D + (1-\alpha_D)p'(\nu,k) \tag{3.14}$$

Here, $p'(\nu,k)$ denotes the conditional speech presence probability under the observation of the STFT of the noisy speech $X(\nu,k)$ while $\tilde{\alpha}_D(\nu,k)$ denotes the time frame and frequency bin dependent smoothing parameter used for the equivalent recursive estimation. As seen from (3.14), the conditional speech presence probability is required for calculating the smoothing parameter. The speech presence probability in a given time frame and frequency bin is determined by the ratio between the local energy of the noisy speech and its minimum within a specified search window of length L frames, as in (3.8). The local energy of the noisy speech is obtained after smoothing its square magnitude STFT, $|X(\nu,k)|^2$, over both

the time and frequency domains as shown below,

$$\begin{aligned} P_x^f(\nu,k) &= \sum_{i=-w}^{w} b(i)|X(\nu,k-i)|^2 \\ P_x(\nu,k) &= \alpha_s P_x(\nu-1,k) + (1-\alpha_s)P_x^f(\nu,k) \end{aligned} \tag{3.15}$$

where $b(\cdot)$ is a windowing function of length $2w+1$ used for frequency averaging, $P_x^f(\nu,k)$ is the frequency smoothed spectrum of $|X(\nu,k)|^2$, and α_s is a recursive parameter used for the smoothing operation in the time domain. $P_x(\nu,k)$ in (3.15) is the time and frequency smoothed spectrum used to represent the local energy of the noisy speech. A minimum value of $P_x(\nu,k)$, denoted as $P_{x\min}(\nu,k)$, is obtained after analysing the time frame window of length L frames. Finally, we compute $P_{xr}(\nu,k) \triangleq P_x(\nu,k)/P_{x\min}(\nu,k)$ which denotes the ratio between the local energy of the noisy speech and its derived minimum.

A Bayesian minimum cost decision rule is designed to compute the conditional speech probability using $P_{xr} \equiv P_{xr}(\nu,k)$,

$$\frac{p(P_{xr} \mid H_1)}{p(P_{xr} \mid H_0)} \underset{H_0'}{\overset{H_1'}{\gtrless}} \frac{c_{10}\Pr(H_0)}{c_{01}\Pr(H_1)} \tag{3.16}$$

where $\Pr(H_1)$ and $\Pr(H_0)$ are the *a priori* probabilities for speech presence and absence in the noisy speech frames, respectively, c_{01} is the cost for deciding speech absence (H_0) when speech utterances are present (H_1) and c_{10} is the cost for deciding speech presence (H_1) during a silence period (H_0). The decision rule in (3.16) reduces to,

$$P_{xr}(\nu,k) \underset{H_0'}{\overset{H_1'}{\gtrless}} \delta \tag{3.17}$$

where $\delta > 0$ is an energy threshold used to determine whether the speech is present or absent in a given frame ν. The speech presence or absence in a time frame can be expressed as an identifier function $I(\nu,k)$ using (3.17),

$$I(\nu,k) = \begin{cases} 0, & \text{if } P_{xr}(\nu,k) < \delta \text{ (speech absent/silence period)} \\ 1, & \text{if } P_{xr}(\nu,k) > \delta \text{ (speech presence period)} \end{cases} \tag{3.18}$$

Using this indicator function, the speech presence probability is estimated as:

$$\hat{p}^{'}(\nu,k) = \alpha_p \hat{p}^{'}(\nu-1,k) + (1-\alpha_p) I(\nu,k) \tag{3.19}$$

where $0 < \alpha_p < 1$ is a smoothing parameter used for recursion.

Finally, having estimated the conditional speech presence probability as above, the smoothing parameter, $\tilde{\alpha}_D$, is computed and the noise PSD estimate is updated using (3.13). As pointed out above, however, because it uses a search window window of L frames, this proposed approach also has a memory in excess of L frames. Consequently, a similar problem as in [38] appears in tracking sudden changes in the noise power.

3.5 Improved Minima Controlled Recursive Averaging(IMCRA)

In 2003, Cohen proposed a noise PSD estimation method which improved upon certain aspects and shortcomings of the MCRA. This method, hence, has been referred to as IMCRA. Like MCRA, it includes averaging of past spectral values using a smoothing factor which is dependent on the speech presence probability. Compared to MCRA, improvements have been made with respect to minimum tracking, speech presence probability calculation and introducing a bias compensation factor. Similar to (3.11) in MCRA, two hypotheses $H_0(\nu,k)$and $H_1(\nu,k)$ are defined, which refer to the case of speech absence and speech presence respectively. The noise PSD estimate is updated exactly as in (3.12) and (3.14), where the conditional speech presence probability $p'(\nu,k)$ in (3.14) is now computed on the basis of a Gaussian model for the speech signal and noise component. Specifically,

$$p'(\nu,k) = \left\{ 1 + \frac{q(\nu,k)}{1+q(\nu,k)} (1+\xi_{\nu,k}) \exp(-\zeta_{\nu,k}) \right\}^{-1} \tag{3.20}$$

where

$$\zeta_{\nu,k} = \gamma_{\nu,k} \frac{\xi_{\nu,k}}{1+\xi_{\nu,k}} \qquad \xi_{\nu,k} = \frac{\sigma_S^2(\nu,k)}{\sigma_D^2(\nu,k)} \qquad \gamma_{\nu,k} = \frac{|X(\nu,k)|^2}{\sigma_D^2(\nu,k)} \tag{3.21}$$

$\gamma(\nu,k)$ and $\xi(\nu,k)$ are the *a posteriori* and *a priori* SNRs respectively and $q(\nu,k)$ is the *a priori* probability of speech absence. The estimation of $q(\nu,k)$ is controlled by the minima values of a periodogram of the noisy speech power spectrum which is smoothed in

both time and frequency domain. The procedure itself includes two iterations of smoothing and minima tracking. A detailed description of the procedure for estimating the *a priori* speech absence probability, $q(\nu, k)$, and noise PSD estimation can be found in [40].

3.5.1 Performance and Limitations of IMCRA

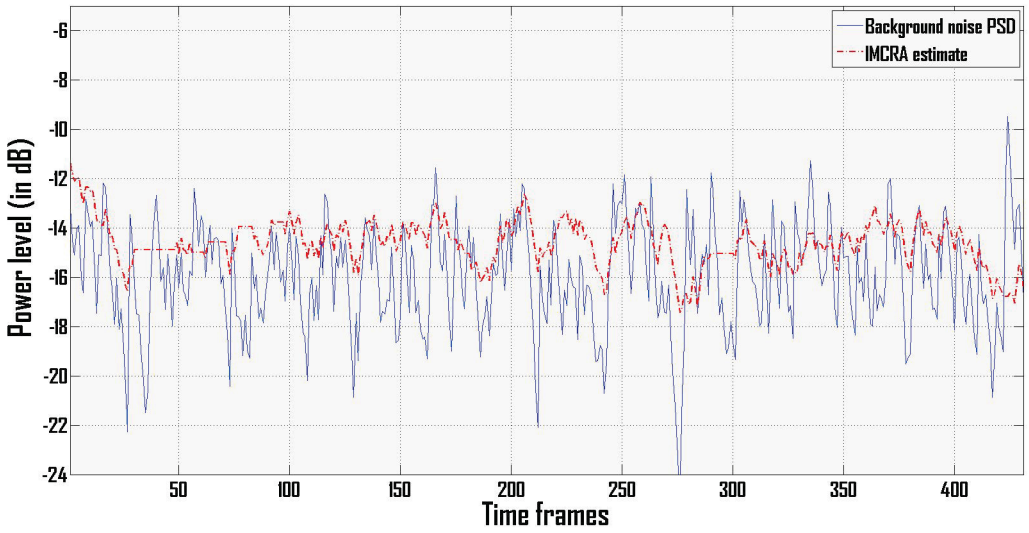
IMCRA, being a soft decision method based on minimum statistics, is able to track the background noise better than the VAD based methods which update only during the speech absence periods. Performing the smoothing and minimum tracking operations in two iterations also allows for larger smoothening windows and smaller windows for minima search. This results in a reduced variance of the minima values and shorter tracking delay in cases of rising noise power when compared to Martin's approach [38] and MCRA [39]. Further analysis of its performance in comparison to Martin's MS method can be found in [40].

The tracking performance of the IMCRA is dictated by the parameters employed for minima tracking operation such as size of search window and smoothing constant etc. These parameters also affect the variance of the estimated noise PSD during periods when speech is present. The trade-off used to reconcile between these issues means that the IMCRA can respond well to noise types with only limited non-stationary behaviour. Abrupt changes in background noise behaviour in cases such as busy street intersection with passing vehicles or a busy airport or train station with intercom announcements are often not tracked properly by the IMCRA. Speech enhancement methods using IMCRA tend to perform poorly in such cases. The noise tracking performance of the IMCRA method for stationary and non-stationary noise types can be observed in Figures 3.1 and 3.2. In Figure 3.1, the temporal dynamics of a particular frequency bin of the actual noise PSD [1] and the IMCRA estimate are plotted for cases of stationary noise types (white and pink). The temporal dynamics of the frequency bin of the actual noise PSD and the IMCRA estimate are plotted for cases of non-stationary noise types (street and babble) in Figure 3.2. The IMCRA's noise estimation is able to track variations in noise levels fairly accurately when they are small and not abrupt (Figure 3.1). IMCRA's noise tracking isn't accurate when the noise level variations are abrupt and large (Figure 3.2).

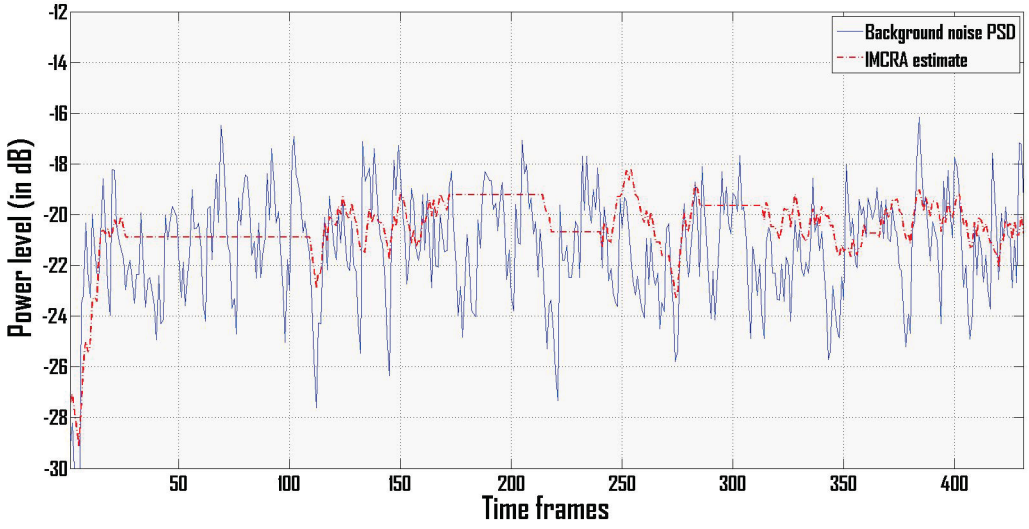
Based on the observations presented in [40], one can concur that Martin's approach [38] and MCRA [39] also face issues with tracking abrupt changes (They show inferior

[1] The "actual" PSD is obtained by performing a recursive averaging operation on consecutive frames of the squared magnitude spectra of the background noise.

performance compared to IMCRA). This limitation with regards to tracking non-stationary behaviour among conventional MS based methods has provided motivation for developing new approaches which can exploit *a priori* information of the non-stationary behaviour of various noise types to estimate the background noise more accurately. The focus of this thesis report will shift towards such estimation methods in the upcoming chapters.

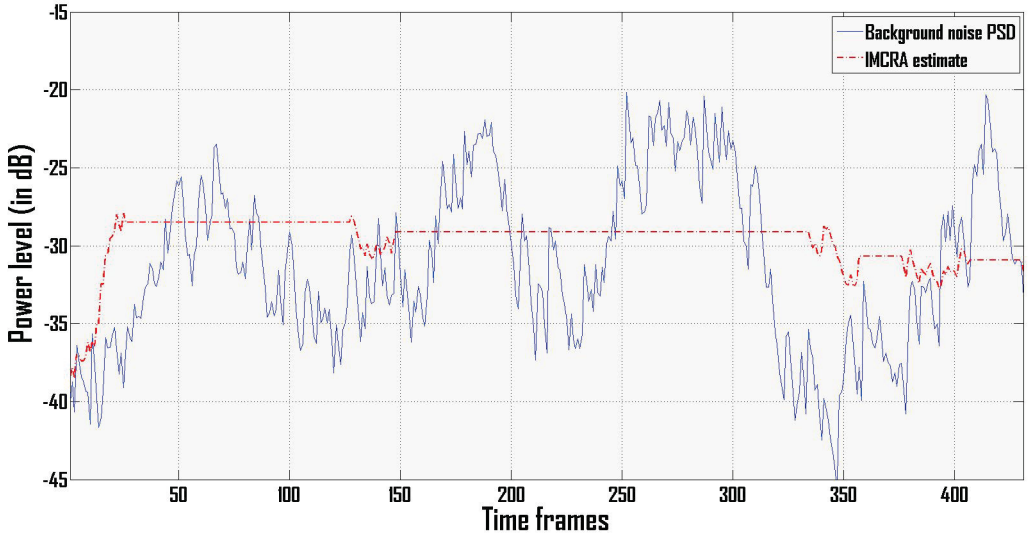


(a) White noise with 5dB SNR

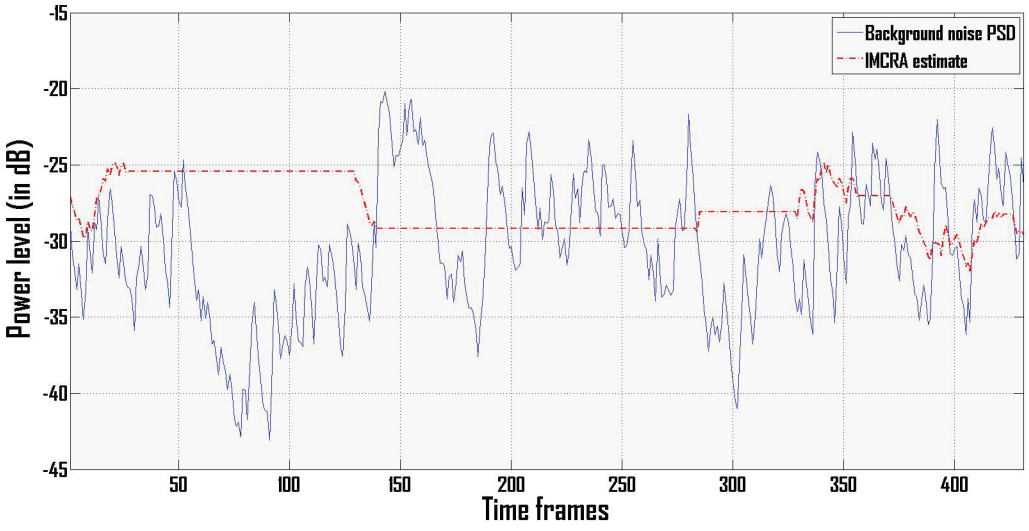


(b) Pink noise with 5dB SNR

Fig. 3.1 Noise tracking of IMCRA for noise types with limited non-stationarity. The graphs contain the temporal trajectory of a frequency bin of the actual background noise PSD present in a speech signal and its IMCRA estimate.



