Pre-Processing of Noisy Speech for Voice Coders

Tarun Agarwal

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

January 2002

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering.

2002/01/24

0-612-79056-8

To my beloved parents and sisters.

Abstract

Accurate Linear Prediction Coefficient (LPC) estimation is a central requirement in low bit-rate voice coding. Under harsh acoustic conditions, LPC estimation can become unreliable. This results in poor quality of encoded speech and introduces annoying artifacts.

The purpose of this thesis is to develop and test a *two-branch* speech enhancement pre-processing system. This system consists of two denoising blocks. One block will enhance the degraded speech for accurate LPC estimation. The second block will increase the perceptual quality of the speech to be coded. The goals of this research are two-fold—to design the second block, and to compare the performance of other denoising schemes in each of the two branches. Test results show that the two-branch system can provide better perceptual quality of coded speech over conventional one-branch (i.e., one denoising block) speech enhancement techniques under many noisy environments.

Sommaire

L'estimation précise des Coefficients de Prediction Linéaire (LPC) est d'une grande importance pour le codage de la voix à faible débit binaire. Sous de mauvaises conditions acoustiques, l'estimation LPC peut devenir incertaine. En conséquence, des perturbations sont introduites pour contribuer à une mauvaise quality de la voix.

Le but de cette thèse est de déveloper et tester un système d'enrichissement de la voix à deux branches. Ce système comprend deux blocs différents pour enlever le bruit sonore. La premierè branche enrichie la voix degradée pour améliorer l'estimation LPC. La seconde branche augmente la quality de perception de la voix codée. L'objectif est de déveloper cette dernière et de la comparer avec d'autres méthodes d'enrichissement dans chacune des deux branches. Des résultats montrent que le système a deux branches (ayant le bloc developé dans l'une de ces branches) offre une meilleure performance que certains systèmes d'enrichissement conventionels, dans une variété d'environments bruyant.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Prof. Peter Kabal, for his valuable guidance, timely suggestions and continuous support throughout my graduate studies at McGill University. Prof. Kabal has been extremely patient and has always motivated me by either pointing to the latest publications or providing the latest computer software. I would also like to thank Dr. Hossein Najafzadeh-Azghandi for constant feedback and providing research directions. I would like to thank Nortel Networks (Canada) for providing me with financial support.

I am truly indebted to fellow researchers and close friends; Mr. Christopher Cave, Mr. Wesley Pereira, Mr. Aziz Shallwani, Mr. Mark Klein and Mr. Paxton Smith. They have provided technical help and critical insight throughout my research. They have always been around to provide useful suggestions, companionship and created a stimulating research environment in the Telecommunications and Signal Processing laboratory. Christopher has been particularly helpful with the French abstract, and I really thank him for motivating me morally and intellectually to complete this thesis—be it by dragging me to the gym or going to Tim Hortons with me to pick up coffee at odd hours of the day!

I would like to extend my thanks to Mr. George Attia, Ms. Sohini Guha, Ms. Karishma Punwani, Ms. Kamakshi Advani and Ms. Shweta Gupta for being with me during good times and bad times. They have always been there to listen to my predicaments and cheer me up during nostalgic moments. I would also like to thank Mr. Vikramaditya for standing up for me on occasions.

Special thanks are also extended to the Sehgals. They have been my guardian in Montreal. They have always been around to give me the love and support.

Although my parents and sisters live in different countries, this research would not have been complete without their love, never-failing moral support, encouragement, trust and confidence in me—renewed either through extended hours on the phone, e-mails or visits to Montreal. They have been there at the back of my mind and in my heart throughout the completion of my two degrees at McGill University. Words cannot express my gratitude to my family. They are my pride and my perpetual source of confidence. My father Mr. Radhey Shyam Agarwal, my mother Mrs. Santosh Agarwal, and my sisters Mrs. Nalini Rastogi and Dr. Nivedita Agarwal, thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the advent of digital cellular telephones, the role of noise suppression in speech coding has gained importance. Recent advances in speech pre-processing algorithms have proven to be successful in improving the quality and intelligibility of speech signals even under harsh acoustic conditions. This importance of speech denoising in not only due to customer expectation of high speech quality (e.g., when using a wireless phone in the car, while the engine is running), but also to improve lower data rate transmission (i.e., using less bandwidth) to accommodate the ever-increasing number of cellular telephone customers. The use of speech enhancement techniques as a pre-processing stage to speech coders is being applied to governmental (military) wireless communication systems as well.

An important problem in some digital communication systems is the degraded performance of Linear Prediction (LP) analysis of speech inputs that have been corrupted by noisy environments. Noise types could be several—ranging from local city din to air-conditioning noise in an office room to HMMWV[1] and CH-47[2]. One of the primary goals of this thesis is to present a suitable speech enhancement system as a pre-processor for voice coding.

The purpose of this chapter is to give the reader a cursory preview of both low bit-rate coders, see Section 1.1, and speech enhancement techniques, see Section 1.2. Later Section 1.3, discusses some of the earlier work done in joint systems (low bit-rate coders appended with noise suppression rules). Section 1.4, mentions the contribution made to

[1] HMMWV: High Mobility Multi-wheeled Vehicle, more commonly known as the hummer is the modern equivalent of the jeep.

[2] CH-47: Chinook-47, army cargo helicopter.

speech enhancement for low bit-rate voice coding. Section 1.5 gives the organization of the chapters.

1.1 Low Bit-Rate Speech Coders

Speech coding algorithms come in three flavours: *waveform coders*, *parametric coders* and *hybrid coders*. The objective behind waveform coders is to yield a reconstructed signal that is closely matched to the original signal. With increasing bit-rate, the reconstructed speech converges towards the original signal. Parametric coders (also known as vocoders), on the other hand, rely on speech-specific analysis-synthesis which is mostly based on source-system model. A basic vocoder is seen in Fig. 1.1.

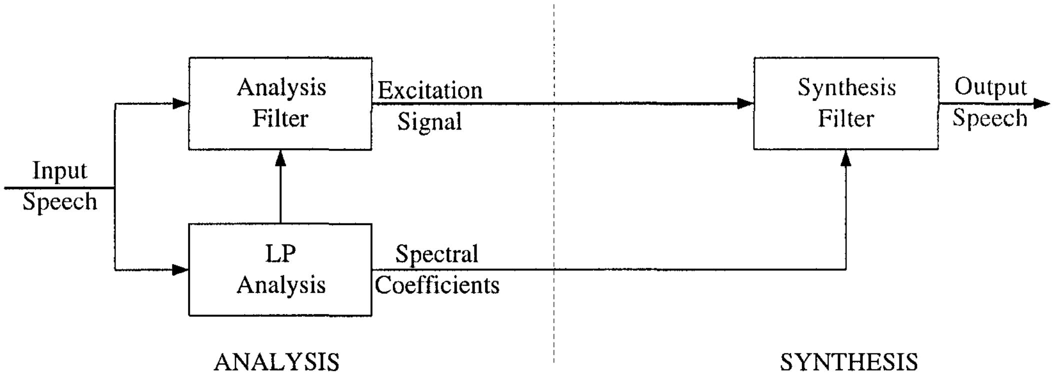


Fig. 1.1 Block diagram of a basic vocoder

Articulatory models represent the human speech production mechanism directly, i.e., distinct human voice-production organs are modelled explicitly. The excitation signal is produced by passing the input speech through the Analysis Filter, see Fig. 1.1. This excitation signal can be modelled by either periodic pulses or random white noise for voiced and unvoiced/fricative[3] segments, respectively. The excitation signal is then filtered through the Synthesis Filter in an attempt to reconstruct speech. Although the reconstructed speech

[3] Speech consists of pressure waves created by the flow of air through the vocal tract. These sound pressure waves originate in the lungs as the speaker exhales. The vocal folds in the larynx open and close quasi-periodically to interrupt this air flow. This results in voiced speech (e.g.,vowels) which is characterized by its periodic and energetic nature. Consonants are an example of unvoiced speech–aperiodic and weaker. These sounds have a noisy nature due to turbulence created by the flow of air through a narrow constriction in the vocal tract, [1].

sounds synthetic, parametric coders have the advantage of using few bits. Why would it be beneficial to use a lower number of bits at the cost of quality? The reasons are several—for storage applications, using few bits means less memory is required; for transmission applications, a lower bit-rate means less bandwidth is used, thus more users can be accommodated within the limit of available frequency. As seen in Fig. 1.2, vocoders offer better

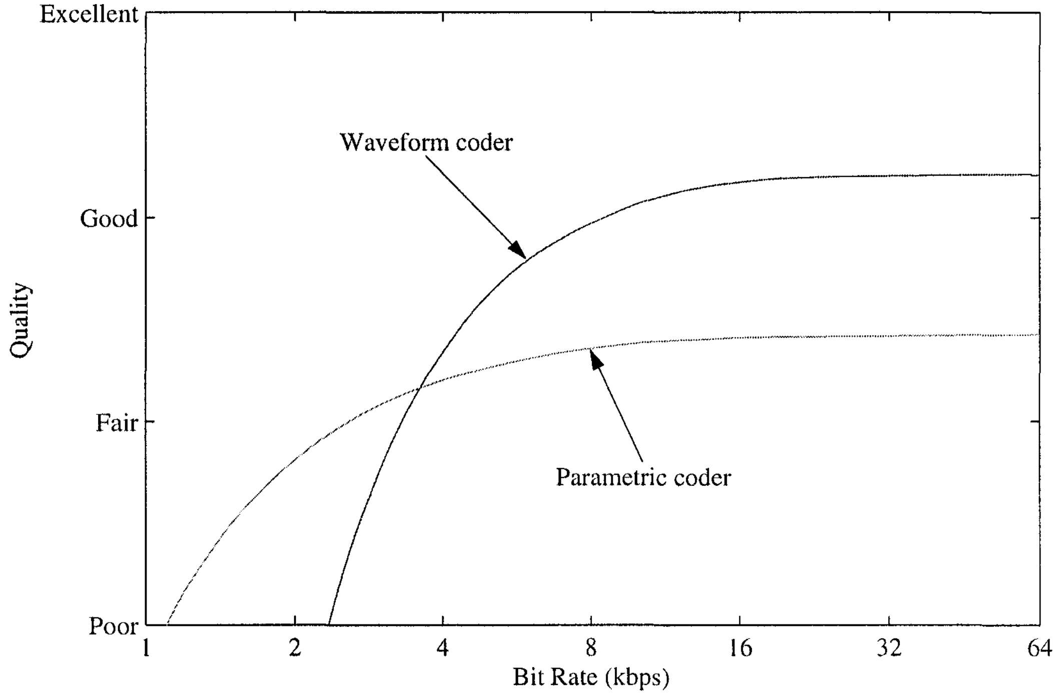


Fig. 1.2 Subjective performance of waveform and parametric coders. Redrawn from [1].

perceptual quality for speech coded at fairly low bit-rates (< 4 kbps).

Traditional pitch-excited LPC vocoders use a fully parametric model to efficiently encode the important information in human speech. Some of the vocoders are Linear Predictive Coding-10 (LPC-10), Mixed Excited Linear Prediction (MELP), Residual Excited Linear Prediction (RELP) which have been adequately summarized in [2]. These vocoders tend to produce intelligible speech at low data rates (800–2400 bps), but they often sound synthetic and generate annoying artifacts such as thumps, buzzes and tonal noises, [3]. These problems exacerbate dramatically in the presence of background noise, [4].

Several researchers, in past and present, have been motivated by the problem of improving speech coder performance under harsh acoustic conditions. The following subsection attempts to highlight some of the current speech enhancement techniques that can be deployed as a pre-processor stage to speech compression.

1.2 Speech Enhancement Techniques

The task of speech enhancement involves processing speech signals for human listening or as preparation for further processing prior to listening. Thus the main objective of speech enhancement is ultimately to improve one or more of perceptual aspects of speech, such as overall quality and intelligibility, or to reduce the degree of listener fatigue, [5]. This section will briefly review some techniques.

A comprehensive overview of various noise suppression techniques can be found in [6]. The listed methods include spectral subtraction, Wiener estimation, Maximum-likelihood estimation, soft-decision method and are adequately summarized in [7]. Most of these algorithms operate in the frequency-domain. A basic problem representation is presented in Fig. 1.3. Clean speech that gets corrupted with background noise is sent through a speech enhancement block (one of those listed above). The purpose of such a block is to estimate the magnitude of speech from the noisy observation.

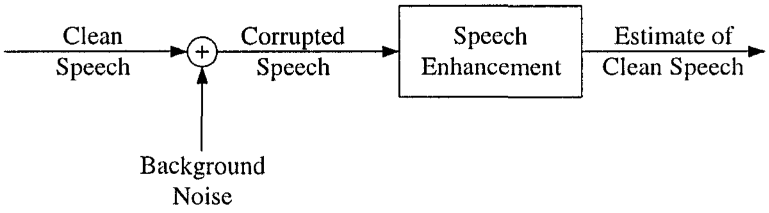


Fig. 1.3 Block diagram of basic speech enhancement.

The differences in the noise suppression algorithms crop up in the cost-function used to compute the *soft-decision gain.* For instance, in the spectral subtraction algorithm, the Short-Time Spectral Amplitude (STSA) is estimated as the square root of the Maximum-Likelihood (ML) estimator of each signal spectral component variance. In systems exploiting Wiener filtering, the STSA estimator is obtained as the modulus of the optimal Minimum Mean-Squared Error (MMSE) amplitude estimator of each signal spectral component. A common feature of these techniques is that the noise reduction process brings about very unnatural artifacts called "musical noise". In 1984, Ephraim and Malah derived an MMSE-STSA estimator that assisted in the reduction of the annoying musical noise. The gain computed in their algorithm is based on the probability of speech absence and is computed for each frequency bin. Later in 1985, they published a paper, [8], where the MMSE of the Log-Spectra Amplitude (MMSE-LSA) is used in enhancing noisy speech. In either case, the enhanced speech sounds very similar, with the exception that the first estimator results in less uniform residual noise, [8]. Another noise suppression rule, of interest is part of the Enhanced Variable Rate Coder (EVRC) [9]. This noise suppression rule tends to leave some residual noise, in the enhanced speech, that aids in masking the annoying signal distortions.

Most of these schemes have been designed to increase the quality, insofar as intelligibility and naturalness of corrupted speech are concerned. It is worthwhile to see research done in joint systems (speech enhancement and low-bit rate coders)—and this is the primary focus of the next subsection.

1.3 Previous Related Work

As mentioned earlier, drastic degradation of LPC parameter estimation, in the presence of acoustic noise, is of major concern. Adding one of the above techniques, as a pre-processing stage to parametric coders, may not suffice for correct LPC parameter estimation. Some of the early work done in this area is seen in [10, 11, 12, 13, 14].

In 1999, Martin and Cox, [15], proposed a modification to Ephraim and Malah's MMSE-LSA algorithm, [8]. They observed that while spectral valleys in between formant frequencies are relatively unimportant for speech *perception*, they are important for LPC *estimation*, [15]. Their modification to MMSE-LSA was aimed at removing more noise from the spectral valleys, thereby improving LPC estimation. However, removing more noise from the valleys has the disadvantage of leaving less residual noise that generally aids in masking the low level distortions. Therefore it seems germane to have one denoising algorithm that will consider accurate LP estimation, while another algorithm that will improve the perceptual quality of the degraded speech. This gives rise to a *two-branch* pre-processing system. In fact, in 1999 Accardi and Cox published a paper, [16], where they consider such a system.

1.4 Thesis Contribution

This thesis attacks the problem of enhancing corrupted speech prior to coding. The goals of the thesis are two-fold:

1. It is desired to tailor the MMSE-LSA algorithm for improved perception of corrupted speech. The new algorithm is derived on the basis of the heuristic approach taken by Martin and Cox in [15]. The underlying cost-function for the new algorithm is subjective quality. This newly derived algorithm is referred to as the MMSE-LSA with Adaptive Limiting Scheme for Perception (ALSP). Details of the derivation are seen in Chapter 3.

2. Several denoising algorithms are tested in each of the two-branches. The performance of the system is evaluated based on subjective and objective measures. These results show the effectiveness of using a two-branch speech enhancement scheme as a preprocessor for low bit-rate speech coding. Some results are also conducted for several noisy environments (such as Hoth[4], office babble, car noise and music) under varying acoustic conditions (i.e., different SNR values).

1.5 Thesis Synopsis

Chapter 2 starts by elaborating basic assumptions and 'tools' necessary for noise suppression algorithms. It then looks at the noise suppression block defined by EVRC. This chapter further details the STSA algorithms (MMSE-STSA and MMSE-LSA) defined by Ephraim and Malah. Some distortion measures (both objective and subjective) that are generally used to qualify the efficiency of noise suppression rules are also studied in this chapter. Some such measures are minimum Euclidean distance between Line Spectral Frequencies (LSFs), percentage correct pitch prediction, SNR and the ITU-T recommendation, P.862—PESQ.