(a) Street noise with 5dB SNR



(b) Babble noise with 5dB SNR

Fig. 3.2 Noise tracking of IMCRA for noise types with rapidly changing non-stationary behaviour. The graphs contain the temporal trajectory of a frequency bin of the actual background noise PSD present in a speech signal and its IMCRA estimate.

Chapter 4

Codebook Assisted Estimation of STP Parameters

In this chapter, the topic of codebook assisted estimation of speech and noise short term predictor (STP) parameters is covered. In Section 4.1, the background and motivation behind codebook-based estimation approaches is presented. Section 4.2 briefly covers the concept of linear prediction and autoregressive modelling with regards to speech signals. In section 4.3, the methodology behind training of codebooks which contain a priori *information in the form of normalized LP coefficients is described. This section also provides a concise overview of the vector quantization algorithm used for creating codebooks. Section 4.4 reviews codebook-based approaches for estimating noise and speech STP parameters.*

4.1 Background

Most speech enhancement methods require an estimate of the background noise PSD to perform noise suppression. These noise PSD estimates are usually obtained via standard noise tracking methods such as the ones presented in Chapter 3. As shown in that chapter, these methods can be ineffective in cases of highly non-stationary background noise. For example, the VAD based hard decision method forsakes speech activity periods and updates its noise estimate only during silence periods [35–37]. Consequently, it cannot track changes in noise behaviour during speech periods. The soft decision based minimum statistics approaches such as Martin's method [38] and Cohen's IMCRA [39, 40] update the noise estimates by tracking the spectral minima in short time buffers of the noisy speech PSD.

Tracking minima in buffer results in a frame lag in the estimated noise spectrum. The time lag hardly affects the estimation when noise spectral characteristics change little with time. However, it can lead to highly inaccurate results when the noise behaviour rapidly changes i.e., non-stationary noise.

Codebook-based approaches attempt to overcome these drawbacks by using *a priori* information of the speech and noise signals in the form of short term predictor (STP) parameters while performing estimation of the noise and speech PSDs [45, 46]. The STP parameters usually consist of the linear predictive (LP) coefficients and their corresponding excitation variance. As a general rule, there is a constraint on the number of possible shapes of spectral envelopes of speech signals due to the physiology involved in speech production. As a result, these envelopes can be modelled by a finite set of representative spectral shapes in the form of LP coefficient vectors obtained from training on large data sets of speech and noise signals. This hypothesis that any spectral envelope of a speech (and noise) signal can be modelled based on a finite set of LP coefficients forms the basis of the codebook-based approach for speech enhancement.

Earlier codebook-based methods used *a priori* information obtained solely from the speech signals while performing speech enhancement. They relied on long term estimates of noise PSD (obtained from VAD) which usually resulted in poor performance when dealing with non-stationary background noise [101, 102]. Simultaneous estimation of speech and noise spectral envelopes on a frame-by-frame basis by codebooks containing *a priori* information for both speech and noise signals can obviate the use of long-term noise estimates while performing noise reduction. In [108], Kuropatwinski and Kleijn proposed an approach based on this simultaneous estimation principle for estimating the excitation variances of the auto-regressive (AR) spectra of speech and noise signals for in a speech coding application. Frame-by-frame estimation of noise PSD makes it possible for this method to track quickly varying noise levels in a noisy speech signal in a satisfactory manner. Speech enhancement methods proposed over the years based on this estimation approach have been reported to perform reasonable enhancement under non-stationary background noise conditions [45–47].

4.2 Spectral Model

Consider the additive noise model in (2.1), reproduced below for convenience as,

$$x[n] = s[n] + d[n] \tag{4.1}$$

where $x[n], s[n]$ and $d[n]$ are the noisy speech, clean speech and background noise signals respectively. Under the assumption of uncorrelated speech and noise signals, the PSD of the noisy speech can be represented as,

$$P_x(\omega) = P_s(\omega) + P_d(\omega), \quad \omega \in [0, 2\pi) \tag{4.2}$$

where $P_s(\omega)$ and $P_d(\omega)$ are the clean speech and background noise PSD, respectively, and ω is the normalized angular frequency.

The PSD shape of signal $y[n]$, where $y \in \{s, d\}$ stands for either the speech or noise, can be modelled in terms of its LP coefficients and corresponding excitation variance as [103],

$$P_y(\omega) = g_y \, \bar{P}_y(\omega) \tag{4.3}$$

where $\bar{P}_y(\omega)$ is the gain normalized spectral envelope and g_y is the excitation gain (or variance). In turn, the gain normalized spectral envelope is given by,

$$\bar{P}_y(\omega) = \frac{1}{|1 + \sum_{k=1}^{p} a_k^y e^{-j\omega k}|^2} \tag{4.4}$$

where $\{a_k^y\}_{k=1}^p$ are the LP coefficients, represented here by vector $\boldsymbol{\theta}_y = [a_1^y, ..., a_p^y]$, and p is the model order chosen.

The model orders for speech and noise signals are chosen based on the desired level of performance. Usually, for speech signals sampled at the rate of 16kHz, the LP coefficient order would be typically chosen in the range $10 \leq p \leq 16$. Higher values of p leads to better modelling of the higher formants in the spectral envelopes. For the noise signals, the model order is generally chosen in the range of $8 \leq p \leq 10$, which is lower than the model order chosen for speech signals.

4.3 Training of Speech and Noise Codebooks

In this thesis, two different codebooks of short-time spectral parameters, one for the speech and the other for the noise, are generated from training data comprised of multiple speaker signals and different noise types. The codebook generation comprises of the following steps:

1. Segmentation of the training speech and noise data into frames of 20-40 ms duration;
2. Computation of LP coefficients $\{a_k^y\}_{k=1}^p$ for each frame using the autocorrelation method [105];
3. Vector quantization of the LP coefficient vectors $\boldsymbol{\theta}_y$ using the LBG algorithm to obtain the required codebook [49].

The LBG algorithm, which is used in our work, forms a set of median cluster vectors which best represent the given input set of LP coefficient vectors. Optimal values have to be chosen empirically for the size of the speech and noise codebooks, considering the trade-off between PSD estimation accuracy and complexity. A small number of quantized vectors (i.e., under representation of LP data) reduces the complexity but may lead to inaccurate estimation of the underlying PSD, in the associated vector space. On the other hand, a larger number of quantized vectors (i.e., over representation of LP data) entails additional complexity and in some cases, may lead to deterioration of performance. In the sequel, we shall represent the speech and noise codebooks so obtained as $\{\boldsymbol{\theta}_s^i\}_{i=1}^{N_s}$ and $\{\boldsymbol{\theta}_d^j\}_{j=1}^{N_d}$, where vectors $\boldsymbol{\theta}_s^i$ and $\boldsymbol{\theta}_d^j$ are the corresponding i-th and j-th codebook entries, and N_s and N_d are the codebook sizes, respectively.

In addition to the codebook vectors generated as above from training on noise data, during the estimation phase, the noise codebook is supplemented by one extra vector. The latter is updated for every frame and contains the LP coefficients of a noise PSD estimate obtained using a conventional MS method [38, 40]. This provides robustness in dealing with noise types which may not have been present in the training set.

4.3.1 The LBG Vectorization Algorithm

This subsection briefly describes the vector quantization method used in this thesis to generate codebooks, based on [49, 104]. Speech and noise codebooks consist of representative vectors of the LP coefficients obtained from speech and noise samples. These codebooks are

generated using a vector quantization (VQ) method known as Generalized Llyod algorithm (GLA) [49]. A vector quantizer of dimension p is usually described as a mapping from the p dimensional data vectors defined in Euclidean space, $\mathbb{R}^p$, to a finite subset $\mathcal{C} \subset \mathbb{R}^p$ which contains N vectors called code vectors. The set $\mathcal{C}$ of all code vectors is referred to as the codebook while N is the codebook size. GLA is one of the most commonly used VQ algorithms for codebook generation. It was first introduced in [49] and is also commonly referred as LBG (as a reference to the initials of the authors who proposed this method).

LBG is an iterative clustering algorithm which strives to produce an optimal codebook for a given data source based on the specified configuration. This algorithm is very similar to the well known K-means method used for vector quantization in data mining [106]. When applying the LBG algorithm, the training data vectors are partitioned into several groups/clusters, each with its own representative centroid, as dictated by the required final codebook size. This centroid vector is used as a representative of that particular group and is also referred to as a code vector. The main goal of LBG algorithm is to minimize the distortion measure between the training vectors and their representative code vectors, i.e., finding the optimal partition and code vectors to minimize the overall distortion.

To achieve this, LBG employs an iterative procedure which is repeated until the average distortion measure calculated over all of the training data vectors is minimized. The main steps of this process can be summarized as follows:

1. An initial codebook containing N code vectors, $\mathcal{C}^0 = \{\mathbf{c}_j \in \mathbb{R}^p | j = 1, 2, ..., N\}$, is first chosen (as explained later);

2. The set of training data vectors of size K ($K \gg N$), $\mathcal{S} = \{\mathbf{s}_i \in \mathbb{R}^p | i = 1, 2..., K\}$, is partitioned into N groups/clusters, each cluster being represented by one of the initial set of code vectors. The partition is performed by using a pre-selected distortion measure (square distance in most cases):

$$\mathbf{s}_i \in \mathcal{P}_{\mathbf{c}_q} \quad \text{if} \quad ||\mathbf{s}_i - \mathbf{c}_q||_2 \leq ||\mathbf{s}_i - \mathbf{c}_j||_2 \quad \forall j \neq q$$

   where $\mathcal{P}_{\mathbf{c}_q}$ is the cluster corresponding to the code vector $\mathbf{c}_q$ and for a vector $\mathbf{a}$, $||\mathbf{a}||_2 = \mathbf{a}^T\mathbf{a}$ is the Euclidean norm, which is used in this thesis as the distortion measure.

3. An average distortion measure known as D^{init} using this initial set of vectors is cal-

culated as follows,

$$D^{\text{init}} = \frac{1}{N}\sum_{j=1}^{N}\sum_{\mathbf{s}_i \in \mathcal{P}_{\mathbf{c}_j}} ||\mathbf{s}_i - \mathbf{c}_j||_2 \tag{4.5}$$

4. For each group, a centroid vector is calculated to obtain an improved iteration of the code vector,

$$\mathbf{c}_j = \frac{1}{S_j}\sum_{\mathbf{s}_i \in \mathcal{P}_{\mathbf{c}_j}} \mathbf{s}_i \tag{4.6}$$

5. A new distortion measure known as D^{upd} is calculated using these updated code vectors as follows,

$$D^{\text{upd}} = \frac{1}{N}\sum_{j=1}^{N}\sum_{\mathbf{s}_i \in \mathcal{P}_{\mathbf{c}_j}} ||\mathbf{s}_i - \mathbf{c}_j||_2 \tag{4.7}$$

6. If $(D^{\text{init}} - D^{\text{upd}})/D^{\text{upd}} \leq \epsilon$, where ϵ is the pre-selected minimum threshold value, the updated set of code vectors, $\mathbf{c}_j \ \forall \ j \in \{1, 2, ..., N\}$, is chosen as the codebook for representing the training vectors. If the condition is not satisfied, Steps 2 to 5 are repeated with the new updated vectors as the initial set;

Classification algorithms such as vector quantization are considered as NP hard problems in most cases. The convergence mentioned in Step 6 may take a long time to be achieved. To limit this computational burden, the LBG algorithm is usually constrained to perform only a fixed number of iterations. An example of a clustering obtained through a vector quantization algorithm for a $p = 3$ dimensional data set is shown in Fig. 4.1.

The initial codebook plays an important role in the speed of convergence of the algorithm as well as the efficiency of the codebook's representation of the training data set. Numerous approaches have been proposed in order to generate the initial codebook in Step 1. The method originally suggested as a part of the LBG algorithm [49] is known as binary splitting. In this method, an initial code vector is obtained as the average of the entire training sequence as shown below,

$$\mathbf{c}^0 = \frac{1}{K}\sum_{\mathbf{s}_i \in \mathcal{S}} \mathbf{s}_i \tag{4.8}$$

This code vector is then split into two vectors, each with an added perturbation to the

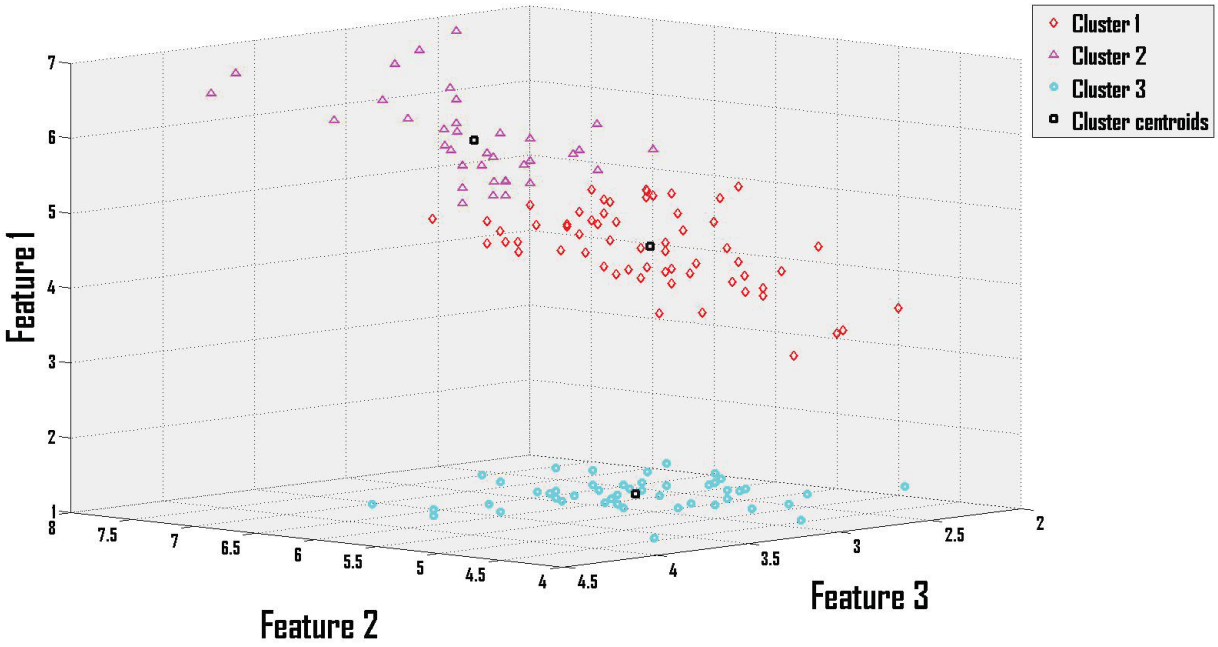


Fig. 4.1 An example of a clustering method similar to the one used in the LBG algorithm. This particular data clusterization is implemented on a 3 dimensional data (three features) obtained from a botanical database known as Fisher's iris data (Source *Mathworks*).

initial average vector as shown below

$$\begin{aligned} \mathbf{c}_1^1 &= (1+\delta)\mathbf{c}^0 \\ \mathbf{c}_2^1 &= (1-\delta)\mathbf{c}^0 \end{aligned} \tag{4.9}$$

where δ is the small perturbation. Code vectors $\mathbf{c}_1^1$ and $\mathbf{c}_2^1$ are then used as the initial codebook vectors to perform the LBG iterative clustering to obtain a codebook with two vectors, $\mathcal{C}_2 = \{\mathbf{c}_j \in \mathbb{R}^p | j = 1, 2\}$. The two vectors from this codebook are then split into four vectors (two from each) as in (4.9). These four vectors are then used as initial codebook vectors to train a codebook containing four vectors, $\mathcal{C}_4 = \{\mathbf{c}_j \in \mathbb{R}^p | j = 1, 2, 3, 4\}$. This process is repeated until a codebook with the desired number of code vectors, $\mathcal{C}_N = \{\mathbf{c}_j \in \mathbb{R}^p | j = 1, 2, ..., N\}$, is obtained. Other commonly used approaches to obtain the initial set of code vectors are the random selection method and the pairwise nearest neighbour (PNN) method [107].

4.4 STP Estimation of Speech and Noise Spectra

Each codebook entry, i.e., $\boldsymbol{\theta}_s^i$ or $\boldsymbol{\theta}_d^j$, can be used to compute a corresponding gain normalized spectral envelope, respectively $\bar{P}_s^i(\omega)$ or $\bar{P}_d^j(\omega)$ by means of relation (4.4). To obtain the final PSD shape as in (4.3), however, the resulting envelope needs to be scaled by a corresponding excitation gain, which we denote as g_s^i and g_d^j, respectively. In this work, we use an adaptive approach whereby the excitation gains for the speech and noise codebooks are updated every frame based on the observed noisy speech magnitudes $|X(\nu, k)|$.

Specifically, for every possible combination of vectors $\boldsymbol{\theta}_s^i$ and $\boldsymbol{\theta}_d^j$ from the speech and noise codebooks, the corresponding gains g_s^i and g_d^j at the ν-th frame are obtained by maximizing the likelihood function obtained from the noisy speech probability. The final optimum values of g_s^i and g_d^j, which can be interpreted as conditional maximum likelihood (ML) estimates, are approximated as in [45, 108]. Specifically, the ML estimates of the speech and noise excitation variances for the codebook combination $[\boldsymbol{\theta}_s^i, \boldsymbol{\theta}_d^j]$ are obtained as,

$$\left\{g_s^i, g_d^j\right\}_{\mathrm{ML}} = \arg\max_{g_s, g_d} p(\boldsymbol{x}|\boldsymbol{\theta}_s^i, \boldsymbol{\theta}_d^j; g_s, g_d) \tag{4.10}$$

where $\boldsymbol{x} = [x[\nu F + 1], \ldots, x[\nu F + N]]^T$ is the observed data vector at the ν-th frame, $p(\boldsymbol{x}|\boldsymbol{\theta}_s^i, \boldsymbol{\theta}_d^j; g_s, g_d)$ is the conditional probability density function of the noisy speech frame $\boldsymbol{x}$ for a codebook combination of $\boldsymbol{\theta}_s^i, \boldsymbol{\theta}_d^j$ with g_s and g_d as the speech and noise excitation variances.

The noisy speech signal vector $\boldsymbol{x}$ is assumed to follow a multivariate zero mean normal distribution. Based on this assumption, the likelihood function in (4.10) can be expressed as,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2} \det(\boldsymbol{R}_{xx})^{1/2}} e^{-(1/2)(\boldsymbol{x}^T \boldsymbol{R}_{xx}^{-1} \boldsymbol{x})} \tag{4.11}$$

where $\boldsymbol{\theta}$ is the parameter vector consisting of a speech and noise codebook pair and their corresponding variances, $\boldsymbol{\theta} = [\boldsymbol{\theta}_s^i, \boldsymbol{\theta}_d^j; g_s, g_d]$, and $\boldsymbol{R}_{xx} = E\{\boldsymbol{x}\boldsymbol{x}^T\}$ is the associated covariance matrix. Under the previous modelling assumptions, the covariance matrix can be written as the sum of the speech and noise covariance matrices, i.e., $\boldsymbol{R}_{xx} = \boldsymbol{R}_{ss} + \boldsymbol{R}_{dd}$, due to lack of correlation between speech and noise signals. From (4.11), the log likelihood

function (LLF) can be computed as,

$$l(g_s, g_d) = \ln p(\boldsymbol{x}|\boldsymbol{\theta}) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|\det(\boldsymbol{R}_{xx})| - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{R}_{xx}^{-1}\boldsymbol{x} \tag{4.12}$$

The equation for the LLF in (4.12) involves the inversion of $N \times N$ matrix $\boldsymbol{R}_{xx}$, which will, in general, significantly increase the computational complexity of the processing. If the length of the speech frames are large enough (i.e, in the range of 32ms to 40ms, the covariance matrices $\boldsymbol{R}_{ss}$ and $\boldsymbol{R}_{dd}$ can be assumed to be circulant and, as such, can be diagonalized by the Fourier transform [45, 109]. The LLF in (4.12) then reduces to,

$$\begin{aligned} l(g_s, g_d) \approx -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{k=0}^{N-1}\ln(g_s\bar{P}_s(\omega_k) + g_d\bar{P}_d(\omega_k)) \\ - \frac{1}{2}\sum_{k=0}^{N-1}\frac{|X(\nu,\omega_k)|^2}{g_s\bar{P}_s(\omega_k) + g_d\bar{P}_d(\omega_k)} \end{aligned} \tag{4.13}$$

where $\omega_k = \frac{2\pi k}{N}$, $\bar{P}_s(\omega_k)$ and $\bar{P}_d(\omega_k)$ are the gain normalized LP spectral envelopes of speech and noise signal respectively, as given by (4.4).

The ML estimates of the excitation variances $[g_s, g_d]$ can be obtained by maximizing the LLF in (4.13) with respect to g_s and g_d respectively. Maximization of the LLF can be interpreted as the minimization of a corresponding Itakura-Saito distortion measure between the observed squared magnitude spectrum and the estimated spectral envelope of noisy speech using the codebook vectors $\boldsymbol{\theta}_s^i$ and $\boldsymbol{\theta}_d^j$ [45, 110], that is,

$$\begin{aligned} \left\{g_s^i, g_d^j\right\}_{\mathrm{ML}} &= \arg\max_{g_s,g_d} l(g_s, g_d) \\ &= \arg\min_{g_s,g_d} d_{\mathrm{IS}}(|X(\nu,\omega_k)|^2, P_x^{ij}(\omega_k)) \end{aligned} \tag{4.14}$$

where the spectral envelope $P_x^{ij}(\omega)$ is given by,

$$P_x^{ij}(\omega) = g_s\bar{P}_s^i(\omega) + g_d\bar{P}_d^j(\omega) \tag{4.15}$$

and the Itakura-Saito distortion measure in (4.14) is given by,

$$d_{\mathrm{IS}}(|X(\nu,\omega_k)|^2, P_x^{ij}(\omega_k)) = \sum_{k=0}^{N-1} \left(\frac{|X(\nu,\omega_k)|^2}{P_x^{ij}(\omega_k)} - \ln \left(\frac{|X(\nu,\omega_k)|^2}{P_x^{ij}(\omega_k)} \right) - 1 \right) \tag{4.16}$$

Equation (4.16) can be further simplified by performing a Taylor series expansion of the $\ln(\cdot)$ term under the assumption that the modelling error between the observed spectrum $|X(\nu,\omega_k)|^2$ and the estimated spectrum $P_x^{ij}(\omega_k)$ is small. The Itakuro-Saito distortion measure then reduces to [45],

$$d_{\mathrm{IS}}(|X(\nu,\omega_k)|^2, P_x^{ij}(\omega_k)) \approx \frac{1}{2} d_{\mathrm{LS}}(|X(\nu,\omega_k)|^2, P_x^{ij}(\omega_k)) \tag{4.17}$$

where $d_{\mathrm{LS}}(|X(\nu,\omega_k)|^2, P_x^{ij}(\omega_k))$ is the log spectral distortion measure given by,

$$d_{\mathrm{LS}}(|X(\nu,\omega_k)|^2, P_x^{ij}(\omega_k)) = \sum_{k=0}^{N-1} \left| \ln \frac{P_x^{ij}(\omega_k)}{|X(\nu,\omega_k)|^2} \right|^2 \tag{4.18}$$

In [45], the excitation variances are determined by performing partial differentiation of the distortion measure in (4.18) with respect to g_s and g_n, setting the result to zero and solving the resulting set of simultaneous equations. The solutions for these simultaneous equations can be represented in a matrix form as given by,

$$\mathbf{C} \begin{bmatrix} g_s^i \\ g_d^j \end{bmatrix} = \mathbf{D} \tag{4.19}$$

where

$$\begin{aligned} \mathbf{C} &= \begin{bmatrix} \left|\left| \frac{\bar{P}_s^i(\omega_k)^2}{|X(\nu,\omega_k)|^4} \right|\right| & \left|\left| \frac{\bar{P}_d^j(\omega_k)\bar{P}_s^i(\omega_k)}{|X(\nu,\omega_k)|^4} \right|\right| \\ \left|\left| \frac{\bar{P}_d^j(\omega_k)\bar{P}_s^i(\omega_k)}{|X(\nu,\omega_k)|^4} \right|\right| & \left|\left| \frac{\bar{P}_d^j(\omega_k)^2}{|X(\nu,\omega_k)|^4} \right|\right| \end{bmatrix} \\ \mathbf{D} &= \begin{bmatrix} \left|\left| \frac{\bar{P}_s^i(\omega_k)}{|X(\nu,\omega_k)|^2} \right|\right| \\ \left|\left| \frac{\bar{P}_d^j(\omega_k)}{|X(\nu,\omega_k)|^2} \right|\right| \end{bmatrix} \end{aligned} \tag{4.20}$$

and $||f(\omega_k)|| = \sum_{k=0}^{N-1} |f(\omega_k)|$.

With the help of the estimated excitation gains at the ν-th frame, for each pair of

speech and noise codebook vectors $\boldsymbol{\theta}_s^i$ and $\boldsymbol{\theta}_d^j$, we can define a complete codebook-based parameter vector $\boldsymbol{\theta}^{ij} = [\boldsymbol{\theta}_s^i, \boldsymbol{\theta}_d^j, g_s^i, g_d^j]$. The unknown parameter vector $\boldsymbol{\theta}$ which best matches the noisy speech spectrum can be estimated through various schemes. In their earlier work [45], Srinivasan and Kleijn first compute the joint spectra $P_x^{ij}(\omega)$, as per (4.15), for every possible vector combination from the speech and noise codebooks using the excitation variances obtained from (4.19). The log-likelihood score is then obtained based on this value of $P_x^{ij}(\omega)$ using (4.12). A search operation is done over the computed log-likelihood scores to find the $\boldsymbol{\theta}^{ij} = [\boldsymbol{\theta}_s^i, \boldsymbol{\theta}_d^j, g_s^i, g_d^j]$ which has the highest value. This estimate of $\boldsymbol{\theta}$ is referred to as the ML estimate of the parameter vector,

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}^{ij}} p(\boldsymbol{x}|\boldsymbol{\theta}^{ij}) \tag{4.21}$$

The flow chart for the ML based estimation scheme is presented in Figure 4.2.