Chapter 3, studies the low-bit rate speech coder—MELP. This chapter also discusses the two-branch systems that aim at improving speech quality of encoded speech in the presence of background noise. In this chapter, a new adaptive scheme is developed that is aimed at maximizing the Mean Opinion Score (MOS) of coded speech.

[4] Hoth noise is that which simulates the acoustic background noise at an average telephone subscriber location, [17].

The purpose of Chapter 4 is to present the test procedure adopted to evaluate the performance of the speech enhancement algorithms that are used as pre-processors to the MELP speech coder. Subjective results based on the A–B Comparison test are used to draw conclusions on the efficiency of the proposed algorithm. These results are further tested for other noisy environments by using some of the objective measures listed in Chapter 2.

In Chapter 5, a brief summary of the thesis, along with potential future work is presented.

Chapter 2

Speech Enhancement

The previous chapter introduced the subject of this thesis—the problem of additive background noise in low-bit rate speech coding. The focus of this chapter is on the importance of Short-Time Spectral Amplitude (STSA) estimate of speech. Section 2.6, summarizes some of the distortion measures that are used to qualify these noise suppression rules.

2.1 Speech Enhancement Basics

The basic structure of STSA speech enhancement system is given in Fig. 2.1. This figure highlights three stages that are necessary in such systems:

1. A spectral analysis/synthesis system, Section 2.1.4.
2. Noise estimation algorithm, Section 2.1.5 and Section 2.1.6.
3. A spectral gain computation algorithm, Section 2.2, Section 2.3,and Section 2.4.

Let $x[n]$ and $d[n]$ denote the speech and the noise processes, respectively. The observed noisy signal, $y[n]$, is given as:

$$y[n] = x[n] + d[n] \qquad 0 \leq n \leq N - 1. \tag{2.1}$$

The objective of a speech enhancement block is to estimate $x[n]$. This is generally accomplished on a frame-by-frame basis by applying a unique gain to each of the frames of $y[n]$. These gains are computed (in either frequency or time domain) by minimizing or

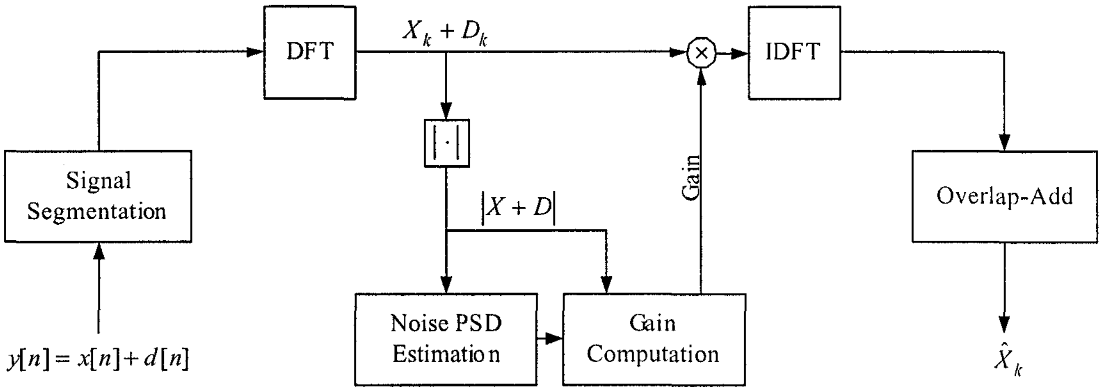


Fig. 2.1 Speech enhancement system description.

maximizing a *cost function.* For instance, Signal-to-Noise Ratio (SNR) is one such cost function. Defining a gain that maximizes the SNR of the output enhanced signal is one suitable criterion.

Let $X_k \triangleq A_k \exp(j\alpha_k)$, $Y_k \triangleq R_k \exp(j\vartheta_k)$ and D_k denote the kth spectral component of the original clean signal $x[n]$, the noise, $d[n]$ and the noisy observation $y[n]$, respectively, in the analysis frame interval $[0,\ N-1]$. A_k, R_k and D_k denote the spectral magnitude of the clean signal, noisy signal and the noise only observations, respectively. α_k and ϑ_k, respectively, denote the phase of the clean and the noisy signal. Recall that Y_k (and likewise X_k and D_k) are given by:

$$Y_k = \sum_{n=0}^{N-1} y[n] \exp\left(-j\frac{2\pi}{N}kn\right) \qquad k = 0, 1, 2, \ldots N-1 \tag{2.2}$$

With these definitions, the problem reduces to estimating the modulus of X_k from the corrupted signal $\{y[n],\, 0 \leq n \leq N-1\}$. In earlier works, it has been noted that the speech phase, α, shows up as a nuisance parameter (i.e., it provides no useful information) in noise suppression and as such is ignored in the estimation problem. Hence, once the modulus of speech is estimated it can be combined with the complex phase of the noisy speech.

In all speech enhancement algorithms assumptions are made to ease the computational and analytical complexity involved in the derivation of the gain block. Some such assumptions are discussed in the following subsection.

2.1.1 Assumptions on Speech and Noise Parameters

In order to derive an estimator for A_k, the *a priori* probability distribution of the speech and noise Fourier expansion coefficients should be known. Generally, the speech and possibly the noise are neither stationary nor ergodic[1] processes, thereby excluding the convenient possibility of obtaining the statistics of Fourier coefficients by examining the long-term behaviour of each process. This problem can be resolved by assuming an appropriate statistical model.

Gaussian Statistical Model

It is sometimes assumed that the complex Fourier expansion coefficients are statistically independent Gaussian random variables. The mean of each coefficient is assumed to be zero, while the variance is assumed time-varying, due to speech non-stationarity. A well known fact about *independent* Gaussian random variables is that they are *uncorrelated* as well [19]. This eases the mathematical formulation of the problem of removing background noise from corrupted speech.

2.1.2 Appropriate Analysis Frame Length

Long analysis frames are good, as they provide better statistical averages. However, due to time-varying characteristics of speech, a shorter window is sometimes desired. A convenient way to tackle this trade-off is to consider the durations of typical phonemes[2] in speech. A typical vowel (voiced phoneme) ranges between 50–400 ms, while a plosive may last for about 10 ms. Changes in the shape of the speech signal, whether gradual or abrupt, result from movements of the vocal tract articulators, which rarely stay fixed in position for more than 40 ms at a time, [20]. Thus in most works, the analysis frame length, T, is usually chosen in the 16–40 ms range (or about 128–320 samples, for speech sampled at 8 kHz). Note that even if a smaller frame length were considered appropriate (to increase temporal resolution at the cost of spectral resolution), statistical independence would still be assumed.

[1] For definition of either stationarity or ergodicity refer to [18].

[2] A phoneme is the smallest phonetic unit in a language that is capable of conveying a distinction in meaning, as the *m* of 'mat' and the *b* of 'bat'.

2.1.3 Input Conditioning: High Pass Filter

Initial pre-processing is needed to condition the input signal against excessive low frequency and other background disturbances that might degrade the quality of a speech codec. Low frequency disturbances include hum at 60 Hz and its harmonics at 120, 180 and 240 Hz. Therefore it is desired to pre-process input narrowband speech[3] (sampling frequency of 8 kHz) with a high-pass filter[4] that will eliminate low frequency noise. In the case of STSA algorithms mentioned in this thesis, speech is first filtered with the high-pass filter prior to parameter extraction and further processing.

2.1.4 Analysis Window Type

Some of the commonly used windows in speech processing are symmetric (e.g., Hamming and Hanning windows) or asymmetric (such as the hybrid Hamming-Cosine window). The goal of asymmetric windows is to reduce the algorithmic delay in speech coders. In this thesis, symmetric windows are used.

Ideally, the window spectrum should have a narrow main-lobe and small side-lobes. However, there is an inherent trade-off between the width of the main-lobe and the side-lobe attenuation. A wide main-lobe will average adjacent frequency components and large side-lobes will introduce contamination (or spectral leakage) from other frequency regions, [22]. Fig. 2.2 shows the spectrum of three windows. It is seen that the main lobe for rectangular window is narrower than that of the hanning window, while its sidelobes are higher.

The third window type, is the smoothed Trapezoidal window. This is used in the EVRC

[3] Narrowband speech may not sound as clear as wideband speech (sampling frequency of 16 kHz, but is intelligible enough to provide toll-quality. Toll-quality is voice quality equal to that of the long-distance public switched network, which charged a toll for each minute of use. Thus, in this thesis, input speech is sampled at 8,000 Hz.

[4] The Enhanced Variable Rate Coder (EVRC) deploys a sixth order Butterworth filter having a 3 dB cutoff frequency of 120 Hz and providing −36 dB/octave gain below 120 Hz is deployed, [21]. The filter coefficients are seen in [9]. It is implemented as three cascaded biquadratic sections and contributes virtually no passband delay.

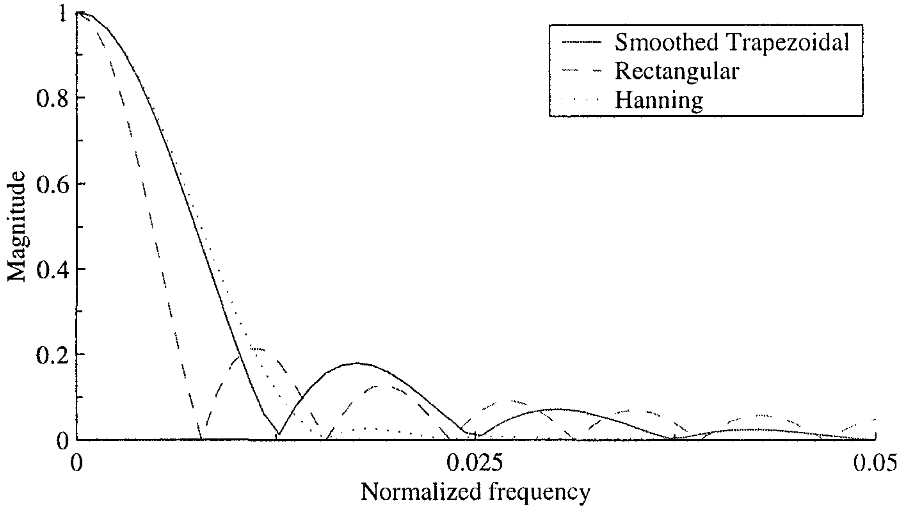


Fig. 2.2 Frequency Response for three symmetric windows. Solid: smoothed Trapezoidal; Dashed: Rectangular; and Dotted: Hanning Window.

noise suppression rule, Section 2.2, and is mathematically expressed as:

$$w(n) = \begin{cases} \sin^2(\pi(n+0.5)/2D), & 0 \leq n < D, \\ 1, & D \leq n < L, \\ \sin^2(\pi(n-L+D+0.5)/2D), & L \leq n < D+L, \\ 0, & D+L \leq n < M, \end{cases} \tag{2.3}$$

where n is the sample index, L is the frame length and D is the overlap length.

In this thesis the analysis window of choice is the smoothed Trapezoidal window. Based on the argument presented in Section 2.1.2, an appropriate window length is chosen. A window of the specified length multiplied with the speech signal produces a *frame of speech* of that length. Speech frames are frequency-transformed for analysis and denoised using a noise suppression rule. Signal reconstruction is performed by the traditional overlap-add method without further windowing.

2.1.5 Noise Power Spectrum Estimation

The basic concern of this subsection is that of estimating the Power Spectral Density (PSD) of noise from the observation of noisy signal over a finite time interval. Frames of "pure" noise are identified using a *Voice Activity Detector*, Section 2.1.6. If $d(n)$ is a stationary random process, its autocorrelation function is:

$$\gamma_{dd}(m) = E\{d^*(n)d(n+m)\}, \tag{2.4}$$

where $E\{\cdot\}$, denotes the statistical average. Then, via the Wiener-Khintchine theorem, [19], the PSD for a wide-sense stationary process is given as the Fourier transform of the autocorrelation function. Since $\gamma_{dd}(m)$ is usually unknown, its Fourier transform cannot be computed. However, from a single realization of the random process, the time-average autocorrelation function, $\widehat{\gamma}_{dd}$, can be computed:

$$\widehat{\gamma}_{dd}(m) = \frac{1}{N}\sum_{n=0}^{N-m-1} d^*(n)d(n+m) \qquad m = 0, 1, \ldots, N-1, \tag{2.5}$$

where N is the observation time.

The corresponding PSD is given as the Fourier transform:

$$P_{dd}(f) = \sum_{m=-(N-1)}^{N-1} \widehat{\gamma}_{dd}(m)e^{-j2\pi fm}. \tag{2.6}$$

Substituting Eq. (2.5) into Eq. (2.6) gives an estimate for $P_{dd}(f)$ in terms of the Fourier transform, $D(f)$, of the sample sequence $d(n)$:

$$P_{dd}(f) = \frac{1}{N}\left|\sum_{n=0}^{N-1} d(n)e^{-j2\pi fn}\right|^2 = \frac{1}{N}|D(f)|^2. \tag{2.7}$$

This well known form of PSD is called the *periodogram*. However, as pointed in [22], this method suffers from bias—producing an inconsistent estimate of true PSD. Other classical non-parametric methods, which make no assumptions about data sequence are generally deployed. One such method is the *Welch periodogram*.

In Welch's method, two modifications, over the existent periodogram, are apparent. Firstly, the data is split into K *overlapping* segments. For each segment the periodogram is computed and is averaged over the number of segments to obtain the final PSD estimate. This reduces the variance of the estimate by a factor of K. Secondly, the segments are windowed prior to computing the autocorrelation function. The effect of this operation is to reduce the frequency resolution by a factor of K, and hence a further reduction in noise estimate variance. In this thesis, the Welch method (with the Trapezoidal window) is used to obtain an estimate of noise PSD.

2.1.6 Voice Activity Detector

Voice Activity Detector (or VAD) returns a '1' in the presence of speech and '0' in absence thereof. Conceptually, such a binary decision is based on some measured or extracted feature of speech compared against some pre-defined (or adaptively changing) thresholds (generally extracted from noise only frames). A VAD must be robust as its accuracy is critical in noise suppression algorithms. Misclassifying speech as noise will erroneously remove speech or result in a poor estimate of noise.

Some of the early VAD algorithms relied on short-time energy, zero crossing rate, and LPC coefficients. Recent work employs Cepstral features, formant shape and a least-square periodicity procedure, [23]. In this thesis, the VAD used in EVRC is considered and a brief outline of it is presented in the sequel.

The VAD used in the EVRC is energy-based, and is an inherent part of the coder's noise suppression block. The input frame is split into two bands and the energy in the bands is compared against two thresholds. Speech is detected if the energy in each band is greater than the corresponding lowest threshold. Other VAD algorithms employ a similar strategy, but the novel part of the EVRC VAD is its dynamic threshold update capability. In order to be able to respond to changes in the background noise level (e.g., if a wireless phone is being used in a car, and the user lowers the window), there is a provision for a forced update of the band noise (and also the corresponding threshold) if more than 35 blocks of noise have elapsed since the last update, [9].

2.1.7 Intuitive Approach to Noise Suppression

A typical speech file contains several pause (between words) and/or silence (speaker stops talking) segments. In speech enhancement, the noise variance is updated during such pauses or silence. A VAD is used to discriminate between speech and silence/pause a VAD is employed. For denoising speech files corrupted with additive white Gaussian noise, one straightforward idea would be to subtract off the estimated noise variance, $(\widehat{D}_k)$, from the power spectrum of the observed noisy signal, (Y_k), to obtain an estimate of the modulus of speech power spectrum $(\widehat{X}_k)$. Mathematically this is represented as:

$$|\widehat{X}_k|^2 = |Y_k|^2 - |\widehat{D}_k|^2. \tag{2.8}$$

However, there are limitations to this subtraction rule, see Section 2.5, and as such the basic problem has been tackled by deriving several fundamentally and theoretically justified noise suppression rules.

2.2 Enhanced Variable Rate Coder Noise Suppression

EVRC is a standard coder for use with the IS-95x Rate 1 air interface (CDMA), [9, 21]. The noise suppression algorithm is an independent module and can thus can be tested separately or as part of other systems. Fig. 2.3 shows an overview of the noise suppression module.

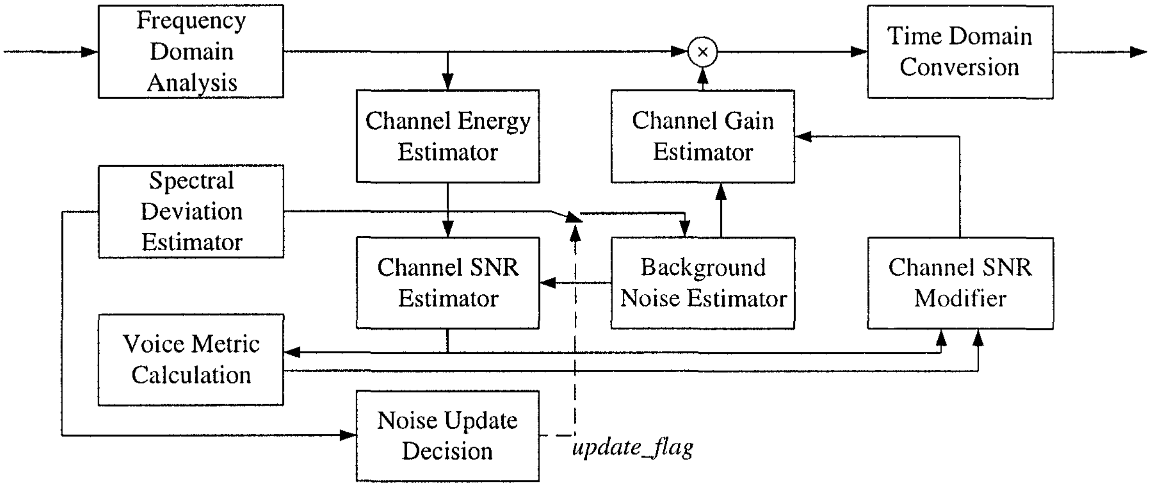


Fig. 2.3 EVRC Noise Suppression Block, redrawn from [21].

The noise suppression filter is implemented block-by-block in the frequency domain using the overlap-add method. The filters are run and updated once each 10 ms (or 80 samples for speech sampled at 8 kHz). A 104 sample vector is formed using 80 new samples of the speech sequence and 24 samples from the end of the last 80 sample block. This vector is pre-emphasized (using the high-pass filter described in footnote 4) and weighted using the smoothed Trapezoidal window, appended with 24 zeros and transformed using a 128-

point FFT. This produces the frequency transformed signal for the nth analysis frame or block, [21].

Essentially, the noise suppression algorithm sums the output of a bank of adaptive filters that span the frequency content into sixteen bands. The width of the bands increases logarithmically. This mimics the frequency resolution of the human ear. The energy present in each of the sixteen bands is estimated by computing the mean magnitude for all frequency bins within that band, $E_b(n,i)$ (where i is one of the sixteen bands, b). The Signal-to-Noise Ratio (SNR) for each band is computed as the ratio of the band energy estimate $(E_b(n,i))$ to the noise energy estimate in that band $(E_N(n,i))$.

The sixteen band gains are computed based on the total noise energy estimate, $\gamma_N(n)$:

$$\gamma_N(n) = -10\log_{10}\left(\sum_{i=0}^{15} E_N(n,i)\right), \tag{2.9}$$

where i is one of sixteen bands in the nth analysis frame. Each of the sixteen band SNR values are quantized and constrained to lie within 6 and 89 dB and are denoted by $\sigma_Q^{''}$.

With the use of Eq. (2.9), individual band gains $(\gamma(b))$, in dB, are computed to be:

$$\gamma_{dB}(b) = \mu_g(\sigma_Q^{''}(c) - 6) + \gamma_N, \tag{2.10}$$

where $\mu_g = 0.39$ is the gain slope. These band gains are constrained to lie in the range -13 to 0 dB, [21]. Eq. (2.10) is converted to sixteen linear band gains:

$$\gamma(b) = \min(1, 10^{\gamma_{dB}(b)/20}). \tag{2.11}$$

These gains are applied to the noisy frequency bands, to which they belong. The result is a 'denoised' frame, that is inverse frequency transformed, combined with the phase of the noisy signal and overlap-added to construct the enhanced speech.

2.3 MMSE of Short-Time Spectral Amplitude

The previous section described the noise suppression algorithm used in the EVRC standard. The purpose of this section is to study the Short-Time Spectral Amplitude (STSA) estimator. The goal of the STSA algorithm is to estimate the modulus of each complex Fourier expansion coefficient of the speech signal in a given analysis frame of noisy speech. Fig. 2.4, shows an overview of the basic steps for enhancement schemes employing STSA estimator methodology. The first two steps, indicated in Section 2.1, are common to all STSA algorithms.

In this section a brief derivation of the MMSE-STSA estimator, proposed by Ephraim and Malah, [24], is outlined. As the name suggests, it estimates $\widehat{A}_k$ by minimizing the mean-squared error with A_k. There are two parts to the algorithm:

- Deriving the MMSE STSA estimator, based on modelling speech and noise spectral components as statistically independent Gaussian random variables, Fig. 2.4.
- Computing a *multiplicative modifier* under uncertainty of signal presence in noisy environment, Fig. 2.6.

2.3.1 Derivation of the MMSE-STSA Estimator

From estimation theory, [25], it is quoted that:

> The minimum mean-square error estimate is always the mean of the *a posteriori* density (the conditional mean).

Hence, $\widehat{A}_k$ is the conditional mean of A_k given Y_k is observed. More conveniently this is expressed as:

$$\begin{aligned}\widehat{A}_k &= E\{A_k|Y_k\} \\ &= \frac{\displaystyle\int_0^{\infty}\int_0^{2\pi} a_k p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k da_k}{\displaystyle\int_0^{\infty}\int_0^{2\pi} p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k da_k},\end{aligned} \tag{2.12}$$

where $E\{\cdot\}$ is the expectation operator and $p(\cdot)$ is the Probability Density Function (PDF). Since the observed data is assumed to be Gaussian distributed, Eq. (2.12) and can be expanded in terms of $p(Y_k|a_k,\alpha_k)$ and $p(a_k,\alpha_k)$:

$$p(Y_k|a_k,\alpha_k) = \frac{1}{\pi\lambda_d(k)} \exp\left\{-\frac{1}{\lambda_d(k)}|Y_k - a_k e^{j\alpha_k}|^2\right\}, \tag{2.13}$$

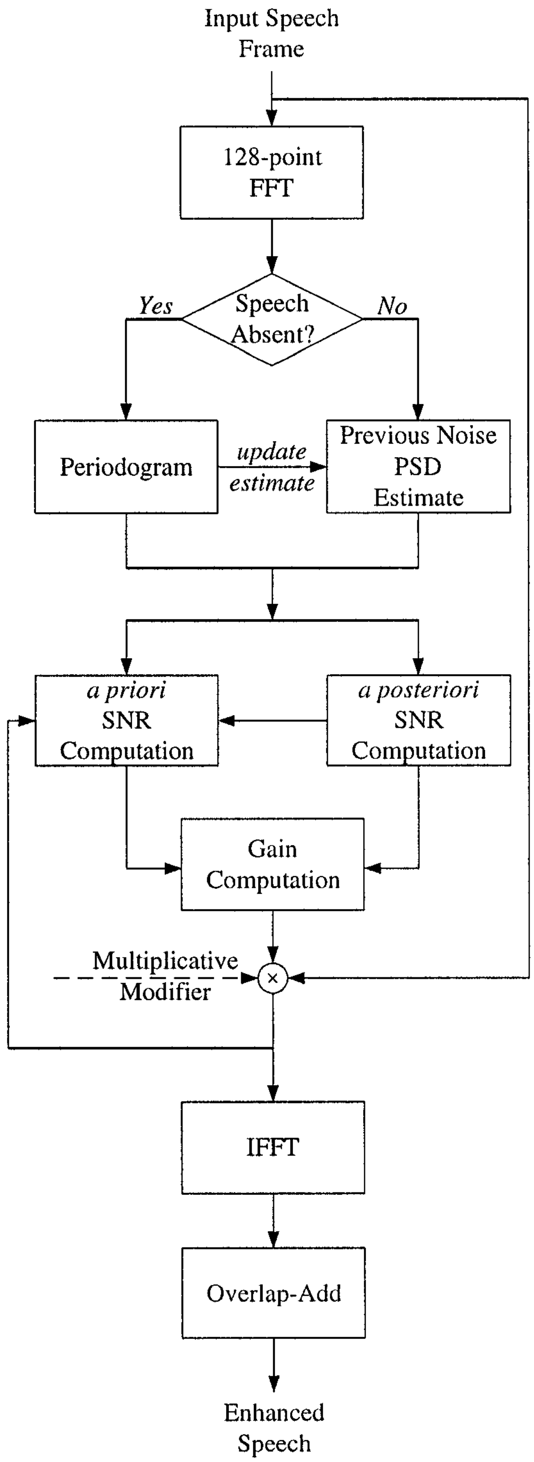


Fig. 2.4 Algorithm for MMSE-STSA.

$$p(a_k, \alpha_k) = \frac{a_k}{\pi\lambda_x(k)} \exp\left\{-\frac{a_k^2}{\lambda_x(k)}\right\}, \tag{2.14}$$

where $\lambda_x(k) \triangleq E\{|X_k|^2\}$, and $\lambda_d(k) \triangleq E\{|D_k|^2\}$, are defined as variances of the kth spectral component of the speech and noise, respectively. Simple substitution of Eq. (2.13) and Eq. (2.14) in Eq. (2.12) gives the desired gain function for the MMSE-STSA estimator, [24]:

$$G_{\text{MMSE}}(\nu_k) = \Gamma(1.5)\frac{\sqrt{\nu_k}}{\gamma_k}\exp\left(-\frac{\nu_k}{2}\right)\cdot\left[(1+\nu_k)I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right], \tag{2.15}$$

where $\Gamma(\cdot)$ is the Gamma function (with $\Gamma(1.5) = \sqrt{\pi}/2$) and $I_0(\cdot)$ and $I_1(\cdot)$ are the zeroth and first order modified Bessel functions, respectively, defined as:

$$I_n(z) \triangleq \frac{1}{2\pi}\int_0^{2\pi}\cos(\beta n)\exp(z\cos\beta)\,d\beta. \tag{2.16}$$

In Eq. (2.15), ν_k is defined as:

$$\nu_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k. \tag{2.17}$$

The gain function, (Eq. (2.15)), is thus, solely parameterized by ξ_k and γ_k. These are interpreted as *a priori* SNR and *a posteriori* SNR values, respectively:

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)}. \tag{2.18}$$

$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)}. \tag{2.19}$$

Essentially, *a priori* SNR is the Signal-to-Noise Ratio of the kth spectral component of the "clean" speech signal, $x[n]$, while *a posteriori* SNR is the kth spectral component of the corrupted signal, $y[n]$. Computation of γ_k is straightforward—ratio of the variance of the noisy speech signal to the estimated noise variance. However, computation of *a priori* SNR is more involved, especially since the knowledge of "clean" signal is seldom available in real systems. *"Decision-Directed"* estimation and Maximum Likelihood estimation are two approaches taken to compute *a priori* SNR, [24]. Owing to certain shortcomings of the latter approach, decision-directed has been exploited in this thesis and is addressed next.

"Decision-Directed" Estimation Approach to compute a priori SNR

From Eq. (2.1) and the Fourier expansion definitions, it is seen that the expected value of the variance of noisy data can be expressed as:

$$E\{R_k^2\} = E\{A_k^2\} + 2E\{A_k\}E\{D_k\} + E\{D_k^2\}. \tag{2.20}$$

Since, the clean speech is assumed uncorrelated to noise, Section 2.1.1, the cross-term is set to zero. This simplifies the analysis and results in the *instantaneous SNR*:

$$\xi_k(n) = E\{\gamma_k(n) - 1\}. \tag{2.21}$$

for the nth analysis frame.

Ephraim and Malah exploit this equivalency of relation between Eq. (2.18) and Eq. (2.21) to propose an estimate for the *a priori* SNR, $\widehat{\xi_k}$:

$$\widehat{\xi_k} = \alpha \frac{\widehat{A}_k^2(n-1)}{\lambda_d(k-n-1)} + (1-\alpha)P\{\gamma_k(n) - 1\}, \qquad 0 \leq \alpha \leq 1, \tag{2.22}$$

where $\widehat{A}_k(n-1)$ is the amplitude estimator of the kth signal spectral component in the $(n-1)$th analysis frame (and hence the feedback in Fig. 2.4 from the multiplier to the *a priori* SNR Computation block) and α is a weighting constant that is deduced from experimental data. The "best" value for α is found to be 0.9, [24]. $P\{\cdot\}$, is an operator defined by:

$$P\{x\} = \begin{cases} x, & x \leq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2.23}$$

where the positive operator, $P\{\cdot\}$, is defined to ensure that the *a posteriori* SNR is less than or equal to one, Eq. (2.22).

Parametric Gain Curves defining MMSE-STSA

From Eq. (2.15) it is seen that the gain depends on ξ_k and γ_k. Several parametric gain curves are seen in Fig. 2.5, where gain is plotted against instantaneous SNR. The gain curves show an increase in gain as γ_k decreases, while maintaining ξ_k constant. This is explained next in the light of the MMSE-STSA estimator compromising between what it knows *a priori* and what it learns from the noisy data.

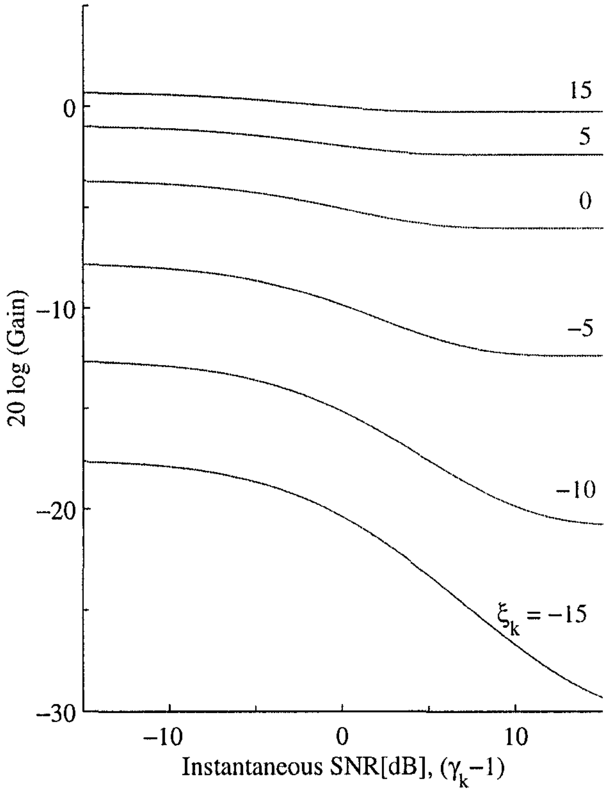


Fig. 2.5 Parametric gain curves describing MMSE-STSA gain function $G_{\text{MMSE}}(\xi_k, \gamma_k)$ defined by Eq. (2.15) for various values of ξ_k (in dB).

For fixed values of $\lambda_d(k)$ and $\lambda_x(k)$, ξ_k is also fixed, Eq. (2.18). The fixed value of $\lambda_x(k)$ determines the most probable realization of $\widehat{A}_k$, which is considered by the estimator (since $\lambda_x(k)$ parameterizes the PDF of a_k, Eq. (2.14)). From Eq. (2.19), it is seen that γ_k is proportional to $\widehat{R}_k$. Thus, for fixed ξ_k and decreasing γ_k, the estimator should compromise between the most probable realization of A_k and decreasing values of R_k, and this is done by increasing G_{MMSE} (since $\widehat{A}_k \triangleq G_{\text{MMSE}} \cdot R_k$).

The derivation assumes that signal *is* present in noisy observation. In reality, there is absolutely no assurance of signal presence in the observed noisy data. Consider the case when the speech energy is less than that of noise. In this case it would be desirable to have a decreasing gain (to attenuate more noise as speech is weakly present) for decreasing γ_k when ξ_k is high (i.e., when $\lambda_d(k)$ is larger than $\lambda_x(k)$). The concern of Section 2.3.2, is to address this issue and describe the Multiplicative Modifier algorithm suggested in [24, 26].

2.3.2 Multiplicative Modifier based on Uncertainty of Speech Presence

Signal absence in noisy observations $\{y[n], 0 \leq n \leq N\}$ is frequent, as speech signals generally contain large portions of silence, [24]. Nevertheless, it does not mean that speech is *never* present in noisy sections. Two statistical models exist for speech absence in noisy observations. The first model assumes that speech is either present or absent, with predefined probabilities. This is a legitimate model under the reasoning that signal absence or presence should be the same over the finite analysis interval. The second model is based on statistically independent random appearance of signal in noisy spectral components.