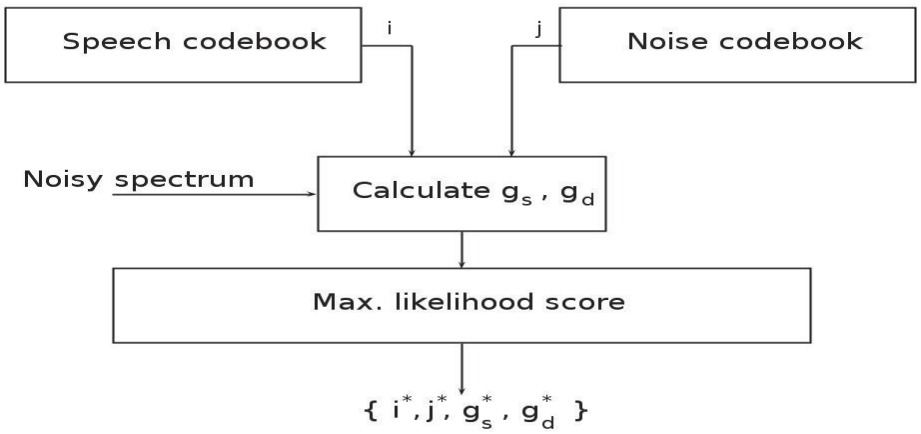


Fig. 4.2 Flow chart for a ML based STP parameter estimation. i^*, j^* form the codebook vector combination which the highest likelihood score with g_s^* and g_d^* as the corresponding excitation variances.

In the ML estimation scheme, the parameter vector $\boldsymbol{\theta}$ is treated as a deterministic parameter while performing its estimation. In later work [46], $\boldsymbol{\theta}$ is treated as a random variable with uniform distribution in order to perform a Bayesian MMSE estimation of its value. In this case, it is shown that the joint estimation can be carried out by using

numerical integration over the range of both speech and noise codebook entries, i.e.,

$$\hat{\boldsymbol{\theta}}_{\text{MMSE}} = \frac{1}{N_s N_d} \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} \boldsymbol{\theta}^{ij} \frac{p(\boldsymbol{x}|\boldsymbol{\theta}^{ij})}{p(\boldsymbol{x})} \tag{4.22}$$

where

$$p(\boldsymbol{x}) = \frac{1}{N_s N_d} \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} p(\boldsymbol{x}|\boldsymbol{\theta}^{ij}). \tag{4.23}$$

These equations provide a fair approximation to the joint MMSE estimate under the assumptions that the codebook size is sufficiently large and the unknown parameter vector $\boldsymbol{\theta}$ is uniformly distributed [46]. In this work, we have used the MMSE approach for estimating $\boldsymbol{\theta}$.

Following the MMSE estimation of the parameter vector $\boldsymbol{\theta}$ based on (4.22), the desired PSDs of clean speech and noise for every ν-th frame can be obtained by employing the AR spectral model equations, as shown below,

$$\begin{aligned} \hat{P}_s^{\text{CB}}(\omega) &= g_s \bar{P}_s(\omega) \\ \hat{P}_d^{\text{CB}}(\omega) &= g_d \bar{P}_d(\omega) \end{aligned} \tag{4.24}$$

with

$$\begin{aligned} \bar{P}_s(\omega) &= \frac{1}{|1 + \sum_{k=1}^{p} a_k^s e^{-j\omega k}|^2} \\ \bar{P}_d(\omega) &= \frac{1}{|1 + \sum_{k=1}^{q} a_k^d e^{-j\omega k}|^2} \end{aligned} \tag{4.25}$$

where $\{a_k^s\}_{k=1}^p$, $\{a_k^d\}_{k=1}^q$ are the LP coefficient vectors of speech and noise signals obtained from the estimated parameter vector $\hat{\boldsymbol{\theta}}_{\text{MMSE}}$, with g_s and g_d being the corresponding excitation gains. The overall speech and noise short term power spectral estimates, $\hat{P}_s^{\text{CB}}(\nu, k)$ and $\hat{P}_d^{\text{CB}}(\nu, k)$, can be obtained by performing this joint codebook estimation for every time frame ν. These estimates can be later utilized in a speech enhancement system to perform noise reduction. Further discussion on this topic in presented in Chapter 5.

Chapter 5

Speech Enhancement with Codebook Estimated Parameters

The focus of this chapter will be on speech enhancement methods which employ codebook-based approach for estimating the background noise spectra. In Section 5.1, a Wiener filter designed using the codebook estimates of speech and noise spectral envelopes is presented. Then the discussion shifts towards the incorporation of the codebook-based speech and noise estimates with a modulation domain based MMSE STSA speech enhancement algorithm which was developed over the course of this thesis research. As mentioned earlier, this new method is referred to as CB-MME.

5.1 Codebook Assisted Wiener Filtering

Wiener filters belong to a class of linear estimators which are used to minimize the mean square error between the clean speech and estimated speech spectrum.

$$\mathcal{E} = \mathrm{E}[(|S(\nu,k)| - |\hat{S}(\nu,k)|)^2] \tag{5.1}$$

where $\mathcal{E}$ is the mean square error between the clean speech magnitude spectrum $|S(\nu,k)|$ and the estimated speech magnitude spectrum $|\hat{S}(\nu,k)|$. Ideally, the gain function of a Wiener filter in frequency domain for a ν^{th} frame is denoted by,

$$H^{\text{ideal}}(\omega_k) = \frac{P_s(\omega_k)}{P_s(\omega_k) + P_d(\omega_k)} \tag{5.2}$$

where $P_s(\omega_k)$ and $P_d(\omega_k)$ are the power spectral densities of the clean speech and noise signals in the ν^{th} time frame and ω_k is the angular frequency denoted by $\omega_k = \frac{2\pi k}{N}$. Since the exact estimates of clean speech and background noise PSDs are not observable, several adaptive schemes have been proposed over the years to estimate an approximate the gain function which can perform reasonable noise reduction while keeping the resultant distortion within a tolerable limit [14, 15, 101].

In [45], a Wiener filter is constructed using the codebook-based PSD estimates of speech and noise. The gain function of such a filter is given by,

$$\hat{H}^{\text{CB}}(\omega_k) = \frac{\hat{P}_s^{\text{CB}}(\omega_k)}{\hat{P}_s^{\text{CB}}(\omega_k) + \hat{P}_d^{\text{CB}}(\omega_k)} \tag{5.3}$$

where $\hat{P}_s^{\text{CB}}(\omega_k)$ and $\hat{P}_d^{\text{CB}}(\omega_k)$ are the codebook-based estimates of the speech and noise PSDs in the ν^{th} time frame. We can perform a recursive smoothing operation on the gain function along its time frames ν to reduce its variance. This helps suppress the resultant distortion in the enhanced signal. The recursion parameter is adjusted empirically based on the performance.

5.2 Codebook-based STSA Estimation in Modulation Domain (CB-MME)

The merits of performing speech processing operations in the modulation frequency domain were already discussed in detail in Chapter 2. Several proposed modulation domain based enhancement schemes such as the Filter bank techniques, spectral subtraction [27] and MMSE-STSA [28] were also presented. However, to the best of our knowledge, modulation domain based methods which employ codebook estimates of noise (and speech) spectra are yet to be implemented. As a part of this thesis, efforts were undertaken to implement a modified MMSE-STSA algorithm in modulation domain which utilizes the codebook-based estimates of speech and noise PSDs to calculate its enhancement gain function.

As seen in Chapter 2, the MME method [28] is an extension of the widely used acoustic domain based MMSE spectral amplitude estimator [9], into the modulation domain. In the MME method, the clean speech modulation magnitude spectrum is estimated from the noisy speech by minimizing the mean square error, denoted as $\mathcal{E}$, between the clean and

estimated speech, i.e.,

$$\mathcal{E} = \mathrm{E}[(|S(t,k,m)| - |\hat{S}(t,k,m)|)^2] \tag{5.4}$$

where $|S(t,k,m)|$and$|\hat{S}(t,k,m)|$ denote the modulation magnitude spectra of the clean and estimated speech, respectively. t, k, and m denote the modulation time frame index, acoustic frequency bin and modulation frequency bin respectively. Using this MMSE criterion, the modulation magnitude spectrum of the clean speech can be estimated from the noisy speech as,

$$|\hat{S}(t,k,m)| = G(t,k,m)|Z(t,k,m)| \tag{5.5}$$

where $G(t,k,m)$ is the MME spectral gain function and $Z(t,k,m)$ is the modulation spectrum of the noisy speech obtained from (2.27). The MME gain function is given by [28],

$$G(t,k,m) = \frac{\sqrt{\pi\zeta}}{2\gamma} \exp\left(\frac{-\zeta}{2}\right) \left[(1+\zeta) I_0\left(\frac{-\zeta}{2}\right) + \nu I_1\left(\frac{-\zeta}{2}\right)\right] \tag{5.6}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified bessel functions of order zero and one, respectively, and the parameter $\zeta \equiv \zeta(t,k,m)$ is defined in terms of the *a priori* and *a posteriori* SNRs ξ and γ.

$$\zeta = \frac{\xi}{1+\xi}\gamma \tag{5.7}$$

The *a priori* SNR of a speech signal in modulation domain is defined as the ratio between the clean speech modulation power spectrum and the modulation power spectrum of the background noise present in the speech file, as given by,

$$\xi(t,k,m) \triangleq \frac{E[|C(t,k,m)|^2]}{E[|D(t,k,m)|^2]} \tag{5.8}$$

where $|C(t,k,m)|^2$, $|D(t,k,m)|^2$ are the squared magnitude modulation spectrum of clean speech and noise signals respectively, and $E[\cdot]$ is the expectation function. The *a posteriori* SNR of a speech signal in modulation domain is defined as the ratio between the noisy speech modulation spectrum ($|Z(t,k,m)|^2$) and the modulation power spectrum of the background noise present in the speech file, as given by,

$$\gamma(t,k,m) \triangleq \frac{|Z(t,k,m)|^2}{E[|D(t,k,m)|^2]} \tag{5.9}$$

Since the exact values of clean speech modulation spectrum and the background noise modulation spectrum are not available unlike the modulation spectrum of noisy speech which is observable, estimation schemes are required to calculate the above mentioned SNR parameters. It is precisely in the calculation of these SNR parameters that we make use of the codebook-based PSD estimates. In this work, the *a posteriori* SNR is estimated as,

$$\hat{\gamma}(t,k,m) = \frac{|Z(t,k,m)|^2}{|\hat{D}(t,k,m)|^2} \tag{5.10}$$

where $|\hat{D}(t,k,m)|^2$ is an estimate of the noise spectrum in the modulation domain. This quantity is obtained by applying the STFT (over frame index ν) to the square-root of the codebook-based noise PSD estimate, and then squaring the result. Specifically, we compute

$$\hat{D}(t,k,m) = \sum_{\nu=-\infty}^{\infty} \sqrt{\hat{P}_d^{\mathrm{CB}}(\nu,k)} w_M(tF_M - \nu) e^{-2j\nu m\pi/M} \tag{5.11}$$

where $\hat{P}_d^{\mathrm{CB}}(\nu,k)$ is the noise PSD estimate obtained at the ν-th frame through codebook-based MMSE estimation.

To reduce spectral distortion the following "decision directed" approach is employed to obtain the value of the *a priori* SNR,

$$\hat{\xi}(t,k,m) = \alpha \frac{|\hat{S}(t-1,k,m)|^2}{|\hat{D}(t-1,k,m)|^2} + (1-\alpha)\frac{|C(t,k,m)|^2}{|\hat{D}(t,k,m)|^2} \tag{5.12}$$

where $|C(t,k,m)|$ is an estimate of the clean speech spectrum in the modulation domain and $0 < \alpha < 1$ is a control factor which acts as a trade-off between noise reduction and speech distortion. Similar to (5.11), $C(t,k,m)$ is obtained by applying the STFT to the square-root of the codebook-based PSD estimate of the clean speech at the ν-th frame.

$$\hat{C}(t,k,m) = \sum_{\nu=-\infty}^{\infty} \sqrt{\hat{P}_s^{\mathrm{CB}}(\nu,k)} w_M(tF_M - \nu) e^{-2j\nu m\pi/M} \tag{5.13}$$

where $\hat{P}_s^{\mathrm{CB}}(\nu,k)$ is the clean speech PSD estimate obtained at the ν-th frame through codebook-based MMSE estimation.

The estimated modulation magnitude spectrum, $|\hat{S}(t,k,m)|$ in (5.5), is transformed to

the acoustic frequency domain by applying inverse STFT followed by OLA synthesis. The resulting spectrum is combined with the phase spectrum of the noisy speech to obtain the enhanced speech spectrum. The latter is mapped back to the time by performing inverse STFT followed by OLA synthesis to obtain the enhanced speech signal. Detailed Evaluation of the performance of this CB-MME scheme is provided in the following chapter. The proposed method is subjected to standard objective evaluations such PESQ and SegSNR and compared with other standard methods studied in this thesis such as MMSE-STSA, MME and Codebook-based Wiener filtering etc.

Chapter 6

Experiments and Results

In this chapter we describe the experiments conducted and the results obtained over the course of this thesis research. In Section 6.1, we begin by providing a brief description of the methodology of the codebook-based spectral estimation and modulation domain processing for speech enhancement. In Section 6.2, we present and discuss some of the experiments performed to analyse various aspects of the codebook-based estimation approach. Section 6.3 focuses on the experimental evaluation of the CB-MME *method proposed in this thesis. This includes comparisons to some standard speech enhancement methods such as MMSE-STSA, MME and codebook-based Wiener filter.*

6.1 Methodology

Speech utterances of two male and two female speakers from the TSP [111] and TIMIT [112] databases are used as clean speech signals for conducting the experiments in this thesis. Noise types derived from the Noisex-92 [113] and Sound Jay [114] databases, including babble, street, and restaurant noise, are used as the background noise samples in our experiments. In addition to these noise types, a non-stationary white Gaussian noise is also considered, which is generated by modulating the amplitude of a standard white Gaussian noise sequence with a specific function as follows,

$$w_{ns}(n) = p_T(n) w_s(n) \tag{6.1}$$

where $w_s(n)$ is the standard white Gaussian noise signal, $p_T(n)$ is the amplitude modulating waveform obtained by concatenating a rectangular pulse train and a sine wave signal, as illustrated in Figure 6.1, and $w_{ns}(n)$ is the resulting amplitude modulated non-stationary white noise.