The MMSE estimator which accounts for uncertainty of speech presence in noisy observation was first developed by Middleton and Esposito, [27] and is based on the second model mentioned above. This model will be used in this thesis.

In Fig. 2.6, steps involved in the derivation of the multiplicative modifier are seen. The spectral components of a speech frame and *a posteriori* SNR are computed earlier as in Fig. 2.4. Derivation of the "likelihood ratio computation" and "*a priori* probability of speech absence" blocks is addressed below.

Since speech in noisy observation is either present or absent, a two-state model based on binary hypothesis testing, [25], has been used by McAulay and Malpass, [7], to derive the "likelihood ratio". The hypotheses under test are:

- Null Hypothesis, H_0^k: speech *absent*: $|R_k| = |D_k|$.
- Alternate Hypothesis, H_1^k: speech *present*: $|R_k| = |A_k + D_k|$.

In view of these hypotheses, Eq. (2.12), can be re-written more explicitly as:

$$\widehat{A}_k = E\{A_k|Y_k, H_1^k\}P(H_1^k|Y_k) + E\{A_k|Y_k, H_0^k\}P(H_0^k|Y_k), \tag{2.24}$$

where $P(H_i^k|Y_k)$, (for $i = 0, 1$), is the probability that the speech in state H_i^k, for the kth spectral component, given that the measured envelope is Y_k. Obviously, $E\{A_k|Y_k, H_0^k\}$ is zero as it represents the average value of A_k in Y_k when speech *is* absent (the null hypothesis). Hence, Eq. (2.24), can be simplified to:

$$\widehat{A}_k = E\{A_k|Y_k, H_1^k\}P(H_1^k|Y_k). \tag{2.25}$$

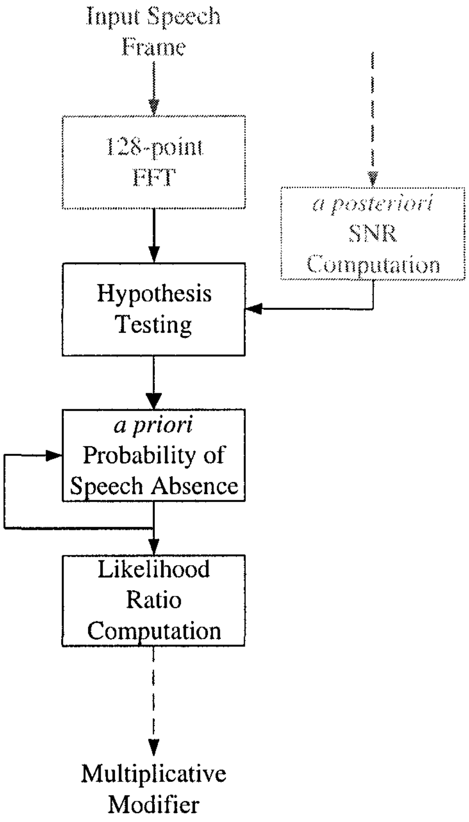

Input Speech Frame
128-point FFT
a posteriori SNR Computation
Hypothesis Testing
a priori Probability of Speech Absence
Likelihood Ratio Computation
Multiplicative Modifier

Fig. 2.6 Algorithm for uncertainty of speech presence.

Note that $E\{A_k|Y_k, H_1^k\}$ replaces $E\{A_k|Y_k\}$, identically, in Eq. (2.12) with an added assertion of the alternate hypothesis. Hence, $P(H_1^k|Y_k)$ defines the *multiplicative modifier on the optimal estimator under the signal presence hypothesis.*

Exploiting Bayes' rule, [25], one obtains:

$$P(H_1^k|Y_k) = \frac{\Lambda(k)}{1+\Lambda(k)} = G_{\mathrm{MM}}(k), \tag{2.26}$$

where

$$\Lambda(k) \triangleq \mu_k \frac{p(Y_k|H_1^k)}{p(Y_k|H_0^k)} \qquad \mu_k \triangleq \frac{P(H_1^k)}{P(H_0^k)} = \frac{1-q_k}{q_k}. \tag{2.27}$$

$\Lambda(k)$ is the likelihood ratio while q_k denotes the *a priori* probability of speech absence in the kth bin.

Apart from knowledge of ξ_k and γ_k, to compute G_{MM}, reliable estimation of prior q_k (for all frequency bins in a given noisy frame) is essential. Malah, Cox and Accardi in 1999, proposed another set of binary hypotheses, [26], for the estimating q_k:

- Null Hypothesis, $\mathcal{H}_0$: speech *present* in the kth bin: $\eta_k \geq \eta_{min}$.
- Alternate Hypothesis, $\mathcal{H}_A$: speech *absent* in the kth bin: $\eta_k \leq \eta_{min}$.

where $\eta_k \triangleq (1 - q_k)\xi_k$, is the new *a priori* SNR, as it only makes sense to give it a value when speech is present, with probability $(1 - q_k)$, in the spectral analysis frame. It is also observed that the PDF of γ_k is parameterized by η_k. This means, that knowledge of γ_k is sufficient for estimating the priors q_k, [26].

The Neyman-Pearson decision rule [28] can be exploited to establish the *uniformly most powerful test*[5] for the problem at hand:

$$\gamma_k \underset{\mathcal{H}_A}{\overset{\mathcal{H}_0}{\lessgtr}} \gamma_{th}, \tag{2.28}$$

where γ_{th} is a threshold to satisfy a desired significance level.

Based on Eq. (2.28), a binary decision is issued to I_k ($I_k = 1$ if $\mathcal{H}_0$ is rejected and $I_k = 0$ if accepted) and used on speech only frames (detected using a VAD) to compute q_k for all k frequency bins in that frame:

$$q_k = \alpha_q q_{k-1} + (1 - \alpha_q)I_k, \tag{2.29}$$

where $\gamma_{th} = 0.8$ and $\alpha_q = 0.98$ have been used based on heuristic results [26].

Fig. 2.7 shows plots of the effect of the multiplicative modifier on MMSE-STSA estimator. Two inferences can be drawn:

1. The decrease in gain as γ_k decreases, for high η_k, is in contrast to the increase in gain (Fig. 2.5). This is explained in the realm of the estimator favouring $\mathcal{H}_A$.
2. The gain decreases as q_k (probability of signal absence) increases. This is intuitive,

[5] If $\boldsymbol{\theta}$ is a random variable with *unknown* density, a Bayes test is not meaningful, [25]. Let χ be the parameter space for $\mathcal{H}_0$ and $\mathcal{H}_A$. Then, it is shown that the Neyman-Pearson tests are the Uniformly Most Powerful (UMP) tests, as the likelihood ratio test for every $\boldsymbol{\theta} \in \chi$ can be completely defined (including the threshold) without knowledge of $\boldsymbol{\theta}$.

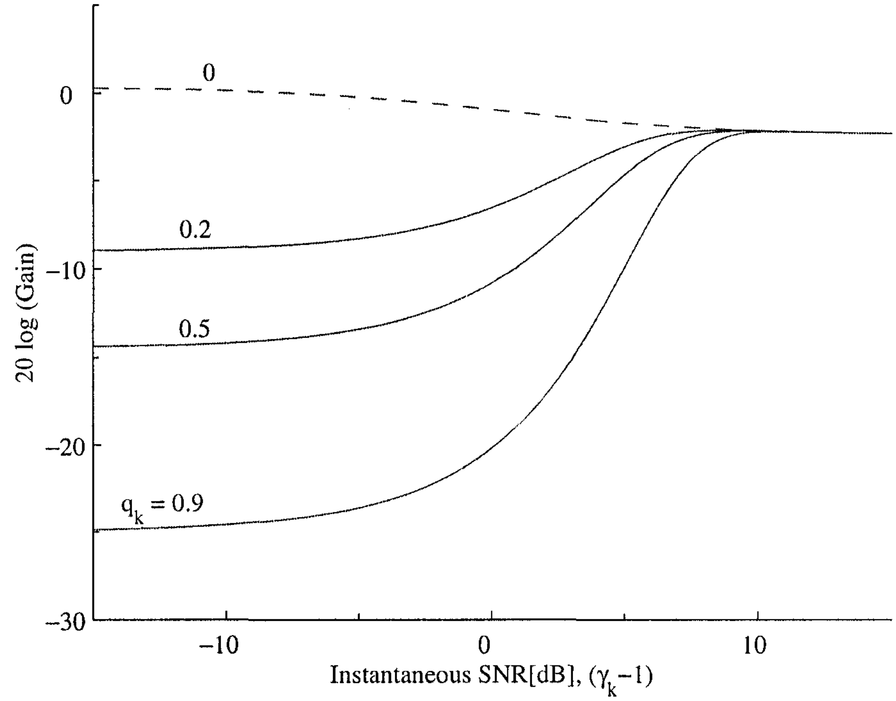


Fig. 2.7 Effect of q_k on Gain for a given ξ_k and varying Instantaneous SNR (γ_k-1).

since the estimator tends to attenuate more noise if the probability of speech absence is high at low values of instantaneous SNR.

Thus, the effect of adding the multiplicative modifier to MMSE-STSA estimator is to enhance speech with prior knowledge of speech presence or absence in each spectral bin.

Ephraim and Malah claim that using the MMSE estimator, which takes into account the uncertainty of signal presence, results in a whitened noise, [24]. However, as will be seen in Section 2.4, modifications to MMSE-STSA estimator lead to a further reduction of this residual noise—making it sound more uniform in silence/pause segments of speech.

2.4 MMSE Log-Spectral Amplitude Estimator

A year after they proposed the MMSE-STSA cost function, Eq. (2.12), Ephraim and Malah looked at MMSE of the *log* spectra in enhancing noisy speech, [8]. The wide-spread use of log-spectra in distortion measures, [29], is the leading motivation to examine the effect of amplitude estimator constrained at minimizing the mean-squared error of the log-spectra. This section will briefly summarize the derivation of the Minimum Mean Square Error Log-Spectral Amplitude (MMSE-LSA) estimator. As much as possible, the same notation, as used in Section 2.3, is used to formulate the MMSE-LSA estimation problem.

2.4.1 Derivation of the MMSE-LSA Estimator

With definitions given in Eq. (2.1) and their respective Fourier transforms, the driving constraint on finding the optimal realization, $\widehat{A}_k$, of the logarithmic A_k, is expressed as:

$$\widehat{A}_k = E\{(\log A_k - \log \widehat{A}_k)^2 | Y_k\}. \tag{2.30}$$

Hence, the estimator is easily shown to be:

$$\widehat{A}_k = \exp\{E[\ln A_k | Y_k\}, \qquad 0 \leq n \leq N-1]. \tag{2.31}$$

and it is independent of the basis chosen for the log.

The evaluation of Eq. (2.31), can be simplified with the *moment generating function*, [19]. For notational convenience, let $Z_k \triangleq \ln A_k$. Then the moment generating function $\Phi_{Z_k|Y_k}(\mu_k)$ of Z_k given Y_k can be expressed as:

$$\begin{aligned} \Phi_{Z_k|Y_k}(\mu_k) &= E\{\exp(j\mu_k Z_k)|Y_k\} \\ &= E\{A_k^{j\mu}|Y_k\}. \end{aligned} \tag{2.32}$$

The first derivative of $\Phi_{Z_k|Y_k}(\mu_k)$, evaluated at $\mu = 0$ generates the first moment of Z_k given Y_k, [19]. As seen earlier in Section 2.3.1, the right hand side of Eq. (2.12) now defines $\Phi_{Z_k|Y_k}(\mu_k)$. Assuming the same Gaussian model, as used previously, with the same definitions for *a priori* and *a posteriori* SNR, it is not too difficult to establish a gain function that satisfies Eq. (2.30)[6]:

$$G_{\text{LSA}}(\xi_k, \gamma_k) = \frac{\xi_k}{1+\xi_k} \exp\left\{\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\right\}, \tag{2.33}$$

where v_k and ξ_k have been previously defined in Section 2.3.1. Note that G_{LSA} is the gain

[6] For detailed derivations refer to [8]

function and is multiplied with the observed noisy data to obtain an estimate of A_k.

Ephraim and Malah did some subjective comparisons between MMSE-STSA and MMSE-LSA to show that the enhanced speech (using the latter estimator) suffers from much less residual noise, while there is no perceptible difference in the enhanced quality of speech itself. Further, the residual noise sounds more uniform (i.e., more *white*). In order to explain this phenomenon, it is worthwhile to look at the parametric gain curves of the present estimator compared to those of MMSE-STSA estimator. The general trend of the curves was explained in Section 2.3.1 and is applicable here as well. In Fig. 2.8 it is also seen that MMSE-LSA offers more attenuation (or lower gain values) for the corresponding instantaneous SNR than its counterpart, MMSE-STSA. This is readily seen by analyzing Jensen's inequality:

$$\widehat{A}_k = \exp\{E[\ln A_k|Y_k]\} \leq \exp\{\ln E[A_k|Y_k]\} = E[A_k|Y_k]. \tag{2.34}$$

From the inequality it is seen that the log estimate of $\widehat{A}_k$ is always less than or equal to the true estimate. Since, the MMSE-LSA works in the log domain to estimate the magnitude of A_k (as opposed to MMSE-STSA algorithm), it is easily seen why the MMSE-LSA gain plots are lower than those of the MMSE-STSA in Fig. 2.8.

The lower value of gain curves further aid in minimizing the effects of residual noise especially at low instantaneous SNR values.

If necessary, this estimator can be further modified by adding the multiplicative modifier—based on uncertainty of speech presence, Section 2.3.2. The enhanced speech, obtained thus, contains much less residual noise, while no difference in the enhanced speech itself is noticed.

Having studied the relevant STSA noise suppression rules, it is worth looking at limitations of these algorithms. Some such limitations are introduction of annoying artifacts and underlying speech distortion. Methods undertaken to reduce these effects are also seen in the following section.

2.5 Limitations of Noise Suppression Algorithms

If an accurate estimate of the noise variance were available, it would seem natural to remove noise completely from noise only or signal+noise frames to attain the enhanced "noise-free"

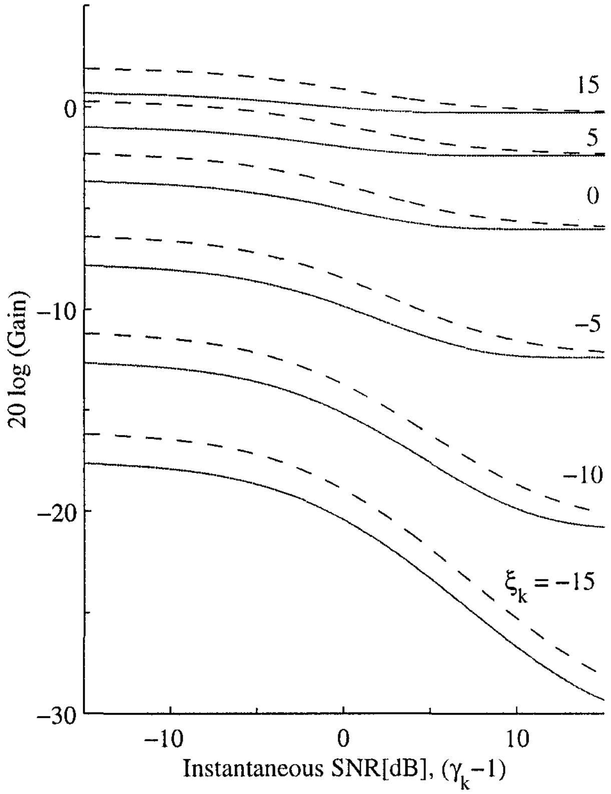


Fig. 2.8 Parametric gain curves describing $G_{\text{MMSE}}(\xi_k, \gamma_k)$ defined by Eq. (2.15) (solid) and G_{LSA} defined by Eq. (2.33) (dashed) for various values of ξ_k (in dB).

signal. However, there are severe underlying limitations to the amount of noise that can be removed. The aim of most enhancing algorithms is to remove most of the noise leaving comfortable residual noise with minimal speech distortion. Removing more noise than is necessary can produce choppy and distorted outputs. However, rèmoving less noise can have the counter effect of suppressing weak energy phonemes. Thus, there is an inherent trade-off that must be considered in most noise removal algorithms.

This section briefly looks at some of the limitations of noise suppression algorithms. Since, the noise variance is obtained as an estimate, clean speech cannot be exactly recovered. As noted earlier, proper estimation of noise variance is heavily dependent on accurate VAD decisions. If a frame is mistaken to be noise only, it will lead to signal cancellation. If a frame is mistaken to be high in speech energy (thus misclassifying it as clean speech segment), on the other hand, it might lead to incomplete speech enhancement. These distortions and artifacts (residual noise) can exacerbate auditory impression rather than enhancing it. Such distortions are especially critical when the input SNR is low. Some of the reported artifacts are:

- Musical noise,
- Timbre[7] changes and loss of signal components,
- Phase distortions and temporal smearing.