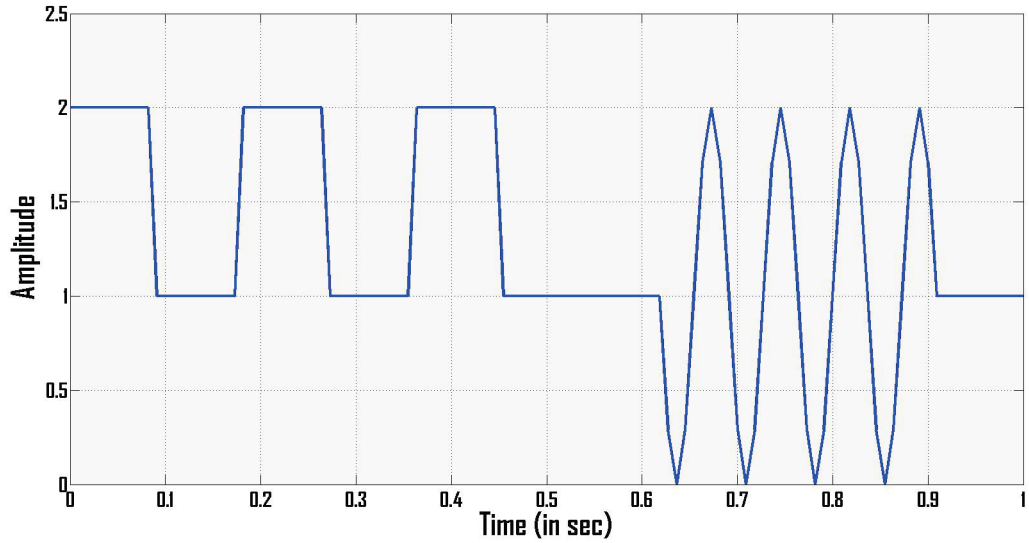


Fig. 6.1 Plot of the modulating waveform signal, $p_T(n)$, used to create the non-stationary white noise signal in (6.1). It is a combination of a rectangular pulse train followed by a sinusoidal signal.

All of the speech and noise files are uniformly sampled at a rate of 16kHz using the function `resample` in MATLAB [115]. The model order for the LP coefficients, i.e, p in (4.4), is set to 10 for both speech and noise codebooks. A 7-bit[1] speech codebook is trained with ∼7.5 minutes of clean speech from the above mentioned sources (i.e., 55 short sentences for each speaker). A 4-bit noise codebook is trained using over a minute of noise data from the available databases (i.e., about 15s for each noise type). For the testing, i.e., objective evaluation of the various enhancement methods, noisy speech files are generated by adding scaled segments of noise to the clean speech. For each speaker, 3 sentences are selected and combined with the four different types of noise, properly scaled to obtain the desired SNR values of 0, 5 and 10dB. For all cases, the speech and noise samples used for testing are kept different from those used to train the two codebooks.

[1] A N-bit codebook contains 2^N codebook vectors.

The parameters used in the codebook-based estimation scheme and modulation domain processing play a crucial role in the performance of the proposed CB-MME method. Fine tuning of these parameters is paramount for the effectiveness of the proposed enhancement method. The acoustic frame duration used for windowing speech samples in (2.3) is chosen to be $N = 512$ samples (32ms), i.e., long enough to validate the approximation used in the log-likelihood calculation (4.13) while short enough to satisfy the stationarity condition. The values of the other analysis parameters are chosen empirically as follows:

- Acoustic frame advance $F = 64$ (4ms) in (2.3): A shorter frame advance (i.e., larger overlap between consecutive frames) is recommended for effective modulation domain enhancement processing. It also provides a greater degree of freedom for choosing the values of frame length and frame advance in the modulation domain.

- Modulation frame duration $N_M = 20$ (80ms) in (2.27): The window length of the modulation frame is chosen to be 80ms following experiments with a range of values from 32 to 256ms). Large frame durations appear to cause some temporal spectral smearing in the resultant enhanced speech spectrum while smaller values of N_M tend to introduce some musical noise artefacts.

- Modulation frame advance $F_M = 2$ (8ms) in (2.3): The modulation frame advance value is chosen empirically after subjectively analysing the performance of CB-MME for a variety of frame advance values.

- Recursion factor in the decision directed method in (5.12) is set to $\alpha = 0.95$.

For the objective evaluation of the enhanced speech, we use the perceptual evaluation of speech quality (PESQ) [116] and the segmental SNR (SegSNR) as performance measures. A brief overview of these evaluation measures is presented in Subsections 6.1.1 and 6.1.2.

6.1.1 Perceptual evaluation of speech quality(PESQ)

PESQ is one of the most widely used measures for the objective evaluation of the performance of a speech enhancement method. PESQ tries to emulate the results of a subjective listening test by predicting the quality of the enhanced signal as it would be perceived by a listener. This tool requires both the enhanced speech, which serves the test signal, and the clean speech, which serves as the reference, while computing the score. The sampling

frequency is also required as an input. The score varies from 0 (worst) to 4.5 (best). Studies have shown that the PESQ measure has a significant correlation with attributes such as the signal distortion, the amount of background noise, and overall quality of the test signal [117]. Hence, a high PESQ measure signifies a good enhancement performance.

6.1.2 Segmental signal to noise ratio (Seg SNR)

Segmental SNR is defined as the average of SNR values calculated over short segments of speech. The SNR for a speech signal frame i of length N can be expressed as,

$$\mathrm{SNR}_i = 10\log_{10}\frac{\mathbf{s}_i^T\mathbf{s}_i}{(\mathbf{s}_i-\hat{\mathbf{s}}_i)^T(\mathbf{s}_i-\hat{\mathbf{s}}_i)} \tag{6.2}$$

where $\mathbf{s}_i = [s(n_i),\ldots,s(n_i+N-1)]$ and $\hat{\mathbf{s}}_i = [\hat{s}(n_i),\ldots,\hat{s}(n_i+N-1)]$ (n_i being the index of the first sample in the i^{th} frame) are $N\times 1$ vectors corresponding to the i^{th} frames of clean speech and enhanced speech signals, respectively. The average segmental SNR can be computed over all the frames as,

$$\mathrm{SegSNR} = \frac{1}{N_f}\sum_{i=0}^{N_f-1}\mathrm{SNR}_i = \frac{1}{N_f}\sum_{i=0}^{N_f-1}10\log_{10}\frac{\mathbf{s}_i^T\mathbf{s}_i}{(\mathbf{s}_i-\hat{\mathbf{s}}_i)^T(\mathbf{s}_i-\hat{\mathbf{s}}_i)} \tag{6.3}$$

where N_f is the total number of frames present in the speech signal. Higher SegSNR values indicate lesser residual background noise and a better performance of the enhancement method in question.

6.2 Evaluation of the Codebook Estimation Method

In this section, we present the experiments performed to evaluate the performance of the codebook approach used to estimate of speech and noise STP parameters, as exposed in Chapter 4. Subsections 6.2.1 and 6.2.2 present the experiments which are performed to choose the sizes of the speech and noise codebooks. Subsection 6.2.3 presents the experiments which are used to evaluate the accuracy of the codebook-based PSD estimates of noise and speech signals.

6.2.1 Codebook Size

The sizes of the speech and noise codebooks are chosen empirically based on the performance of the codebook in the STP estimation task. To this end, experiments are performed to evaluate different codebooks that are generated over a range of sizes. The average distortion measured between the data set vectors, consisting of gain normalized LP coefficients, and their representative codebook vectors serves as a marker for indicating how well the data set vectors are represented by the codebook in question.

The average distortion measure between a data set, $\mathcal{S} = \{\mathbf{s}_i \in \mathbb{R}^p | i = 1, 2..., K\}$, and a representative codebook, $\mathcal{C} = \{\mathbf{c}_j \in \mathbb{R}^p | j = 1, 2, ..., N\}$, is given by,

$$D^{\text{dist}} = \frac{1}{N} \sum_{j=1}^{N} \sum_{\mathbf{s}_i \in \mathcal{P}_{\mathbf{c}_j}} ||\mathbf{s}_i - \mathbf{c}_j||_2 \tag{6.4}$$

where $\mathcal{P}_{\mathbf{c}_j}$ denotes the cluster of the data set vectors which are represented a codebook vector $\mathbf{c}_j$ in $\mathcal{C}$.

A training data set consisting of gain normalized LP coefficients is derived from 7.5 minutes of clean speech data. Codebooks of size $N = 2^n$, where $n \in \{0, 1, \ldots, 9\}$, are trained on this data set and their respective distortion measures are computed as well. These distortion measures are then normalized using the distortion measure of the 0-bit codebook as the reference as follows,

$$D_n^{\text{norm}} = 20 \log_{10} \frac{D_n^{\text{dist}}}{D_0^{\text{dist}}} \tag{6.5}$$

where D_n^{dist} is the distortion measure for the n-bit codebook and D_n^{norm} is the normalized distortion measure for a n-bit codebook. A plot of the normalized distortion measure as a function of the codebook size is shown in Figure 6.2. As seen in the plot, the distortion measure keeps reducing as we train codebooks with larger sizes. This reduction is quite steep at the beginning of the plot (i.e., larger negative slope) and becomes more gradual (smaller slope) as we move towards 6 and 7 bit codebook sizes.

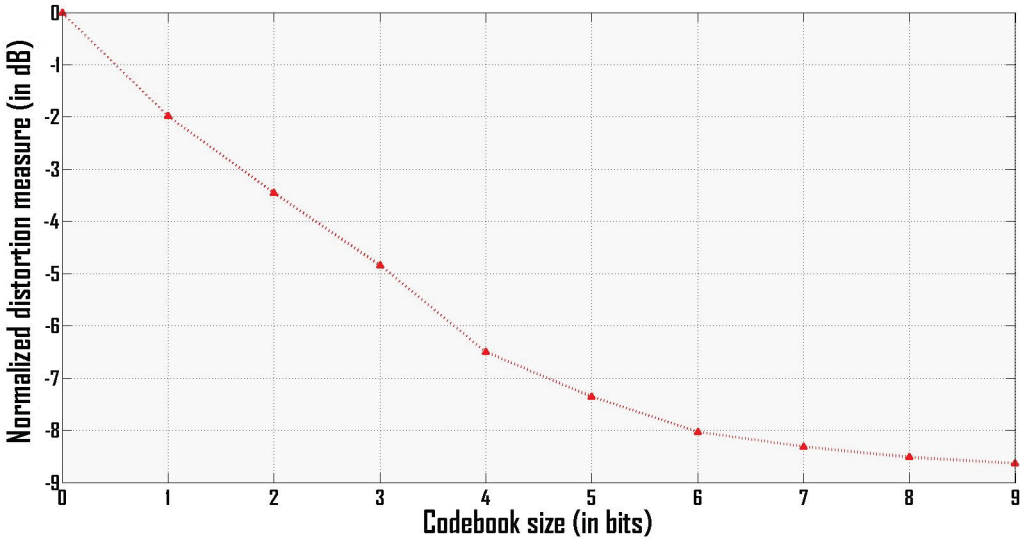


Fig. 6.2 Plot of normalized distortion measure with respect to codebook size. The distortion decreases as we increase the size of the codebook being trained. Hence, a larger codebook tends to represent the data set better than a smaller one.

Similar experiments are performed by training noise codebooks on a data set derived from over a minute of noise data, as explained earlier in Section 6.1. The distortion measures obtained are presented in the Fig. 6.3.

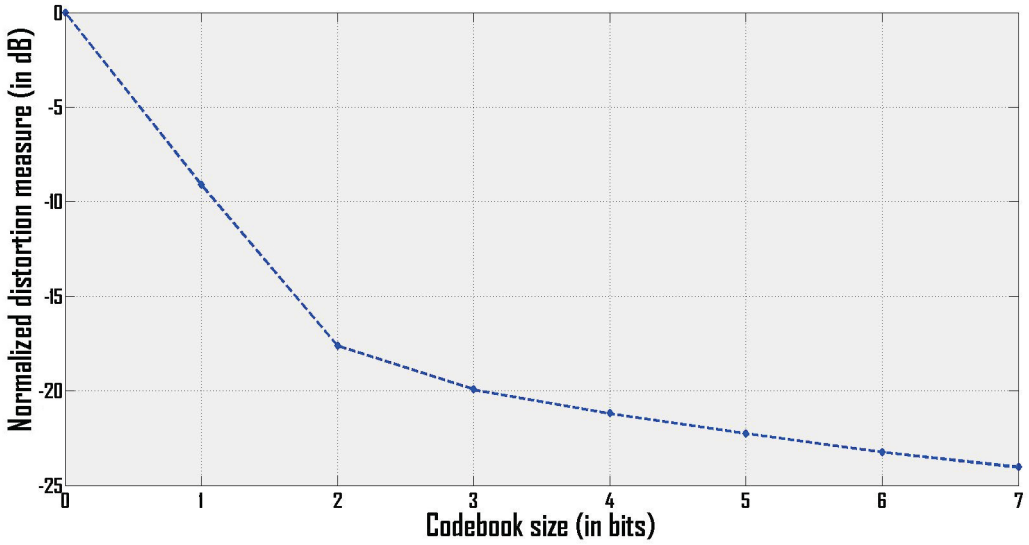


Fig. 6.3 Plot of normalized distortion measure with respect to the noise codebook size. The distortion measure decreases rapidly initially and slows down for sizes > 2 bits.

6.2.2 Computational Complexity of the Joint Estimation Scheme

The computational complexity of an algorithm provides a measure of the time taken by that algorithm to run completely. It is usually measured in terms of the number of computer instructions or operation cycles (e.g., floating point operations) needed to execute the algorithm and it plays a crucial role in determining its range of applications and usefulness. Indeed, in the real world, low complexity generally translates into faster execution and lower implementation costs. Computational complexity is often used as a performance measure of the algorithm along with the memory requirements. In some cases both measures are at odds and a tradeoff needs to be reached between the two.

Assuming that the speech and noise codebooks have been successfully trained, we focus on the complexity of the codebook search for the speech and noise PSD estimation, which is the computational bottleneck in our proposed enhancement approach as illustrated in Table 6.1.

Table 6.1 Distribution of processing time spent for the different operations in the proposed CB-MME method while enhancing a noisy speech signal of 6s in duration (as measured on a desktop computer equipped with a single Intel i7 core).

Operation	Elapsed time
Joint PSD estimation	119.78 s
Modulation transform	4.44 s
Filtering in modulation domain	9.58 s
Inverse modulation transform	22.49 s

The joint estimation scheme evaluates likelihood values for every possible combination of the speech and noise codebook vectors. The total number of iterations required to perform this operation will be a product of the number of vectors present in the speech and noise codebooks. Hence, the computational complexity of the estimation scheme is represented by an exponential function with the size of the codebooks (in bits) as the exponent ,as shown below,

$$C(m, n) = \mathrm{O}(2^{m+n}) \tag{6.6}$$

where $\mathrm{O}(\cdot)$ is the "big O" notation [118], while m and n are the sizes of the speech and noise codebooks in bits.

The time durations for performing joint estimation for different combinations of speech and noise codebook sizes are presented in Fig. 6.4. As seen in the plots presented, the processing time increases with the size of the speech and noise codebooks. To improve the time efficiency of the codebook-based joint estimation, it is advisable to choose small sizes for both speech and noise codebooks. However, this is at odds with the findings presented in Subsection 6.2.1 which recommend a larger codebook size for better representation of

the data set. In the end, the codebook sizes are selected as a tradeoff between these requirements.

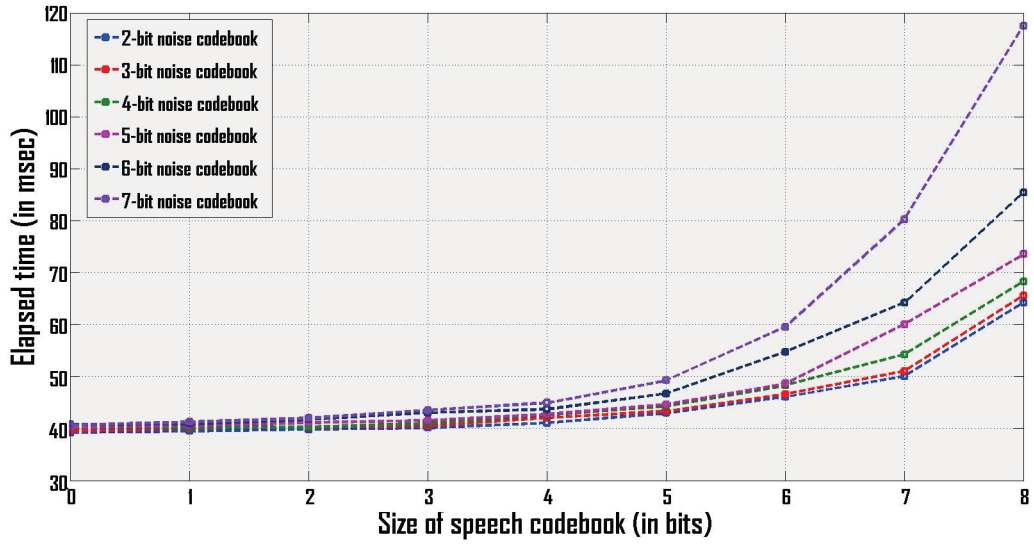


Fig. 6.4 Plots of the elapsed time for making the joint PSD estimation with respect to the sizes of speech and noise codebooks. The `tic` and `toc` commands in MATLAB are used to measure the elapsed times.

6.2.3 Accuracy of Codebook Estimation

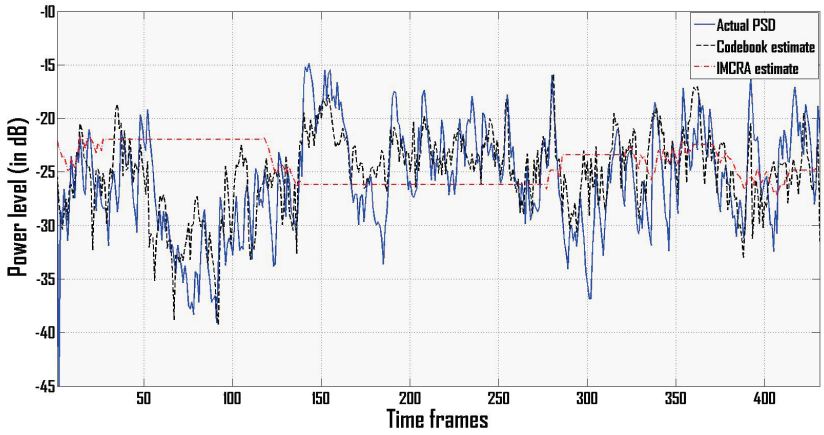
In this subsection, we present the experiments performed to test the accuracy of the codebook estimates of speech and noise PSDs. The codebook-based estimates of speech and noise are compared with the actual PSDs of the clean speech and noise signals present in the noisy speech signal.

Firstly, the noise tracking ability of the codebook-based estimation scheme is tested by plotting the temporal trajectory of a randomly chosen frequency bin of the codebook noise PSD estimate along with the actual noise PSD[2] and the corresponding IMCRA PSD estimate. A selection of speech utterances from a female speaker within the TIMIT database, corrupted by non-stationary noise types, i.e., street, babble and restaurant, were used as

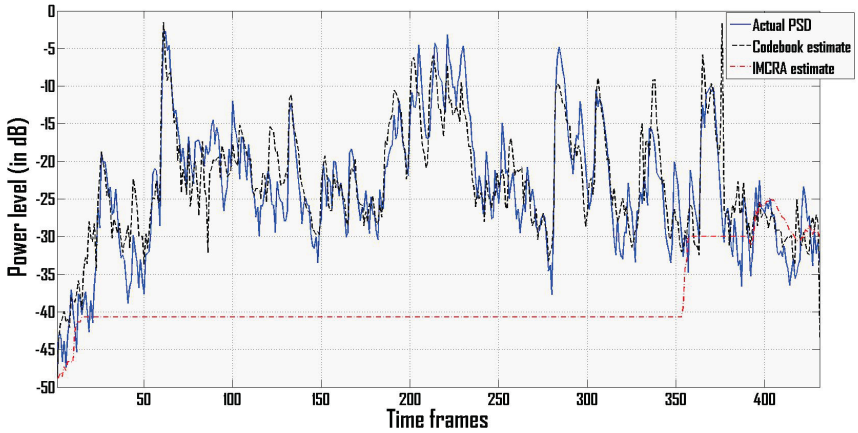
[2]The "actual" PSD is obtained by performing a recursive averaging operation of consecutive frames of the squared magnitude spectra of the background noise with smoothing factor set to 0.85. Note that in practice, only the noisy speech is available so that this actual PSD cannot be obtained from the observation; it represents an idealized PSD estimate.

the noisy speech signal for this noise tracking experiment. From the observation of Fig. 6.5a, 6.5b, and 6.5c, it is clear that the codebook-based estimation scheme offers a better performance at tracking the non-stationarity present in the background noise compared to a conventional estimation scheme like the IMCRA.

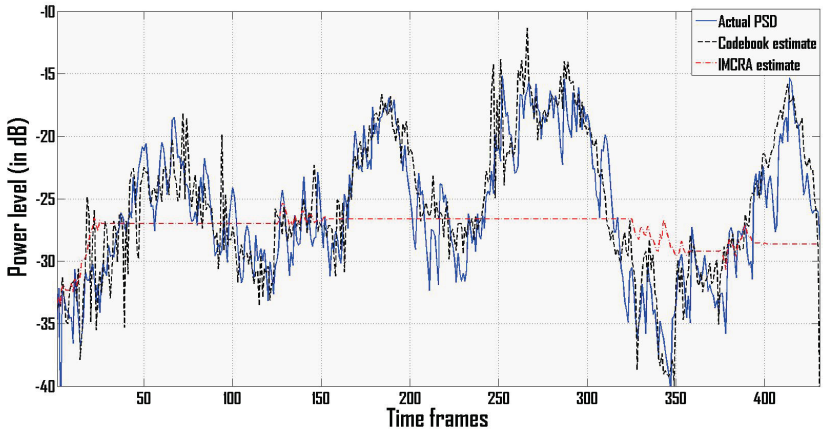
The codebook method's ability to estimate the PSD shapes of various speech and noise spectra is demonstrated by plotting them with the actual PSD of the signals in question. Specifically, Fig. 6.6, 6.7, and 6.8 contain the plots of the estimated and actual PSD obtained for selected frames and for different noise types at 0dB SNR. The clean speech used in this experiment is a selection of utterances from a female speaker within the TIMIT database. Three different non-stationary noise types, i.e., babble, street and restaurant are used as the corrupting background noise. Based on the observation of the plots obtained from these cases, we can conclude that the codebook scheme does seem to estimate the overall PSD shapes (i.e. spectral envelopes) of speech and noise signals in a satisfactory way although the fine details of the harmonic structure are lost.



(a) Babble noise PSD at 0dB SNR

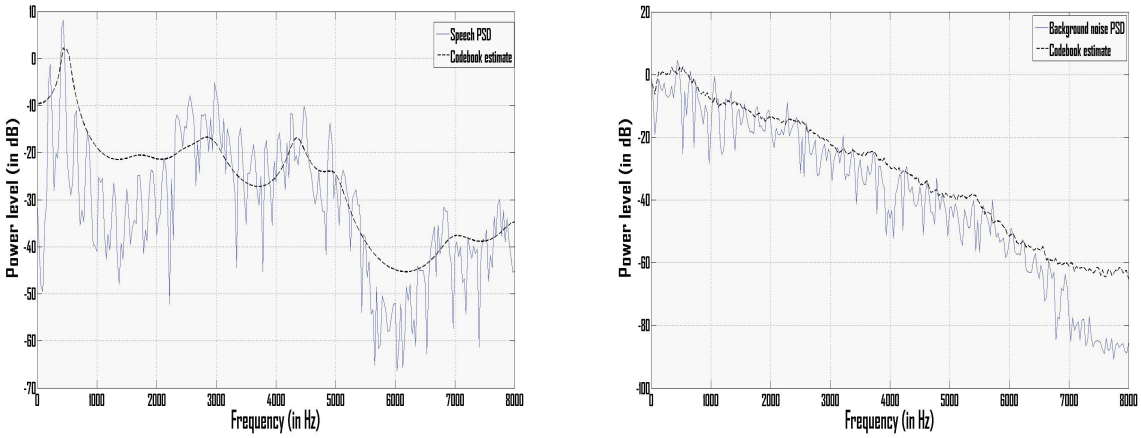


(b) Restaurant noise PSD at 0dB SNR



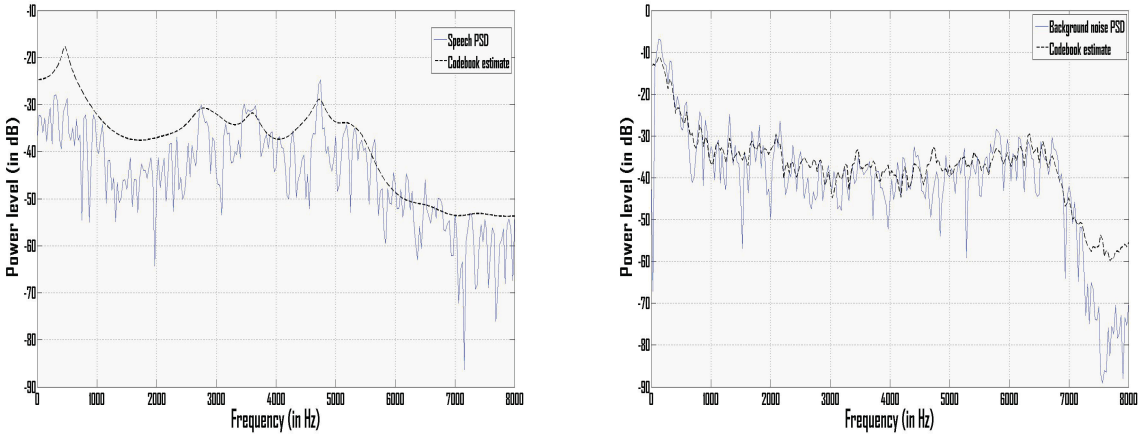
(c) Street noise PSD at 0dB SNR

Fig. 6.5 Noise tracking ability of the codebook-based method for three different noise types with rapidly changing non-stationary behaviour. The graphs contain the temporal trajectory of a frequency bin of the actual background noise PSD present in a speech signal (obtained from a female speaker in the TIMIT database) along with its codebook-based and IMCRA estimates.