The following subsections are geared at describing these artifacts, and possible ways to reduce them.

2.5.1 Musical Noise

The most common introduced artifact is *Musical Noise*—short spurious bursts of isolated frequency components that appear randomly across enhanced speech's spectrum. They originate from the magnitude of short-time observed spectrum, $|Y_k|$, which shows strong fluctuations in low SNR areas—a well known feature of the periodogram. After passing such a frame through a spectral attenuator, the part of the spectrum that originally contained noise now contains succession of randomly spaced peaks corresponding to the maxima of $|Y_k|$. Between these peaks, the short-time spectrum values are strongly attenuated since they are close to or below the estimated noise variance. As a result the residual noise now contains low-energy spurious bursts of frequency components that appear randomly across the short-time frames, Fig. 2.9[8]. The presence of pure tones in residual noise gives rise to a "musical" perception.

Reduction of Musical Noise

Reduction of this type of artifact has been proposed by use of a *spectral floor*, which aids by masking the randomly scattered sinusoidal peaks. The spectral floor is a positive threshold value for the magnitude estimate. Therefore, the spectral floor can be considered as an *overestimate* of the spectral component variance. Although widely used, this method does

[7] The combination of qualities of a sound that distinguishes it from other sounds of the same pitch and volume.

[8] The plot seen in Fig. 2.9, was generated by sending noise only signal through the power subtraction noise suppression rule.

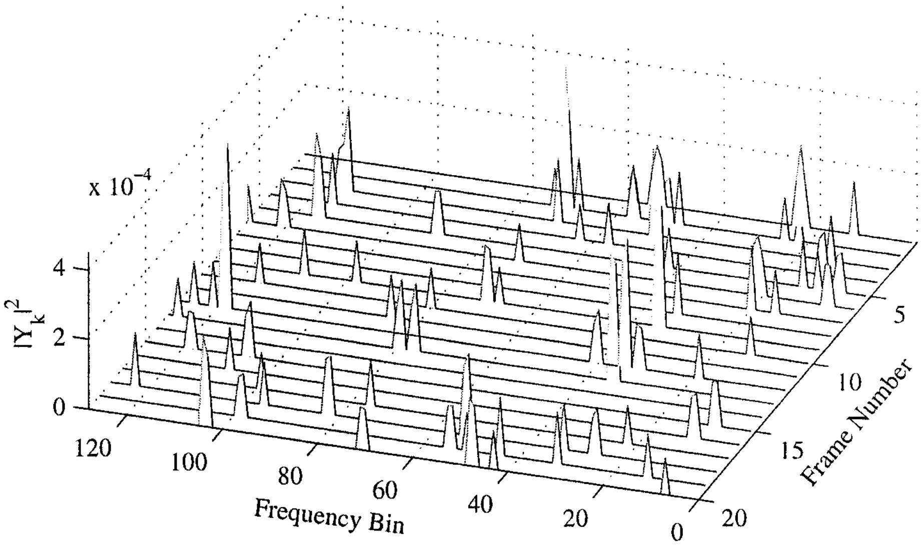


Fig. 2.9 Musical Noise for 20 Frames.

not eliminate musical noise completely. The effect is to reduce its annoying effects to the listener.

Reduction of Musical Noise in EMSR

The Ephraim and Malah MMSE-STSA and MMSE-LSA Noise Suppression Rule (collectively called EMSR) have been reported to attenuate the musical noise phenomenon considerably without introducing audible distortions. This feature explains, why it is an excellent choice for restoration of musical recordings and other corrupted speech. Olivier Cappé attempts to explain how MMSE-STSA counters the musical noise effect, [30]. There are two features in EMSR that assist in the suppression of musical noise:

A. Smoothing effect on *a priori* SNR using the Decision-Directed method.

Consider Fig. 2.10, which is used to illustrate two properties of the *a priori* SNR, ξ_k:

- When γ_k hovers close to 0 dB (i.e., noise only frame), ξ_k emerges as a highly smoothed version (due to the smoothing behaviour of the decision-directed method, Eq. (2.22), used to compute ξ_k) of γ_k. In other words, the variance of ξ_k is much smaller than the variance of γ_k for noise-only frames. This is seen on the left side (frames 0–80) of Fig. 2.10.

- When γ_k is above 0 dB (i.e., a noise+signal frame), ξ_k tends to follow γ_k with a delay of one frame. Frames 80–130 in Fig. 2.10 are indicative of such a trend.

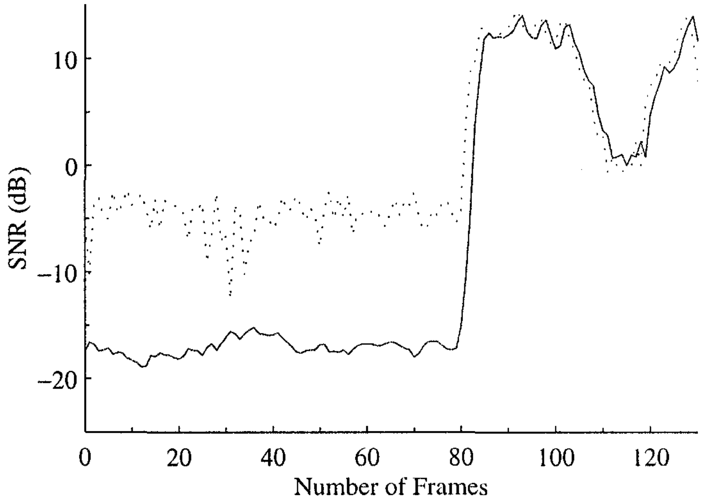


Fig. 2.10 SNR in successive short-time frames. Dotted: *a posteriori* SNR, γ_k; Solid: *a priori* SNR, ξ_k. For the first 80 frames the analyzed signal is pure noise. For the next 50 frames an analyzed signal of 20 dB SNR emerges.

Recall that G_{MMSE} is largely a function of ξ_k, hence the attenuation function does not exhibit large variations over successive short-time frames (due to the reduced variance of ξ_k during noise-only frames). As a consequence, the musical noise (sinusoidal components appearing and disappearing rapidly over successive frames) is considerably reduced. In the EMSR approach, another effect aids in reducing this artifact. In frequency bands containing noise only, it is seen that the average value of ξ_k is about -15 dB (or 0.2 in linear scale), while γ_k is rather high. In that case, the improbable high values of γ_k are assigned increased attenuation (seen more explicitly in Fig. 2.5, where signal with large values of γ_k are attenuated more). This over-attenuation is all the more important as ξ_k is small. Thus, values of the spectrum higher than the average noise-level are "pulled down". This feature is particularly important where the background noise is highly non-stationary. The use of EMSR avoids musical noise whenever noise exceeds its average statistics.

B. Residual Noise Level.

Despite the smoothing performed by the decision-directed method, there are some irregularities that might generate low-level perceptible musical noise. One straightforward way is to assure that the *a priori* SNR is above a certain threshold, ξ_{min}. Usually ξ_{min} is chosen to be larger than the average *a priori* SNR in the short-time noise only frame. As seen in Fig. 2.10, the average value of ξ_k is around -15 dB in noise only frames. Thus, to eliminate low-level musical noise it is recommended to set a lower limit of at least -15 dB on ξ_k (i.e. ξ_{min} is set to -15 dB or higher), [30].

2.5.2 Timbre Change and Loss of Signal Components

It may happen that the level of spectral component of noise is lower than the noise variance estimate. In such cases, some portion of the signal at that frequency will be cancelled. This is particularly hazardous for low-energy voiced speech or even fricatives. Perceptually this will result in a change in timbre of the desired signal. One intuitive way to overcome this potential problem would be to *underestimate* noise level. Obviously, this will lead to incomplete cancellation of noise. Therefore, a trade-off on the spectral floor threshold value is required such that sufficient degree of noise is cancelled (without introducing artifacts) and at the same time assuring that potential loss of timbre effects are reduced.

2.5.3 Phase Distortions and Temporal Smearing

As has been noted earlier, phase information in speech appears as a nuisance parameter and is, thus, often ignored during derivation of speech enhancement algorithms. Usually, an estimate of the magnitude of speech is made and is given the phase of the noisy signal. This will lead to some error in the processed signal and can cause two audible artifacts: roughness and temporal smearing, [31]. *Roughness* is usually heard during sustained sounds such as vowels (e.g. vowels). *Temporal Smearing*, on the other hand, identifies itself as pre- and post-echoes during transition from voiced-unvoiced (or vice-versa) segments. In 1985, Vary derived a direct relation between maximum phase error and noise-to-signal ratio for Gaussian noise. The relation is summarized below, [32]:

$$E[\Delta\Phi_{max}] = \sin^{-1}\left(\sqrt{\frac{\pi}{2}}\cdot\frac{N(f)}{S(f)}\right). \tag{2.35}$$

This equation is seen plotted in Fig. 2.11. One point of interest, on the curve, is when $E[\Delta\Phi_{max}]$ takes on the value of $\pi/4$ (with a corresponding SNR of 2.49 dB).

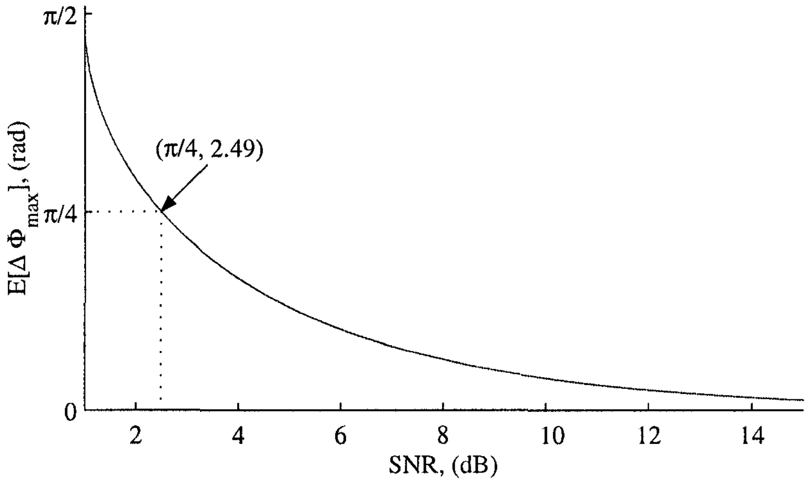


Fig. 2.11 Expected maximum phase error vs. input Signal-to-Noise-Ratio.

It has been shown that while the ear is relatively insensitive to phase, the threshold at which a random phase error becomes audible is about $\pi/4$, [32]. For larger phase errors, speech takes on rough quality. This is to say that as long as the input SNR of corrupted speech is more than 2.49 dB, it can be denoised without phase error becoming audible.

The effects of phase error of less than $\pi/4$ can be heard as temporal smearing—the result of time aliasing. Recall that in most algorithms, the magnitude-estimate of speech is combined with the phase of the noisy observation, ϑ_k. In order to avoid artifacts resulting from phase errors, it would be required to estimate phase (apart from estimating the magnitude from the noisy observation). Since, this phase information is not readily available, temporal smearing results and shows up as pre- or post-echoes in reconstructed signal. Subjective tests show that the effects of temporal smearing are detected by a certain group of trained listeners only, [31]. Therefore, its elimination is considered beyond the scope of this thesis.

2.6 Distortion Measures

A good distortion measure is one that would strongly correlate with the subjective quality of the enhanced speech i.e., small and large distortion should correspond to a good and bad subjective quality. However, this is not always true. For instance, it is sometimes more pleasing to leave residual noise in valleys of formant structures and this will be at the cost of reduced SNR gain. Nevertheless, these measures are used extensively in speech processing for a variety of purposes, [33]. In speech enhancement systems, they are primarily used to define the constraint for derivation of gain curves and evaluating the performance of the system. These distortion measures can be classified into two main categories: *subjective distortion measures* and *objective distortion measures.*

Subjective Distortion Measures

Subjective measures rely heavily on the opinion of a group of listeners to judge the quality or intelligibility of processed speech. These measures are often time consuming and costly as they require proper training of listeners. In addition to this, a constant listening environment (e.g., playback volume), identically tuned output device (e.g., headphones and/or speakers) are necessary. Nevertheless, subjective test results present the most accurate system performance, insofar as intelligibility and speech quality are concerned, as they are determined perceptually by the human auditory system. These tests can be structured under two types of evaluation procedures: *speech intelligibility testing* and *speech quality evaluation.*

- Speech Intelligibility Testing
  - Diagnostic Rhyme Test (DRT)
  - Modified Rhyme Test (MRT)
  - Diagnostic Medial Consonant Test (DMCT)
  - Diagnostic Alliteration Test (DALT)
- Speech Quality Evaluation
  - AB Test
  - Diagnostic Acceptability Measure (DAM)
  - Mean Opinion Score (MOS)
  - Degradation Mean Opinion Score (DMOS)
  - Comparison Mean Opinion Score (CMOS)

While the DRT is the most commonly used intelligibility test (as it computes the percentage of words or phonemes that are heard correctly), MOS and DAM are widely used for speech quality evaluation and strive to rate the naturalness of processed speech, [34].

2.6.1 Mean Opinion Score

MOS is a test for measuring the acceptability or quality of speech over a communication system. Such a score requires the listener to judge the overall quality of a communication system or link on a five category scale for purposes of telephone communication, Table 2.1.

Table 2.1 Verbal categories and the five point scale for MOS.

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

The MOS ratings are reliable as they are based on human responses and perception. A large number of listeners is required, so that a reasonable assessment can be made about system performance. This can be time consuming and expensive. Hence, various objective measures have been developed that are aimed at returning the same result as would have been returned by exhaustive subjective testing. These objective measures simulate listeners' perception and the testing environment through complex non-linear functions. Some such objective measures that have been standardized by International Telecommunications Union (ITU-T) are Perceptual Speech Quality Measure (PSQM)[9], Perceptual Audio Quality Measure (PAQM), Perceptual Analysis Measurement (PAMS), Perceptual Evaluation of Audio Quality (PEAQ)[10] and Perceptual Evaluation of Speech Quality (PESQ). Following a competition (hosted by ITU-T study group 12), PAMS and PSQM attained the highest performance. These have been combined to produce PESQ and has replaced P.861 in early 2001 with P.862, [35]. PESQ has been shown to provide a score that is strongly correlated

[9] PSQM was adopted in 1996 as ITU-T recommendation P.861.

[10] PEAQ became ITU-R recommendation BS.1387 in 1999.

to the subjective ratings obtained by MOS[11].

2.6.2 Perceptual Evaluation of Speech Quality

The purpose of this subsection is to present a brief overview of PESQ and its scope and applications.

Overview of the PESQ algorithm

A high-level structure of the model is seen in Fig. 2.12.

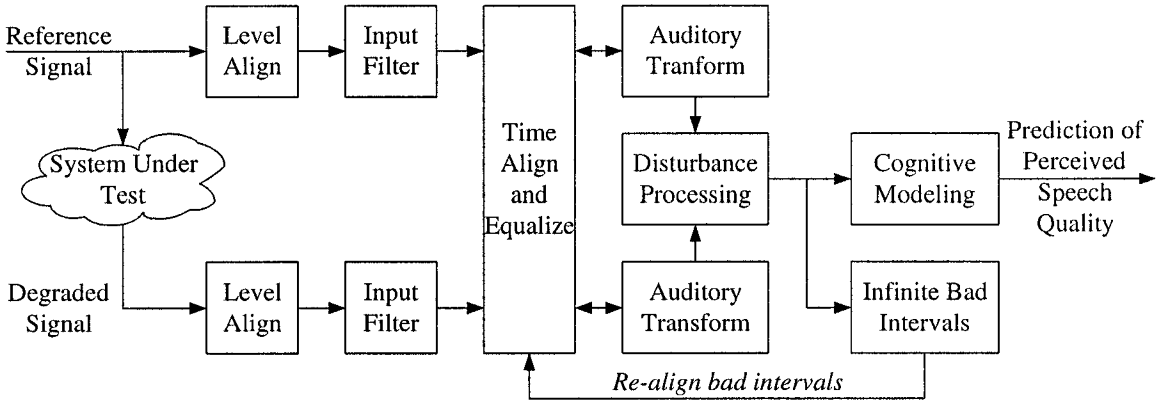


Fig. 2.12 Structure of perceptual evaluation of speech quality (PESQ) model. Redrawn from [35].