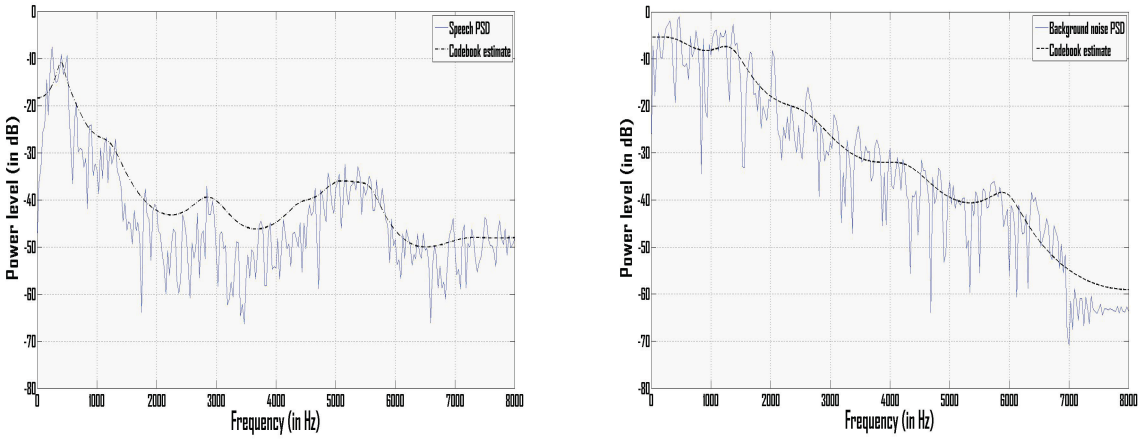
(a) (b)

Fig. 6.6 Codebook-based PSD estimate and actual PSD of (a) desired speech and (b) background noise for time frame $\nu = 200$ and noise type = babble noise.



(a) (b)

Fig. 6.7 Codebook-based PSD estimate and actual PSD of (a) desired speech and (b) background noise for time frame $\nu = 10$ and noise type = restaurant noise.



(a) (b)

Fig. 6.8 Codebook-based PSD estimate and actual PSD of (a) desired speech and (b) background noise for time frame $\nu = 100$ and noise type = street noise.

6.3 Performance Evaluation of CB-MME

The enhancement performance of CB-MME is evaluated by conducting objective evaluation experiments on the enhanced signals. In objective experiments, a test measure, which serves as the indicator on the quality of enhancement, is obtained through application of a rating algorithm on the enhanced signal. In most cases, the clean speech signal may be used as the reference signal in the rating algorithm.

In this work, as explained earlier, the perceptual evaluation of speech quality (PESQ) and segmental SNR measurements are used as the performance evaluators. Four non-stationary noise types, i.e., babble, non-stationary white, restaurant, and street noise are considered for the performance evaluation of CB-MME. The clean speech signal is corrupted by these background noise types at three SNR levels: 0, 5 and 10dB. Other enhancement methods, including the MMSE-STSA [9], MME [28] and codebook-based Wiener filter [45] methods are also evaluated for comparison with CB-MME. The following methods under comparison are labelled as follows for conciseness:

- MMSE - MMSE-STSA estimator.
- MME - Modulation domain based MMSE-STSA estimator.

- CB-WE - Wiener filter with codebook-based PSD estimates.

6.3.1 Perceptual evaluation of speech quality (PESQ)

Tables 6.2, 6.3, 6.4 and 6.5 contain the PESQ measures obtained for the different noise types mentioned above. Besides the scores of the MMSE, MME, CB-WE and the CB-MME, the evaluation score of the original noisy signal (Noisy) is also presented for keeping track of the relative improvement following the enhancement procedure. As mentioned earlier in this chapter, speech utterances which are three sentences (6s) long are used as test samples for calculating the PESQ score. The final PESQ at a particular SNR for each noise type is calculated as an average of the PESQ scores obtained from using four different test signals (one from each speaker whose utterances are used to train the speech codebook).

Table 6.2 PESQ values for non-stationary white noise

SNR level	Noisy	MMSE	MME	CB-WE	CB-MME
0 dB	1.75	1.78	2.04	2.11	**2.24**
5 dB	2.06	2.19	2.46	2.50	**2.58**
10 dB	2.42	2.51	2.83	2.89	**2.98**

Table 6.3 PESQ values for street noise

SNR level	Noisy	MMSE	MME	CB-WE	CB-MME
0 dB	1.72	1.85	1.95	2.00	**2.07**
5 dB	2.01	2.17	2.30	2.28	**2.40**
10 dB	2.35	2.58	2.73	2.69	**2.82**

Table 6.4 PESQ values for restaurant noise

SNR level	Noisy	MMSE	MME	CB-WE	CB-MME
0 dB	1.78	1.84	1.87	1.90	**2.04**
5 dB	2.13	2.20	2.27	2.29	**2.37**
10 dB	2.39	2.47	2.57	2.62	**2.75**

Table 6.5 PESQ values for babble noise

SNR level	Noisy	MMSE	MME	CB-WE	CB-MME
0 dB	1.67	1.83	1.93	1.98	**2.07**
5 dB	2.04	2.19	2.30	2.34	**2.43**
10 dB	2.35	2.52	2.65	2.71	**2.82**

6.3.2 Segmental signal -to- noise ratio (Seg SNR)

Tables 6.6, 6.7, 6.8 and 6.9 contain the SegSNR measures obtained for different noise types under same conditions as in Subsection 6.3.1.

Table 6.6 SegSNR values for non-stationary white noise

SNR level	Noisy	MMSE	MME	CB-WE	CB-MME
0 dB	-2.02	-1.19	0.57	0.91	**1.63**
5 dB	1.55	2.60	3.75	4.21	**5.04**
10 dB	4.16	5.19	6.47	7.13	**8.22**

Table 6.7 SegSNR values for street noise

SNR level	Noisy	MMSE	MME	CB-WE	CB-MME
0 dB	-2.75	-0.96	0.47	0.79	**1.09**
5 dB	0.72	1.35	1.91	1.85	**2.94**
10 dB	3.17	4.78	6.01	5.93	**7.54**

Table 6.8 SegSNR values for restaurant noise

SNR level	Noisy	MMSE	MME	CB-WE	CB-MME
0 dB	-2.44	-2.31	-0.59	-0.21	**0.71**
5 dB	1.14	1.43	2.07	2.35	**3.67**
10 dB	3.91	4.50	6.33	6.61	**7.28**

Table 6.9 SegSNR values for babble noise

SNR level	Noisy	MMSE	MME	CB-WE	CB-MME
0 dB	-3.02	-2.24	-0.85	-0.42	**0.47**
5 dB	0.84	1.28	2.36	2.59	**3.16**
10 dB	3.40	4.54	6.04	6.57	**7.19**

6.3.3 Discussion

It can be seen from the results presented in Subsections 6.3.1 and 6.3.2 that the proposed CB-MME method performs better than the MME, CB-WE and MMSE methods in most cases, for both objective performance metrics under consideration. Informal listening tests concur with these objective results. The speech enhancement methods (MME and MMSE) which use conventional approaches (IMCRA) for noise estimation seem to suppress the

stationary elements present in the background noise but fail to reduce the non-stationary portions properly. This is evident when listening to the enhanced signals processed by these methods. The codebook-based methods, i.e., CB-WE and CB-MME, suppress the non-stationary elements in the background noise much better than the MMSE and MME. However, this seems to come at the expense of some distortion in the resulting enhanced speech. This is probably caused by the spectral mismatch between the codebook-based speech PSD estimate and the actual one. The lack of fine structure in the codebook-based estimates speech and noise PSDs due to the low modelling order used in LP coefficient vectors can also play a part in creating this distortion, but at this time, this remains a conjecture.

Within codebook-based methods, the proposed CB-MME outperforms the CB-WE in most cases. The added advantage of performing the enhancement in the modulation domain seems to reflect on the results. This trend is also noticeable when comparing conventional methods, MMSE and MME. The MME performing better than MMSE by both performance metrics.

Chapter 7

Summary and Conclusion

This chapter provides some concluding remarks for this thesis. Section 7.1 presents a brief summary of this thesis work, while Section 7.2 lists suggestions for possible future work in this area.

7.1 Summary

In this thesis, we proposed a new speech enhancement method that uses codebook-based noise and speech estimates, in order to more effectively perform noise suppression in the modulation domain in the presence of non-stationary noise. Below, we provide a chapter wise sequential overview of the all the pertinent topics in the lead upto the proposal of the said method.

In Chapter 1, a concise summary on the background and issues of quality enhancement for single channel speech signals contaminated by background noise was initially presented. A literature survey on the topics related to the research conducted in this work, such as modulation domain processing and codebook-based speech enhancement was also presented.

Chapter 2 delved upon the topic of modulation domain processing. A brief review of the conventional acoustic domain methods was first presented. This was followed by the presentation of relevant background material on the modulation domain and its significance in speech processing. A few methods which perform speech enhancement in modulation domain were also briefly covered as well.

Chapter 3 covered the topic of noise PSD estimation in speech enhancement. A short

review of conventional noise estimation methods such as VAD based methods and the minimum statistics based methods was provided. The focus then shifted towards certain drawbacks of these conventional methods pertaining to the PSD estimation of non-stationary noise types. These drawbacks served as the motivation for working with alternative noise estimation schemes which can track non-stationary behaviour better than the conventional methods.

In Chapter 4, a detailed description of the codebook-based approach which was used to estimate the PSDs of background noise and speech in this work was presented. The description delved upon various aspects of the codebook-based approach, such as, the training of speech and noise codebooks, the joint estimation of speech and noise STP parameters using the codebooks, etc.

The proposed CB-MME enhancement method which incorporates codebook-based PSD estimates in the modulation domain enhancement, was finally presented in Chapter 5. A brief review of an earlier codebook-based Wiener filter was also provided in this chapter for the purpose of comparison.

Chapter 6 presented the experiments performed to evaluate the codebook-based estimation approach and the proposed CB-MME method. Objective tests (PESQ and SegSNR) applied on the CB-MME showed that in most cases this proposed method performed better than the other methods (MMSE-STSA, MME and CB-WE) taken into consideration.

7.2 Future Work

A list of possible directions for future work based on this thesis is presented in brief below:

- *Gain estimation for codebook vectors*: In Chapter 4, we covered gain adaptation for every combination of speech and noise codebooks while estimating the PSD shapes for speech and noise signals. The maximum likelihood gain estimation relies on a number of assumptions, such as, sufficiently long frames, small modelling error between actual spectrum and codebook estimated spectrum, etc. Inaccurate estimation of the STP parameters may occur if these assumptions do not hold true. Employing an iterative optimization algorithm like stochastic gradient descent with the gain estimates from (4.19) as starting values and the LLF as the cost function may result in better approximation of the gains.

- *Cost function used for designing the CB-MME gain*: In Chapter 5, the CB-MME gain is calculated by minimizing the mean square error between the clean speech and estimated speech modulation spectra as shown in (5.4). STSA estimators which employ different cost functions as the the Log MMSE STSA estimator [10], the Weighted-Euclidean STSA (WE STSA) estimator [11], the β-SA estimator [12] and the Weightedβ-SA (Wβ-SA) estimator [13] are known to show better performance than the MMSE-STSA estimator in noise reduction. Implementation of these estimators in the modulation domain while using the codebook estimates of speech and noise PSDs may result in even better performance than CB-MME.

- *Distortion in codebook-based methods*: As mentioned in Chapter 6, speech enhancement using the codebook-based estimates of speech and noise PSD creates some slight distortion in the enhanced speech despite performing good noise reduction. This is probably caused by the lack of fine harmonic structure in the spectral envelopes estimated using codebook methods. Incorporating a pitch predictor with the codebook speech estimate as suggested in [48] may result in less distortion.

References

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, Springer Edition, 2004.

[2] K. Farrrel, R. J. Mammone, and J. L. Flanagan, "Beamforming microphone arrays for speech enhancement," *Proc. IEEE Intern. Conf. on Acoust. Speech, and Signal Process.*, vol. 1, pp. 285-288, Mar. 1992.

[3] D. Yellin and E. Weinstein, "Multichannel signal separation: Methods and analysis," *IEEE Trans. on Signal Process.*, vol. 44, pp. 106-118, Jan. 1996.

[4] H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Trans. on Audio and Electroacoustics*, vol. 16, pp. 165-168, Jun. 1968.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 27, pp. 113-120, Apr. 1979.

[6] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE Intern. Conf. on Acoustics, Speech, and Signal Process.*, vol. 4, pp. 208-211, Apr. 1979.

[7] M. M. Sondhi, C. E. Schmidt, and L. R. Rabiner, "Improving the quality of a noisy speech signal," *Bell Syst. Tech. J.*, vol. 60, pp. 1847-1859, Oct. 1981.

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

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 32, pp. 1109-1121, Dec. 1984.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 33, pp. 443-445, Apr. 1985.

[11] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. on Speech and Audio Process.*, vol. 13, pp. 857-869, Aug. 2005.

[12] C. H. You, S. N. Koh, and S. Rahardja, "β-order mmse spectral amplitude estimation for speech enhancement, " *IEEE Trans. on Speech and Audio Process.*, vol. 13, pp. 475-486, Jul. 2005.

[13] E. Plourde and B. Champagne, "Generalized Bayesian estimators of the spectral amplitude for speech enhancement," *IEEE Signal Process. Letters*, vol. 16, pp. 485-488, Mar. 2009.

[14] J. Chen, J. Benesty and Y. Huang, "New insights into the noise reduction Wiener filter," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 14, pp. 1218-1234, Jul. 2006.

[15] T. V. Bogaert, S. Doclo, J. Wouters, and M. Moonen,"Speech Enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoust. Society of America*, vol. 125, pp. 360-371, Oct. 2008.

[16] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *IEEE Intern. Conf. on Acous., Speech, and Signal Process.*, vol. 12, pp. 177-180, Apr. 1987.

[17] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. on Speech and Audio Process.*, vol. 6, pp. 373-385, Jul. 1998.

[18] Z. Goh, K. C. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. on Speech and Audio Process.*, vol. 7, pp. 510-524, Sep. 1999.

[19] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. on Speech and Audio Process.*, vol. 3, pp. 251-266, Jul. 1995.

[20] P. S. K. Hansen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Noise reduction of speech signals using rank-revealing ULLV decomposition," *Proceed. in EUSIPCO*, vol. 2, pp. 967-970, Sep. 1996.

[21] F. Jabloun and B. Champagne, "A perceptual signal subspace approach for speech enhancement in colored noise," *IEEE Intern. Conf. on Acous., Speech, and Signal Process.*, vol. 1, pp. 569-572, May 2002.

[22] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. on Speech and Audio Process.*, vol. 11, pp. 700-708, Nov. 2003.

[23] K. Hermus and P. Wambacq, "Assessment of signal subspace based speech enhancement for noise robust speech recognition," *IEEE Intern. Conf. on Acous., Speech, and Signal Process.*, vol. 1, pp. 945-948, May 2004.

[24] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 25, pp. 235-238, Jun. 1977.

[25] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 32, pp. 236-243, Apr. 1984.

[26] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Education, 2002.

[27] K. Paliwal, K. Wojcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun.*, vol. 52, pp. 450-475, May 2010.

[28] K. Paliwal, B. Schwerin, and K. Wojcicki,"Speech enhancement using minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Commun.*, vol. 54, pp. 282-305, Feb. 2012.

[29] S. So and K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Commun.*, vol. 53, pp. 818-829, Jul. 2011.

[30] N. Kowalski, D. Depireux, and S. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses to moving ripple spectra," *Journal of Neurophysiology*, vol. 76, pp. 3503-3523, Nov. 1996.

[31] D. Depireux, J. Simon, D. Klein, and S. Shamma, "Spectrotemporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 85, pp. 1220-1234, Mar. 2001.

[32] S. Shamma, "Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method," *Network: Computation in Neural Systems*, vol. 7, pp. 439-476, Jul. 1996.

[33] S. Bacon and D. Grantham, "Modulation masking: Effects of modulation frequency, depth, and phase," *Journal of Acoust. Society of America*, vol. 85, pp. 2575-2580, 1989.

[34] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal of Acoust. Society of America*, vol. 95, pp. 1053-1064, 1994.

[35] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Letters*, vol. 6, pp. 1-3, Jan. 1999.

[36] A. Sangwan, W. P. Zhu, and M. O. Ahmad, "Improved voice activity detection via contextual information and noise suppression," *IEEE Intern. Symp. on Circuits and Systems*, vol. 2, pp. 868-871, May 2005.

[37] C. C. Hsu, T. E. Lin, J. H. Chen, and T.-S. Chi, "Voice activity detection based on frequency modulation of harmonics," *IEEE Intern. Conf. on Acoust., Speech and Signal Process.*, pp. 6679-6683, May 2013.

[38] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Process.*, vol. 9, pp. 504-512, Jul. 2001.

[39] I. Cohen and B. Berdugo,"Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Letters.*, vol. 9, pp. 12-15, Jan. 2002.

[40] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Process.*, vol. 11, pp. 466-475, Sep. 2003.

[41] T. Gerkmann and R. C. Hendriks,"Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 20, pp. 1383 -1393, May 2012.

[42] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," *Proceedings of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Process.*, vol.3, pp. 1875-1878, Jun. 2000.

[43] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Speech enhancement using a-priori information," *INTERSPEECH*, pp. 1405-1408, Sep. 2003.

[44] M. Kuropatwinski and W. B. Kleijn, "Estimation of the short-term predictor parameters of speech under noisy conditions," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 14, pp. 1645-1655, Sep. 2006.

[45] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short- term predictor parameter estimation for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 14, pp. 163-176, Jan. 2006.

[46] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 15, pp. 441-452, Feb. 2007.

[47] T. Rosenkranz, "Modeling the temporal evolution of LPC parameters for codebook-based speech enhancement," *Intern. Symp. on Image and Signal Process. and Analysis*, pp. 455-460 , Sep. 2009.

[48] T. Rosenkranz and H. Puder, "Integrating recursive minimum tracking and codebook-based noise estimation for improved reduction of non-stationary noise," *Signal Process.*, vol. 92, pp. 767-779, Mar. 2012.

[49] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communications*, vol. 28, pp. 84-95, Jan. 1980.

[50] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th Edition, Academic Press, 2000.

[51] E. Plourde,"Bayesian Short-time Spectral Amplitude Estimators for Single-Channel Speech Enhancement," Ph.D. dissertation, McGill University, Montreal, Canada, 2009.

[52] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. on Speech and Audio Process.*, vol. 2, pp. 345-349, Apr. 1994.

[53] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Process.*, vol. 2003, pp. 668-675, Feb. 2003.

[54] H. Dudley, "Remaking speech," *Journal of Acoust. Society of America*, vol. 11, pp. 169-177, Oct. 1939.

[55] A. R. Moller, "Unit Responses in the Rat Cochlear Nucleus to Tones of Rapidly Varying Frequency and Amplitude," *Acta Physiologica Scandinavica*, vol. 81, pp. 540-556, Apr. 1971.

[56] N. Suga, "Analysis of information-bearing elements in complex sounds by auditory neurons of bats," *International Journal of Audiology*, vol. 11, pp. 58-72, Apr. 1972.

[57] C. Schreiner and J.Urbas, "Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF)," *Hearing Research*, vol. 21, pp. 227-241, 1986.

[58] N. Viemeister, "Temporal factors in audition: A systems analysis approach," *Psychophysics and Physiology of Hearing*, pp. 419-427, 1977.

[59] T. Houtgast, "Frequency selectivity in amplitude-modulation detection," *Journal of the Acoust. Society of America*, vol. 85, pp. 1676-1680, 1989.

[60] S. Sheft and W. Yost,"Temporal integration in amplitude modulation detection," *Journal of the Acoust. Society of America*, vol. 88, pp. 796-805, 1990.

[61] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *Journal of the Acoust. Society of America*, vol. 95, pp. 2670-2680, 1994.

[62] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes, "*Proceed. in Intern. Conf. on Spoken Language Process.*, vol. 4, pp. 2490-2493, Oct 1996.

[63] T. Houtgast and H. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *Journal of the Acoust. Society of America*, vol. 77, pp. 1069-1077, 1985.

[64] T. Chi, Y. Gao, M. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *Journal of the Acoust. Society of America*, vol. 106, pp. 2719-2732, 1999.

[65] M. Elhilali, T. Chi, and S. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, pp. 331-348, Oct. 2003.

[66] B. Schwerin, "Modulation Domain Based Processing for Speech Enhancement," Ph.D. thesis, Griffith University, Brisbane AUS, 2013.

[67] L .E. Atlas and M. S. Vinton, "Modulation frequency and efficient audio coding," *Intern. Symp. on Optical Science and Tech.*, pp. 1-8, International Society for Optics and Photonics, 2001.

[68] J. K. Thompson and L .E. Atlas, "A non-uniform modulation transform for audio coding with increased time resolution," *IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, vol. 5, pp. 397-400, Apr. 2003.

[69] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Process.*, vol. 2, pp. 578-589, Oct. 1994.

[70] C. Nadeu, P. P.- Leal, and B. H. Huang, "Filtering the time sequences of spectral parameters for speech recognition," *Speech Commun.*, vol. 22, pp. 315-332, Sep. 1997.

[71] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, pp. 117-132, Aug. 1998.

[72] V. Tyagi, I. A. McCowan, H. Bourlard, and H. Misra, "On factorizing spectral dynamics for robust speech recognition," *Eurospeech*, Sep. 2003 .

[73] X. Xiong, E. S. Chng, and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 16, pp. 1662-1674, Nov. 2008.

[74] X. Lu, S. Matsuda, M. Unoki, and S. Nakamura, "Temporal contrast normalization and edge-preserved smoothing of temporal modulation structures of speech for robust speech recognition," *Speech Commun.*, vol. 52, pp. 1-11, Jan. 2010.

[75] S. V. Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," *Proceed. in Int. Conf. for Spoken Lang. Process.*, pp. 3205-3208, Nov. 1998.

[76] N. Malayath, H. Hermansky, S. Kajarekar, and B. Yegnanarayana, "Data-driven temporal filters and alternatives to GMM in speaker verification," *Digital Signal Process.*, vol. 10, pp. 55-74, Jan. 2000 .

[77] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," *IEEE Intern. Conf. on Acoust., Speech and Signal Process.*, vol. 1, pp. I, May 2006.

[78] T. Kinnunen, K. -A. Lee, and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification," *ICSA Speaker and Lang. Recognition Workshop (ODYSSEY)*, pp. 30, Jan. 2008.

[79] H. J. M. Steeneken and T. Houtgast , "A physical method for measuring speech transmission quality," *Journal of Acous. Society of America*, vol. 67, pp. 318-326, 1980.

[80] K. Payton and L. Braida, "A method to determine the speech transmission index from speech waveforms," *Journal of Acous. Society of America*, vol. 106, pp. 3637-3648, 1999.

[81] S. Greenberg and T. Arai, "The relation between speech intelligibility and the complex modulation spectrum," *INTERSPEECH*, pp. 473-476, Sep. 2001.

[82] R. Goldsworthy and J. Greenberg, "Analysis of speech-based Speech Transmission Index methods with implications for nonlinear operations," *Journal of Acous. Society of America*, vol. 106, pp. 3679-3689, 2004.

[83] D. -S. Kim, "A cue for objective speech quality estimation in temporal envelope representations," *IEEE Signal Process. Letters*, vol. 11, pp. 849-852, Oct. 2004

[84] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," *IEEE Intern. Conf. on Acoust., Speech, and Signal Process.*, vol. 7, pp. 156-159, May 1982.

[85] H. Hermansky, E. Wan, and C. Avendano, "Speech enhancement based on temporal processing," *IEEE Intern. Conf. on Acoust., Speech, and Signal Process.*, vol. 1, pp. 405-408, May 1995.

[86] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," *IEEE Intern. Conf. on Acoust., Speech, and Signal Process.*, vol. 1, pp. 121-124, Mar. 1992.

[87] C. Avendano, H. Hermansky, M. Vis, and A. Bayya, "Adaptive speech enhancement using frequency-specific SNR estimates," *3rd IEEE Workshop on Interactive Voice Tech. for Telecomm. Applications*, pp. 65-68, Sep. 1996.

[88] N. Mesgarani and S. Shamma, "Speech enhancement based on filtering the spectrotemporal modulations," *IEEE Intern. Conf. on Acoust., Speech, and Signal Process.*, vol. 1, pp. 1005-1008, Mar. 2005.

[89] L. Zadeh, "Frequency analysis of variable networks," *Proceed. of IRE*, vol. 38, pp. 291-299, Mar. 1950.

[90] T. Kailath, "Channel characterization: time-variant dispersive channels," *Lectures on Communi. System Theory*, (edited by EJ Baghdady), pp. 95-123, McGraw-Hill, New York, USA, 1961.

[91] Y. Zhang and Y. Zhao, "Spectral subtraction on real and imaginary modulation spectra," *IEEE Intern. Conf. on Acoust., Speech, and Signal Process.*, pp. 4744-4747, May 2011.

[92] B. Schwerin and Y. Paliwal, "Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement," *Speech Commun.*, vol. 58, pp. 49-68, Mar. 2014.

[93] Y. Wang and M. Brookes, "A subspace method for speech enhancement in the modulation domain," *Proceed. in Euro. Signal Process. Conf. (EUSIPCO)*, pp. 1-5, Sep. 2013.

[94] J. Ramirez, J. M. Grriz, and J. C. Segura, *Voice activity detection. fundamentals and speech recognition system robustness*, INTECH Open Access Publisher, 2007.

[95] R. Martin and D. Kolossa, "Voice Activity Detection, Noise Estimation, and Adaptive Filters for Acoustic Signal Enhancement," *Techniques for Noise Robustness in Automatic Speech Recognition*, pp. 51-85, 2012.

[96] A. Benyassine, E. Shlomot, H.-Y Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "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 applications," *IEEE Communi. Magazine*, vol. 35, pp. 64-73, Sep. 1997.

[97] ETSI and GSM, "06.94: Digital cellular telecommunications system (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *Euro. Telecommuni. Standards Inst.(ETSI)*, Feb. 1999.

[98] R. Martin, "Spectral subtraction based on minimum statistics," *Power*, vol. 6, pp. 8, 1994.

[99] S. Rangachari, "Noise Estimation for Highly Non-stationary Environments," Masters thesis, The University of Texas at Dallas, Dallas, U.S.A, 2004.

[100] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. on Signal Process.*, vol. 39, pp. 1732-1742, Aug. 1991.

[101] T. Sreenivas and P. Kirnapure, "Codebook constrained Wiener Filtering for speech enhancement," *IEEE Trans. on Speech and Audio Process.*, vol. 4, pp. 383-389, Sep. 1996.

[102] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. on Speech and Audio Process.*, vol. 6, pp. 445-455, Sep. 1998.

[103] J.R. Markel and A. H. Gray Jr., *Linear prediction of speech*, Vol. 12, Springer Science & Business Media, 2013.

[104] G. Ghodoosipour, "A Codebook-Based Modelling Approach for Bayesian STSA Speech Enhancement," Ph.D. thesis, McGill University, Montreal, Canada, 2014.

[105] D. O'shaughnessy, *Speech communication: human and machine*, Universities press, 1987.

[106] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[107] W. H. Equitz, "A new vector quantization clustering algorithm," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 37, pp. 1568-1575, Oct. 1989.

[108] M. Kuropatwinski and W. B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding," *Proceed. in IEEE Intern. Conf. on Acoust., Speech, and Signal Process.*, vol. 1, pp. 669-672, May 2001.

[109] U. Grenander and G. Szego, *Toeplitz Forms and their Applications*, 2nd Ed., Chelsea, 1984.

[110] R. M. Gray, A. Buzo, A. H. Gray Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. ON Acoust., Speech, Signal Process.*, vol. 28, pp. 367-376, Aug. 1980.

[111] P. Kabal, McGill University, "TSP speech database," Tech. Rep., 2002.

[112] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT: acoustic-phonetic continuous speech corpus," Linguistic Data Consort., 1993.

[113] Rice University, "Signal processing information base: noise data."

[114] Sound Jay, "Ambient and special sound effects." Available online: http://www.soundjay.com/ambient-sounds-2.html.

[115] T. P. Krauss, L. Shure, and J. N. Little "Signal Processing Toolbox for use with MATLAB," 1994.

[116] ITU-T. P.862, "Perceptual evaluation of speech quality (PESQ): and objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep., 2000.

[117] Y.Hu and P.C. Loizou, "Evaluation of objective measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 16, pp. 229-238, Jan. 2008.

[118] J. E. Hopcroft and J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley Boston, 1983.