Firstly, the reference signal and degraded signal are level aligned such that they have identical playback levels. The signals are then filtered (using standard FFT) with an input filter that mimics bandpass telephone filters (about 330–3300 Hz). This is followed with time-alignment and equalization. The auditory filters deployed by PSQM are then used to process the ensuing signals. The auditory transformation also involves equalization for linear filtering in the system and for gain variation. Two distortion parameters are extracted for the disturbance, which is the difference between the transforms of the two signals. They are then aggregated in frequency and time domain, and mapped to a prediction of subjective

[11] For 22 known ITU benchmark experiments the average correlation was 0.935. For an agreed set of eight experiments used in the final validation—experiments that were unknown during the development of PESQ—the average correlation was also 0.935, [36].

MOS, [36, 35]. Note that the range of MOS mapping is from −0.5 to 4.5, in contrast to the range listed in Table 2.1.

Scope and Application

A range of applications and test conditions have been identified for which PESQ has been validated, [36]. Some of these include waveform codecs (e.g., G.711, G.726, G.727), CELP and hybrid codecs (e.g., G.728, G.729, G.723.1), mobile codecs (e.g., GSM FR, EFR, AMR, EVRC) and for various other network applications. Although PESQ has *not* yet been validated for effects and artifacts generating from noise reduction algorithms, it has been used in this thesis to assess the quality of noisy degraded (and enhanced speech) *coded* signal against clean unprocessed signal. Certain other applications of the PESQ measure that have not been fully characterized yet include music quality, wideband telephony, listener echo, very low bit-rate vocoders below 4 kbps (e.g., MELP) and acoustic and head-and-torso simulator measurements.

Objective Distortion Measures

Contrary to subjective measures, this class of distortion measures is computed from the time or frequency characteristic of a speech signal. Objective tests are less expensive to administer, give more consistent results and are not subject to human failings of administrator or subject. If an objective measure could be found which was highly correlated with human preference, its utility would be undeniable. The fundamental problem in research into objective quality measures requires a strong understanding of the complexities of the speech perception process. A useful distortion measure would be required to possess a little degree of the following properties: 1) be subjectively meaningful—a small or large distortion corresponds to good or bad subjective quality, respectively; 2) be tractable or amenable to mathematical analysis; 3) be computationally efficient, [29]. An extensive performance analysis of a multitude of objective distortion measures is given in [34].

The most common distortion measure is the conventional squared error, error power or error energy. However, for low-bit rate speech coders, such a distortion measure may not be subjectively meaningful. In particular, a large distortion in a squared error sense does not necessarily imply poor quality. For instance, a fricative (e.g., "shh") is essentially a broadband noise process and any representative waveform will sound the same. To overcome this problem, there are other distortion measures that are geared towards low-bit rate coders—Euclidean distance between LSF parameters, percent correct pitch and bandpass voicing estimation being some. These measures have been used in this thesis to measure performance of speech enhancement as pre-processors for vocoders, and as such are briefly discussed in the following few sections.

2.6.3 Signal-to-Noise Ratio

As the name invariably implies, SNR is the ratio of the signal energy to the noise energy:

$$\mathrm{SNR} = 10\log_{10}\frac{\sum\limits_{n=-\infty}^{\infty} s\left[n\right]^2}{\sum\limits_{n=-\infty}^{\infty}\left(s\left[n\right]-\hat{s}\left[n\right]\right)^2}\ \mathrm{dB}, \tag{2.36}$$

where $s\left[n\right]$ is the original/enhanced signal and $\hat{s}\left[n\right]$ is the 'noisy' signal. However mathematically simple, the SNR measure carries with it the drawback of being a poor estimator of subjective quality. The SNR of a speech signal is primarily governed by the high energy segments, e.g., voiced speech. However, noise has a greater perceptual effect in the weaker energy segments, [22]. A high SNR value, is thus, not necessarily indicative of good perceptual quality of the speech.

2.6.4 Segmental Signal-to-Noise Ratio

The $\mathrm{SNR}_{\mathrm{seg}}$ in dB is the average SNR (also in dB) computed over short frames of the speech signal. The $\mathrm{SNR}_{\mathrm{seg}}$ over M frames of length N is computed as:

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{M}\sum_{i=0}^{M-1} 10\log_{10}\left[\frac{\sum\limits_{n=iN}^{iN+N-1} s^2\left[n\right]}{\sum\limits_{n=iN}^{iN+N-1}\left(s\left[n\right]-\hat{s}\left[n\right]\right)^2}\right]\ \mathrm{dB}, \tag{2.37}$$

where the $\mathrm{SNR}_{\mathrm{seg}}$ is determined for $\hat{s}\left[n\right]$ over the interval $n = 0,\ldots,NM-1$. Typically, the frame length is 15–25 ms (or N between 120–200 for speech sampled at 8 kHz). This distortion measure weights soft and loud segments of speech equally and thus circumvents the problem by modelling perception better.

If a frame were weak in speech energy, the corresponding SNR can take on large negative values; having an unwarranted effect on the overall estimate. By either removing weak speech energy frames or introducing an SNR threshold (usually 0 dB), the effects of negative SNR can be reduced. Similarly, a high SNR_{seg} can be a result of frames that are high-energy voiced segments. It has been noted in, [22] that human auditory perception cannot distinguish among frames with an SNR greater than 35 dB, and hence this value can be used as a threshold to prevent positive bias in segmental SNR.

However, SNR measures are meaningful only for coding schemes where the "noise" is deemed additive. Thus, these measures are useful with subband and transform coders but not for vocoders which introduces speech correlated noise components, [34, 37]. In addition, it must be noted that these vocoders focus only on the magnitude of the speech spectrum, as the human auditory system is relatively insensitive to phase distortion. Therefore, the "processed" speech might sound mechanical yet still be intelligible. For the latter type of speech coders, objective measures based on spectral and LPC parameters are useful. These are briefly summarized below.

2.6.5 Weighted Euclidean LSF Distance Measure

The accuracy of LP filter coefficients is fundamental to the design of an accurate parametric coder. Since, in most works, LSFs[12] are used to represent these coefficients, it would be useful to have a metric that computes the distance of enhanced spectra from clean spectra in the LSF sense. One such metric is the Euclidean distance, which computes the $\mathcal{L}_2$ norm between the test and the reference vector, [1]:

$$\Delta^2_{LSF} = \frac{1}{10M} \sum_{m=1}^{M} \sum_{i=1}^{p} \left[c_i (f_{i,m} - \hat{f}_{i,m})\right]^2, \tag{2.38}$$

where $f_{i,m}$ and $\hat{f}_{i,m}$ are the ith LSFs of the mth index of the test and reference vector respectively, c_i are the weights assigned to the ith LSF coefficient, M is the total number of speech signal frames and p is the order of the LP filter. For a 10^{th} order LP filter, the

[12] LP coefficients are either quantized or interpolated prior to transmission. However, small changes in coefficients might result in large changes in the power spectrum and possibly yield an unstable LP filter. A number of one-to-one mappings have been developed. One of the most commonly used representations is the *Line Spectral Frequencies* (LSFs).

fixed weights c_i are given by:

$$c_i = \begin{cases} 1.0, & \text{for } 1 \leq i \leq 8, \\ 0.64, & \text{for } i = 9, \\ 0.16, & \text{for } i = 10. \end{cases} \tag{2.39}$$

The human ear is more sensitive to differences in the lower frequency range than the higher frequency range. Therefore, the lower frequency bins are emphasized more by assigning them larger weight.

Clearly, lower values of Δ^2_{LSF} are indicative of enhanced spectra being closer to clean spectra. In Section 3.2.1, another STSA algorithm is discussed, that minimizes the Euclidean distance between the LSF parameters.

2.6.6 Percentage Correct Pitch Estimation

Pitch is computed once per frame and is used to determine the spacing between the harmonic structure of speech spectrum. In an attempt to give an objective metric to pitch representations, Guilmin *et al.* in 1999, deemed a pitch detection to be correct when its value does not differ of more than 5% from the exact value and the result is averaged only on frames where speech is present. Obviously, the goal of a good pre-processor is to have a large percentage of correct pitch estimation after it has been applied to acoustically impaired speech signal, [38].

2.6.7 Percentage Correct Bandpass Voicing Estimation

In the MELP speech coder the analysis frequency band is split into five subbands. Each subband is assigned a bandpass voicing strength based on the normalized autocorrelation value (see Section 3.1.1 for details of computation). These strengths are quantized to either a 1 or 0, depending on pre-defined thresholds. Thus, each band has five quantized bandpass values. In order to evaluate a percentage for correct bandpass voicing strengths, the quantized values for the original clean and degraded/enhanced signal are computed. The value of the voicing strength, per band, is considered correct if there is no bit inversion. The result is averaged only on frames where speech is present, [38]. Once again, larger values are indicative of better efficiency of speech enhancement techniques.

2.7 Summary

This chapter focussed on derivation and study of speech enhancement techniques deployed in the frequency domain. The EVRC noise suppression block, MMSE-STSA and MMSE-LSA were studied in some detail. Some of the limitations (such as musical noise and timbre changes) of noise suppression schemes was also looked at. This chapter also described some of the objective and subjective distortion measures relevant for low-bit rate speech coders. Such distortion measures are either used as cost functions for derivation of suppression rules, or are used to evaluate their performance under acoustic conditions.

Chapter 3

Application: Pre-Processor for Vocoders

In Chapter 2, two STSA (MMSE-STSA and MMSE-LSA) noise suppression algorithms were described. Such systems are particularly useful in the arena of low-bit rate digital communications. Since the first US and NATO[1] standard LPC-10 speech coder, low bit-rate speech coding has greatly improved at the 2.4 kbps data rate. New vocoders provide increased intelligibility and quality but, nevertheless are sensitive to background additive noise. One of the primary concerns of this chapter is one such vocoder—Mixed Excited Linear Prediction (MELP) speech coder. The parameters required by the MELP encoder/decoder are presented in Section 3.1.1/Section 3.1.2. In Section 3.1.3 it will be seen how the coding parameters are disturbed in noisy environment and how it might affect reconstruction of speech at the decoder. Some specialized pre-processors that aim towards reducing background noise while assuring good parameter extraction are the motive of Section 3.2. Later Section 3.2.1 looks at another noise suppressor that aims at tracking LP parameters better, while Section 3.2.3 describes the algorithm developed in this thesis with the objective of improving the perceptual quality of degraded speech. Furthermore, a comparison of the various noise suppression rules will be studied in Section 3.2.4.

[1] NATO: North Atlantic Treaty Organization.

3.1 The Mixed Excitation Linear Prediction Coder

Legacy tactical secure voice communications use either the 2.4 kbps Linear Predictive Coding (LPC-10) algorithm or the 16 kbps Continuously Variable Slope Delta Modulation (CVSD) algorithm for speech compression. About two or three decades ago these algorithms were considered the state of the art in narrow band speech coding.

In March of 1996, the U.S. government DoD[2] Digital Voice Processing Consortium (DDVPC) announced selection of the 2.4 kbps MELP voice coding algorithm as the next standard for narrow band secure voice coding products and applications. Kohler has presented valuable comparisons (e.g. MOS, DMOS, DRT and DAM scores) of MELP against the 4.8 kbps federal standard (FS1016) Code Excited Linear Prediction (CELP), the 16 kbps CVSD and the venerable federal standard 2.4 kbps LPC-10, [39]. In that paper, Kohler publishes tests on quality, intelligibility, recognizability and communicability. The target device for the new federal coder is an 80 MHz processor coupled with four megabits of memory, of which the MELP coder demands only 51% and 77%, respectively. In this section, details of MELP are described succinctly.

Any parametric coder depends on the accuracy of LPC analysis filter and residual signal. As was seen in Chapter 1, the LP coefficients model the human vocal tract, while the residual signal models the excitation signal—periodic pulses for voiced and random white noise for unvoiced speech segments. MELP, much like any other traditional LPC vocoder, relies on accurate LP parameter estimation to compress speech. The following five features assist MELP in achieving supremacy over its predecessors, Fig. 3.1:

1. **Mixed Excitation**: This is implemented using a multi-band mixing model that uses an adaptive filtering structure to simulate frequency-dependent voicing strengths. The primary effect is to reduce the "buzz" that is associated with LPC vocoders, especially in broadband noise.

2. **Aperiodic Pulses**: These are mostly used during transitions from voiced/unvoiced (or vice-versa) regions in speech segments. This allows the decoder to reproduce erratic glottal pulses without introducing tonal sounds. In the case of voiced segments, the MELP coder can synthesize using either periodic or aperiodic pulses.

[2] DoD: Department of Defence.

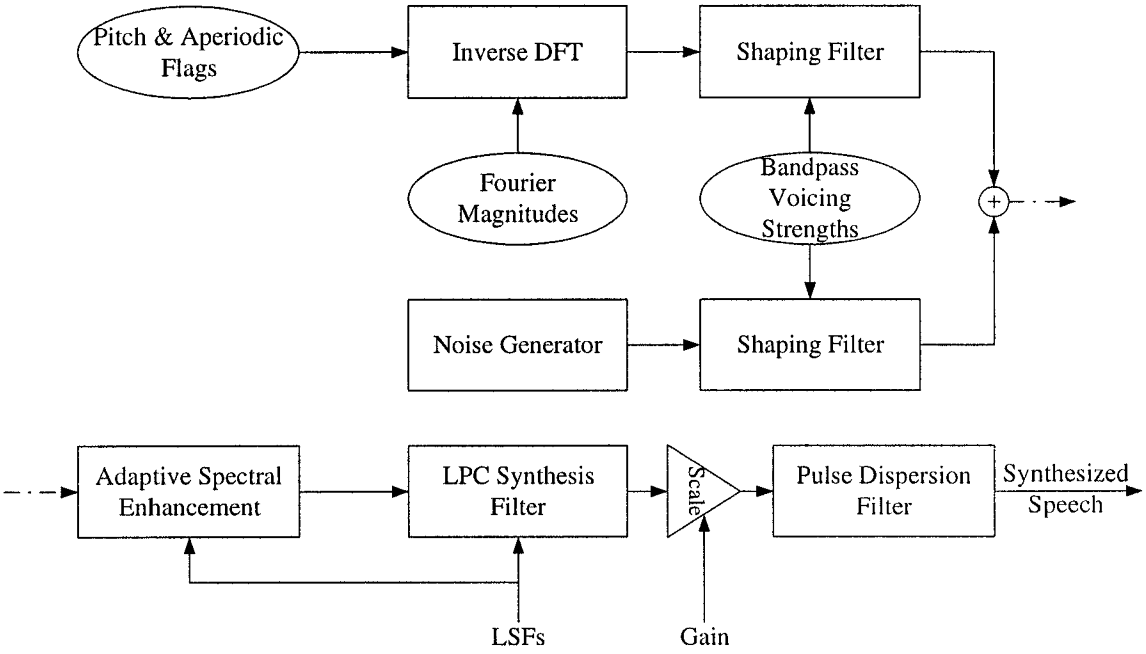


Fig. 3.1 MELP decoder block diagram. Redrawn from [40].

3. **Adaptive Spectral Enhancement**: This is based on the poles of the LP synthesis filter and is exploited to enhance the formant structure of the synthetic speech so as to match it to the bandpass waveforms.

4. **Pulse Dispersion**: This is implemented using a fixed filter based on spectrally-flattened triangle pulses. The filter assists in reducing some of the harsh qualities of the synthetic speech by spreading the excitation energy within a pitch period.

5. **Fourier Magnitude Modelling**: The first 10 Fourier coefficients are extracted from the peaks of the Fourier transform of the prediction residual signal. Since the human ear is most sensitive at lower frequencies, the first ten coefficients assist in increasing quality of synthetic speech—especially for male speakers in the presence of background noise, [40].

3.1.1 Speech Coding Parameters—The Encoder

The input signal is high-pass filtered using a 4^{th} order Chebychev type II filter, having a cutoff frequency of 60 Hz and a stopband rejection of 30 dB. The output of this filter will be referred to as input speech signal throughout the following description. The MELP coder takes the input signal and segments it into frames of 180 samples using a Hanning window. For each signal frame the vocoder extracts 10 LP coefficients, 2 gain factors, 1 pitch value, 5 bandpass voicing strength values, 10 Fourier magnitudes and an aperiodic flag, Fig. 3.2.

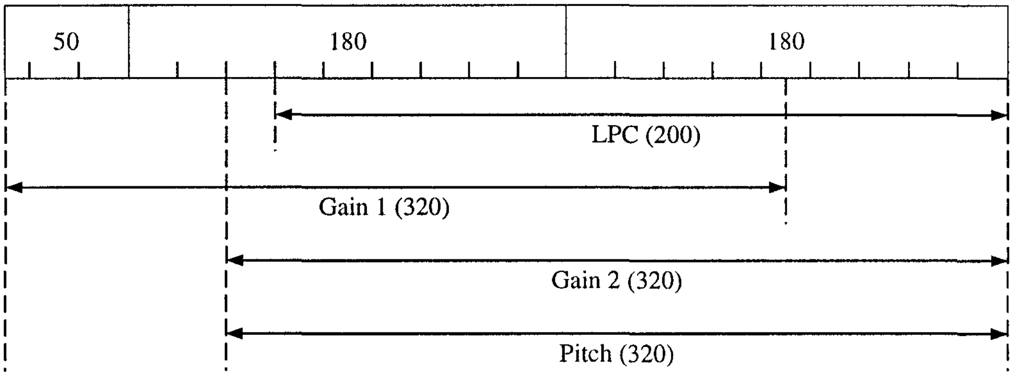


Fig. 3.2 Utilization of data in the input buffer of the MELP coder. Numbers indicate frame sizes.

The following few sections briefly describe how MELP extracts these parameters and use them to reconstruct speech.

LP Coefficients

A 10^{th} order linear prediction analysis is performed on the input speech signal using a 200 sample Hamming window. The traditional autocorrelation analysis is implemented using Levinson-Durbin recursion, [22]. In addition, a bandwidth expansion factor of 0.994 is applied to each prediction coefficient $a_i, i = 1, 2, 3, \ldots, 10$, where each coefficient is multiplied by 0.994^i. Conversion of these LPC parameters to LSF is not very straightforward, however, Kabal and Ramachandaran proposed an efficient way using Chebychev polynomials, [41]. These LSF parameters are sorted in ascending order with a minimum separation of 50 Hz. The linear prediction *residual* signal is computed by filtering the input speech signal with the prediction filter coefficients.

Pitch Values

There are several steps before a *final* value is assigned to pitch. First the coder finds the integer value of the pitch, followed by a fractional pitch refinement. The integer pitch is calculated using the *input* speech signal. The signal is processed using a 1 kHz, 6^{th} order Butterworth lowpass filter. The integer pitch value, P_1, is computed as the *maximum* of the signal autocorrelation and can take on values of lag ranging from 40–160 ms, [40].

A refined pitch measurement is made on the 0–500 Hz passband (see below for details on this passband). Two pitch candidates are considered for this refinement and selected from the current and previous frames. For each candidate an integer pitch search is made over lags from five samples shorter to five samples longer than the candidate. A fractional refinement is computed using an interpolation formula (presented in [40]) around the optimum integer pitch lag. This produces two fractional pitch candidates and their corresponding normalized autocorrelation values. The maximum of the two autocorrelations is selected as the fractional pitch, P_2. For computation of the *final* pitch, P_3, an integer pitch search is performed over lags of 5 samples shorter to 5 samples longer than P_2, rounded to the closest integer. This measurement is done on lowpass filtered *residual* signal. Before assigning P_3 the final pitch value, fractional refinement and pitch doubling checks are made.

Bandpass Voicing Strength Values

Five bandpass voicing strengths, $V_b(i), i = 1, 2, \ldots, 5$, are determined based on five 6^{th} order Butterworth filters with passbands of 0–500, 500–1000, 1000–2000, 2000–3000 and 3000–4000 Hz, respectively. The normalized autocorrelation value that results from P_2 is saved as the lowest bandpass voicing strength, $V_b(1)$. For each remaining band, the bandpass voicing strength is the larger of the autocorrelation value, computed at P_2, as determined by the fractional pitch procedure and the time envelope for the corresponding bandpass signal.

The peakiness of the residual signal is computed as the ratio of the $\mathcal{L}_2$ norm to the $\mathcal{L}_1$ norm of the residual signal. If the peakiness exceeds 1.34, $V_b(1)$ is set to 1. However, if it exceeds 1.6, $V_b(i), i = 1, 2, 3$ are set to 1.

Gain Factors

The input signal gain is measured twice per frame using a pitch adaptive window length. Two window lengths are determined based on values of $V_b(1)$, [40]. The gain in each window correspond to G_1 and G_2 and are computed as the Root Mean Square (RMS) value of the signal in the corresponding window.

Fourier Magnitude Calculation

This analysis measures the Fourier magnitudes of the first 10 pitch harmonics of the prediction residual generated by the quantized prediction coefficients. The coder performs a 512-point FFT on a frame of Hamming windowed vector of 200 samples. Finally, the complex FFT is transformed to magnitudes and the harmonics are found with a spectral peak-picking algorithm. These magnitudes are normalized to have an RMS value of 1.0. If fewer than 10 harmonics are found using the peak-picking algorithm, the remaining are forced to unity.

Aperiodic Flag

The aperiodic flag is set to 1 if $V_b(1) < 0.5$ and to 0 otherwise. When set, it tells the decoder that the pulse component of the excitation should be aperiodic, rather than periodic.

Transmission Format

Prior to transmission these parameters are quantized using Multi-Stage Vector Quantization (MSVQ). The transmission rate is 2400 bps ± 0.01%. Since all frames contain 54 bits, the frame length is 22.5 ms ± 0.01%, Table 3.1. For speech sampled at 8 kHz, this would imply a frame length of 180 samples, Fig. 3.2.

To improve the performance of MELP coder, unused parameters during unvoiced mode are used for Forward Error Correction (FEC), [40].

3.1.2 The Decoder

The received bits are first unpacked and assembled into parameter codewords. Pitch is decoded first, as it (augmented with aperiodic flag information) dictates reception of either voiced or unvoiced speech segment at the receiver, Fig. 3.1. For a voiced segment, the

Table 3.1 Bit allocation in a MELP frame for voiced and unvoiced speech segment.

Parameters	Voiced	Unvoiced
LSF's	25	25
Fourier magnitudes	8	-
Gain (2 per frame)	8	8
Pitch, overall voicing	7	7
Bandpass voicing	4	-
Aperiodic flag	1	-
Error protection	-	13
Synchronization bit	1	1
Total bits / 22.5 ms Frame	54	54

ten Fourier magnitudes (Table 3.1) are inverse transformed to obtain the *pulse* excitation followed by multiplication with the square root of the associated pitch. For an unvoiced segment, on the other hand, a uniform number generator is used to obtain the *noise* excitation pattern. The pulse and noise excitation signals are then filtered and summed to obtain the mixed excitation signal. It is this very part of the vocoder that is unique to MELP. Such a mixing of pulse and noise assists in countering the effects of tonal thumps and buzz in reconstructed speech. Prior to synthesizing speech, the mixed excitation signal is spectrally enhanced. The filter coefficients of this enhancement stage are generated by bandwidth expansion of LP filter transfer function. Synthesized speech is obtained by filtering the spectrally enhanced excitation signal with the LP synthesis filter (whose coefficients are generated from the transmitted LSFs), gain modification and filtering with a pulse dispersion filter, [40].

3.1.3 Speech Coding Parameters in Noisy Environment

It is important to note that the residual signal is used for the computation of all the key parameters (apart from LPC parameters) involved in the MELP coder. Hence, its accuracy is crucial to the coder's optimal performance insofar intelligible speech reconstruction is concerned. Further, as noted above, the residual signal is computed as the output of the LP analysis filter. This in turn requires an accurate estimation of these parameters. How these parameters degrade under background noise are of concern in this subsection.

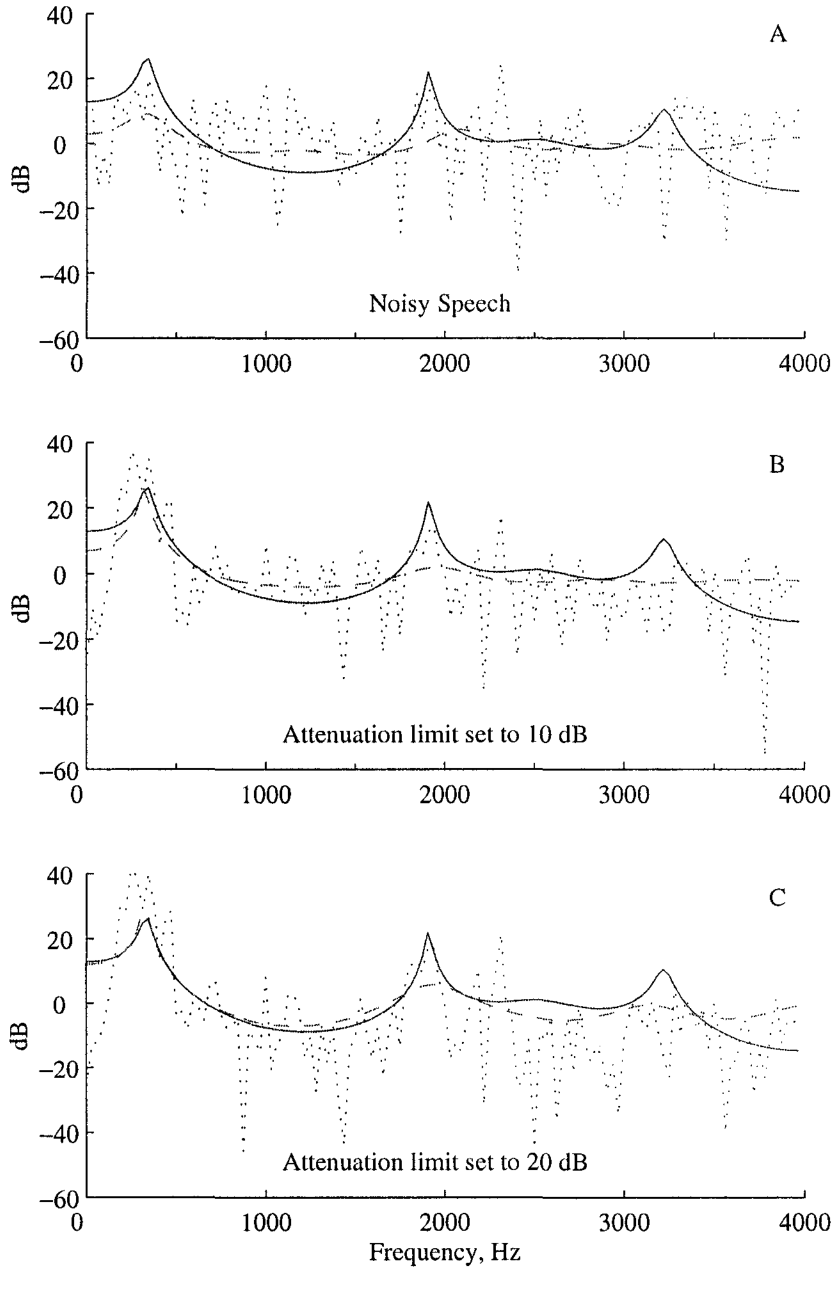


Fig. 3.3 Solid: LPC Spectrum of Clean signal; Dashed: LPC of Noisy/Enhanced speech; Dotted: Magnitude squared DFT of Noisy/Enhanced speech.

After applying one of two Ephraim and Malah Suppression Rules (EMSRs) as front-end processors to parametric speech coders, it is noticed that most degradations and loss of intelligibility are due to errors in the spectral parameters. In 1999, Martin and Cox tackled the problem of speech enhancement for vocoders. More specifically they noted that while spectral valleys in between formant frequencies are *not* important for speech perception (thus filling them with noise will improve auditory impression), they are important for LPC estimation, [15]. A low noise attenuation does not remove a sufficient amount of noise from the disturbed signal and a high noise attenuation will also distort speech components. Thus, it is not quite obvious how aggressive the noise suppression rule should be to trade-off between perception and estimation.

As an example, consider Fig. 3.3-A which shows the magnitude squared DFT coefficients (dotted), the LPC spectrum for a given frame of noisy speech data (corrupted with additive white noise, such that the overall SNR is 6 dB[3]) and the LPC spectrum of the same *clean* speech frame (solid) for reference. In Fig. 3.3-B, the same graphs are seen with minimum gain set to -10 dB, while Fig. 3.3-C depicts graphs when the minimum gain is limited to -20 dB[4]. Comparing Fig. 3.3-B and Fig. 3.3-C shows that the latter tracks the clean speech more closely. In particular, note that the first formant (the peak) matches the clean signal formant better than in case B.

Therefore, it can be concluded that a high maximum noise reduction is beneficial for enhancement of low SNR speech when it is input to parametric speech coders. However, for high SNR speech it might lead to undesirable speech distortions and, especially during speech pause, it might result in musical noise. It is therefore of prime interest to optimize the maximum attenuation (or minimum gain value) as a function of input speech SNR such that the LPC coefficients track the original LPC structure better.

3.2 Specialized Pre-Processors

In Section 3.1.1, it is noted that different speech parameters are obtained from different properties of speech. Therefore, it seems natural to have different speech enhancement pre-processors to accurately extract these different parameters, Fig. 3.4. In 1999, Accardi

[3] To compute the amount of noise to be added to obtain a particular SNR (in dB) refer to Section 4.1.2.

[4] The corrupted speech was enhanced using the MMSE-LSA, seen earlier in Section 2.4, due to its close relation to Itakura-Saito distortion measure, [34]. The gain is restrained to either -10 dB or -20 dB.

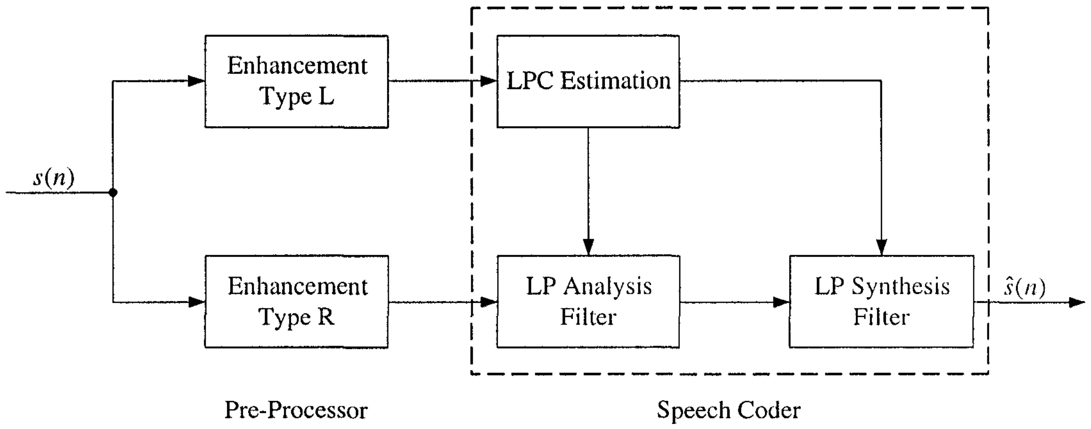


Fig. 3.4 Specialized pre-processor for speech coders. In the original proposal Accardi refers to Type L as Type 1 and Type R as Type 2.

and Cox proposed using one enhancement scheme for computation of LP parameters (Type L)[5] and another scheme for computation of residual signal (Type R), [16], for their target low-bit rate IS-641 speech coder[6]. For Type L enhancement, Accardi proposes, what he calls a magnitude-squared core estimator: $E\{|X_k|^2X_k\}$ in terms of *a priori* SNR defined by Ephraim and Malah in Section 2.3.1:

$$\begin{aligned} E\{|X_k|^2|Y_k\} &= \left(\frac{S_{xx}[k]}{S_{xx}[k]+S_{dd}[k]}\right)^2|Y_k|^2+\frac{S_{dd}[k]S_{xx}[d]}{S_{xx}[k]+S_{dd}[k]}, \\ &= \left(\frac{\eta_k}{1+\eta_k}\right)^2|Y_k|^2+\left(\frac{\eta_k}{1+\eta_k}\right)S_{dd}[k], \end{aligned} \tag{3.1}$$

where X_k and Y_k are the spectral components of the clean and the noisy speech. S_{xx} and S_{dd} are the power spectral densities of the clean speech and noise only signal, respectively, and η_k is interpreted as the *a priori* SNR.

For Type R enhancement scheme he proposes a gain function G_{SS} using a Signal Sub-

[5] Although, in the original proposal Accardi refers to Type L as Type 1 and Type R as Type 2. However in Chapter 4, the nomenclature of Type L and Type R will be used.

[6] The IS-641 speech coder is the Enhanced Full Rate (EFR) speech codec that has been standardized in 1996 for the North American TDMA digital cellular systems (IS-136). The codec is based on ACELP algorithm and has a gross bit-rate of 13.0 kbps (7.4 kbps for source coding and 5.6 kbps for channel coding). The resulting speech quality is close to that of the wireline telephony as compared with G.726 32 kbps ADPCM speech codec, and is robust to errors arising from typical cellular operating conditions such as transmission errors, environmental noises and tandeming of speech codecs, [42].

space Approach proposed by Ephraim and Van Trees, [43]:

$$G_{\mathrm{SS}} = \sqrt{\exp\left(-\nu\frac{\sigma_d^2}{\lambda_x[k]}\right)}, \tag{3.2}$$

where ν is a constant that determines the level of aggression of the enhancement, σ_d is the noise variance and λ_x is the kth eigenvalue of the clean speech. G_{SS} is multiplied with the Fourier magnitude of the corrupted speech, combined with the phase of noisy speech and inverse FFT transformed and overlap-added with previous enhanced frames.

However, Accardi proposed Eq. (3.2) in terms of the *a priori* SNR, ξ_k, the probability of signal absence, q_k, and the multiplicative modifier, G_{MM}, (these were discussed earlier in Section 2.3.2):

$$G_{\mathrm{SS}} = G_{\mathrm{MM}} \cdot \exp\left(-\frac{\nu_k}{\xi_k}\right)\bigg|_{\xi_k=\frac{\eta_k}{1-q_k}}. \tag{3.3}$$

In [44], Accardi compares the formant structure of a certain speech segment in an attempt to show how well his Type L enhancement scheme tracks the formant structure of the original clean speech in the same frame. However, as will be seen in Section 3.2.1, there is a more reliable strategy for tracking the formant structure. As for Type R enhancement, Accardi argues the validity of Eq. (3.3) by performing some informal listening tests and establishing that pitch prediction errors are not terribly annoying in the decoded speech—manifesting themselves as subtle clicks in the worst case. In Section 3.2.3, a new model is derived with the motivation of maximizing MOS so as to perceptually enhance the lower branch in Fig. 3.4.

3.2.1 Adaptive Limiting Scheme for LP Estimation

This scheme strives to improve the LP parameter estimation and, thereby, to increase the robustness and intelligibility of coded speech. The motivation behind their work was presented in Section 3.1.3.

As seen in Section 2.5.1, it is advised to limit ξ_k between 0.1 and 0.2 to avoid low-level structured musical noise. This means that less attenuation is assigned to low SNR frames and hence, the spectral valleys are less attenuated. As a result the overall spectral shape of speech sound is affected and an accurate estimation of spectral parameters is disturbed. On the other hand, not limiting ξ_k will yield considerable amounts of annoying musical noise.

In 1999, Martin and Cox attack this problem by presenting an *Adaptive Limiting Scheme for LP Estimation* (referred to as MMSE-LSA with ALSE in this thesis) to compute a lower bound on the *a priori* SNR, [15]. Whenever a speech pause is detected (using a VAD) ξ_k is set to a constant minimum value of 0.12. If an active speech segment is detected the limit changes adaptively (as it is a function of SNR):

$$\xi_{min1} = \begin{cases} \xi_{minP}, & \text{for pause segments,} \\ \xi_{minP} \cdot \exp(-5) \cdot (0.5 + \text{SNR})^{0.65}, & \text{for active speech segments,} \end{cases} \tag{3.4}$$

where $\xi_{minP} = 0.12$ (in their proposal) and SNR is the Signal-to-Noise Ratio of the analysis frame. This preliminary limit is smoothed by a first order recursive system to provide smooth transition between active and pause frames:

$$\xi_{min}(m) = 0.9\xi_{min}(m-1) + 0.1\xi_{min1}(m), \tag{3.5}$$

where m is the index of the analysis frame. This resulting ξ_{min} is then used as lower limit for ξ_k.

3.2.2 Formulation of Adaptive Limiting Scheme—A Heuristic Approach

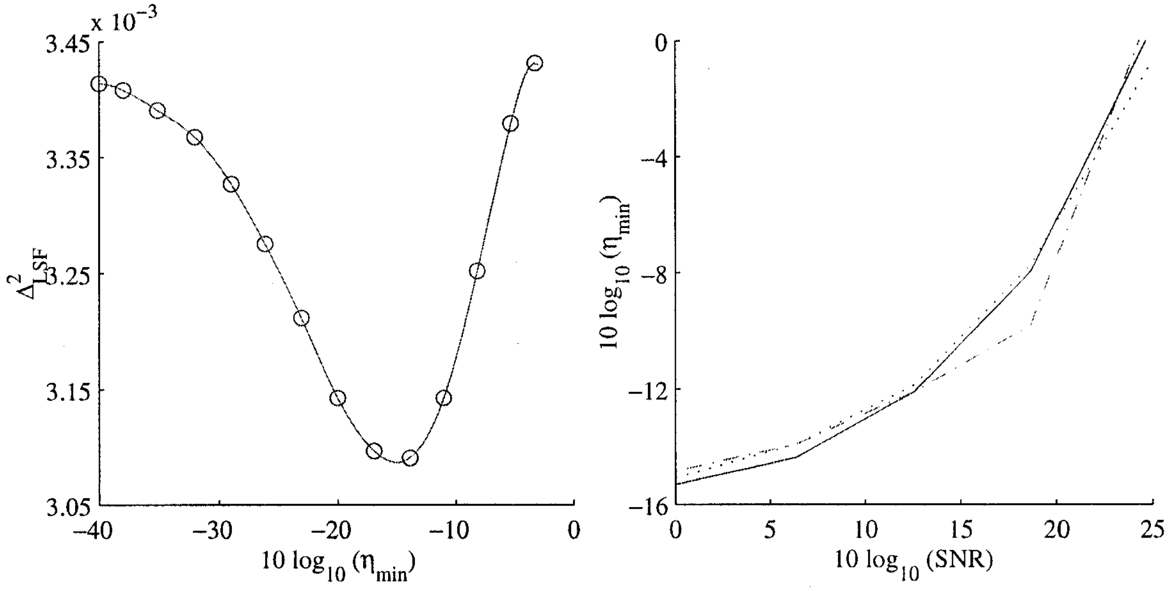
The proposal on the adaptive lower limit for ξ_k, made by Martin and Cox in [15], is not very intuitive. In 2000, Martin *et al.* outlined the methodology that was carried out to optimize the lower limit on ξ_k, [45]. The purpose of this sub-section is to highlight the heuristic approach that they deployed.

1. Sixty seconds of male and female speech files were corrupted with computer generated white noise at various SNRs: 0, 6, 12, 18, 24 dB.

2. These files were processed separately with MMSE-LSA estimator for twelve fixed lower limits on ξ_k: -3, -8, -11, -17, -20, -23, -26, -29, -32, -35, -38, -40 dB.

3. The enhanced files were passed through the Adaptive Multi-Rate (AMR) speech coder (operating at 4.75 and 12.2 kbps modes) and LSF parameters for each frame were recorded (five per frame). These parameters were used in conjunction with the corresponding five LSF parameters of the clean spectra to compute Δ^2_{LSF} using Eq. (2.38).

4. For each speech file with the specified input SNR, the Δ^2_{LSF} was plotted against the pre-defined lower limits on ξ_k. A cubic spline was used to interpolate the measured points. The minimum point on this curve was recorded, Fig. 3.5(a).
5. After computing the optimum for all other SNR values the resulting minima is plotted against input SNR, Fig. 3.5(b).



(a) LSF distortion measure Δ^2_{LSF} vs. $10\log_{10}(\eta_{min})$ for input female speech with 0 dB SNR. Redrawn from [45].

(b) $10\log(\eta_{min})$ vs. speech $10\log(SNR)$. Dotted: Male Speech; Dash-Dotted: Female Speech; Solid: Approximation of Eq. (3.6). Redrawn from [45].

Fig. 3.5 Heuristic approach to defining Eq. (3.6).

Consider Fig. 3.5(a), which exemplifies the procedure. A sixty second female speech file is corrupted with computer generated white noise such that the input SNR is 0 dB. This noisy speech is enhanced using the MMSE-LSA noise suppression rule twelve times. Each time the lower limit for ξ_k is chosen to be one of the twelve values listed in point 2 above. Each of the resulting twelve 'denoised' speech files is coded using the AMR speech coder.

The minimum Euclidean distance (Δ^2_{LSF}) is computed for each of the twelve coded files. These values of Δ^2_{LSF} are plotted against the twelve lower limits on ξ_k. The plot for this example of 0 dB female speech is seen in Fig. 3.5(a). Similarly, other Δ^2_{LSF} vs. the lower limits on ξ_k graphs are plotted (not shown here) for speech corrupted at the other four SNR values (see point 1, in the above description of the heuristic approach). The lower limit on ξ_k for which the minimum occurs, is plotted against the input SNR in Fig. 3.5(b). In this example the minimum value of Δ^2_{LSF} is 3.1x10^{-3} at $\xi_{min} \approx 15$ dB. This point is recorded and plotted in Fig. 3.5(b).

Likewise, the procedure is executed with male speech to obtain the dashed curve in Fig. 3.5(b). The solid curve in the figure represents the line of best fit between male and female heuristic data, and is mathematically expressed in Eq. (3.6).

3.2.3 Proposed Adaptive Limiting Scheme for Perception

Owing to the strategy employed by Martin and Cox, it is worthwhile to derive another lower bound on the *a priori* SNR such that MOS is maximized. As seen earlier, the lower branch in Fig. 3.4 serves to model the excitation signal, also known as the excitation signal. Since the Mean Opinion Score (MOS) gives the listeners' opinion on reconstructed speech, it seems legitimate to maximize such a score. It is a well known fact that MOS ratings are very time consuming to obtain as they require a large number of listeners and exhaustive data collection, Section 2.6.1. As noted earlier, several perceptual quality evaluation standards have been released by ITU-T, of which PESQ is the latest released (February 2001) perceptual measure.

In this thesis, work was carried out to maximize MOS. For the method proposed in this thesis, steps 1–5 (from Section 3.2.2, but with 12 s of speech files) were followed with MOS (taken as output of the PESQ), instead of Δ^2_{LSF}, as the cost function. Even though PESQ has *not* been validated for noise reduction systems yet, it seems the most reasonable solution to obtaining MOS ratings for the purposes of deriving an adaptive lower bound on ξ_k. For step 3, two speech coders were used instead of the AMR speech coder: the 8 kbps G.729[7] and the MELP speech coder. As noted earlier in Section 2.6.1, PESQ has *not* been

[7] The ITU-T recommendation describes the 8 kbps toll quality speech coder that is based on Conjugate Structure Algebraic Code LP (CS-ACELP) algorithm. The coder operates on 10 ms of speech frames. For every frame, the speech signal is analyzed to extract the parameters of the CELP model (LP filter coefficients, adaptive and fixed codebook indices and gains). The bits are encoded and transmitted. At

validated for low bit-rate speech coders (such as those with bit rate < 4 kbps). The results obtained with G.729 are very close in comparison to those obtained with MELP as the underlying test coder—Fig. 3.6 compared to Fig. 3.7.

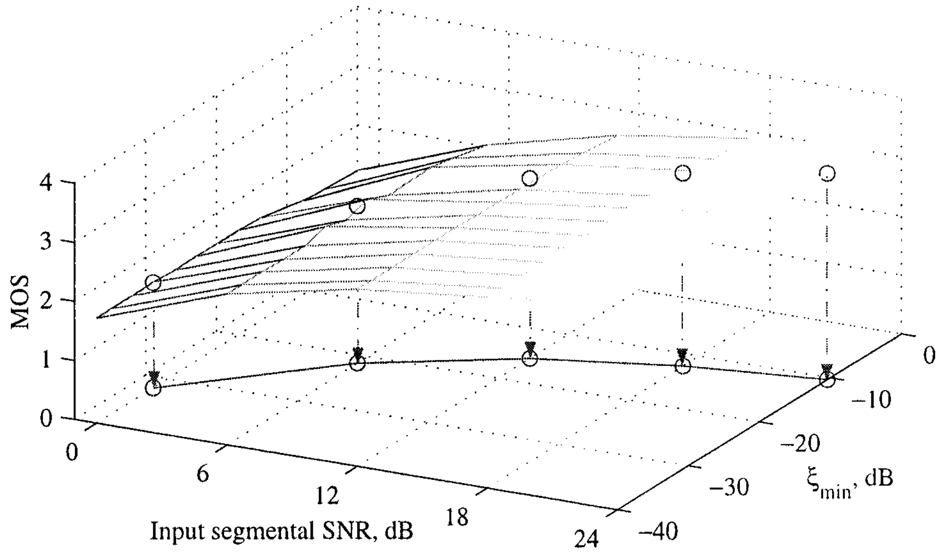


Fig. 3.6 Heuristic approach to defining ALSP for G.729 speech coder.

The line of best fit (solid line) for the two test cases is:

$$\xi_{min} = 0.0013(\text{SNR})^3 - 0.1(\text{SNR})^2 + 2.5(\text{SNR}) - 32, \tag{3.6a}$$

$$\xi_{min} = 0.0013(\text{SNR})^3 - 0.1(\text{SNR})^2 + 2.5(\text{SNR}) - 38, \tag{3.6b}$$

where ξ_{min} will be called the *Adaptive Limiting Scheme for Perception* (also referred to as MMSE-LSA with ALSP) in this thesis. Eq. (3.6a) corresponds to G.729 speech coder while Eq. (3.6b) when MELP coder is used instead. Note, that the only difference between the two equations is an offset of 6 dB. Thus, even though PESQ has not been validated for low bit-rate coders, results generated in conjunction with the 8 kbps G.729 seem to correlate well with the MELP coder. In this thesis, Eq. (3.6b) is applied to all the frames irrespective of them being a combination of speech and noise or noise only, prior to encoding with the MELP vocoder.

the decoder end, speech is reconstructed by filtering the excitation signal with the synthesis filter. After computing the reconstructed speech, it is further enhanced with a post-filter. In addition to the 10 ms frame delay, there is an additional 5 ms delay due to lookahead, [46].

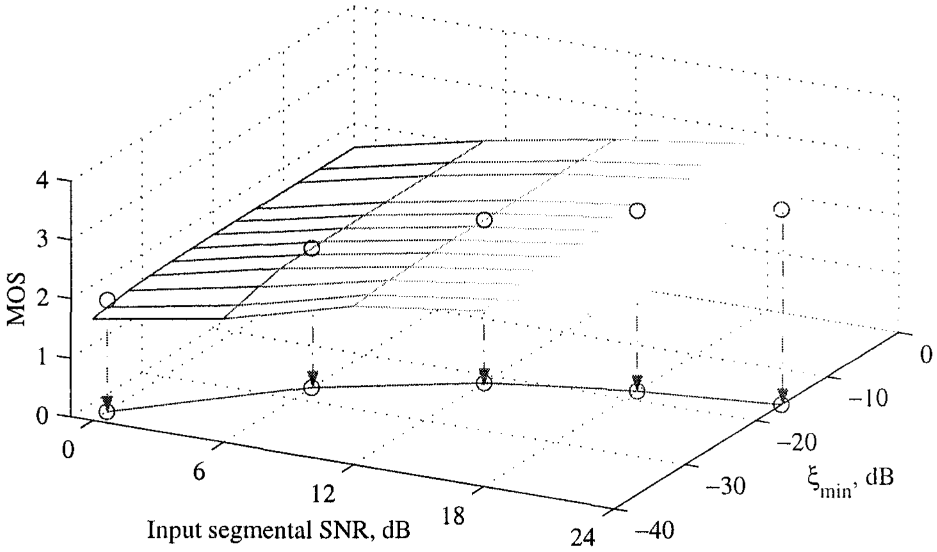


Fig. 3.7 Heuristic approach to defining ALSP for MELP speech coder.

A detailed analysis of the results obtained in the heuristic approach taken to find Eq. (3.6b), can be seen in Fig. 3.8. In Fig. 3.8-A five plots of MOS (female speech corrupted with computer generated white noise at five levels input SNR) are seen plotted against the twelve different values of the lower limit on ξ_k. The lower limits on ξ_k corresponding to the maxima are plotted against the corresponding levels of input SNR in Fig. 3.8-B. Similarly, male speech is used to obtain the dashed curve. The line of best fit defines the equation of the lower limit to be used on ξ_k, and is seen in Eq. (3.6b).

Thus, there are two new lower bounds on ξ_{min}: one that assures more accurate LSF estimation (using MMSE-LSA with ALSE) and the other that maximizes the MOS ratings (using MMSE-LSA with ALSP). Using these as Type L and Type R enhancement schemes in Fig. 3.4, respectively, will serve to reconstruct speech that is encoded at low-bit rates. Chapter 4, presents results using a subjective measure (A–B Comparison Test, see Section 4.3.1) and some objective measures discussed in Section 2.6.

The aim of the following section is to compare the various noise suppression rules discussed so far.

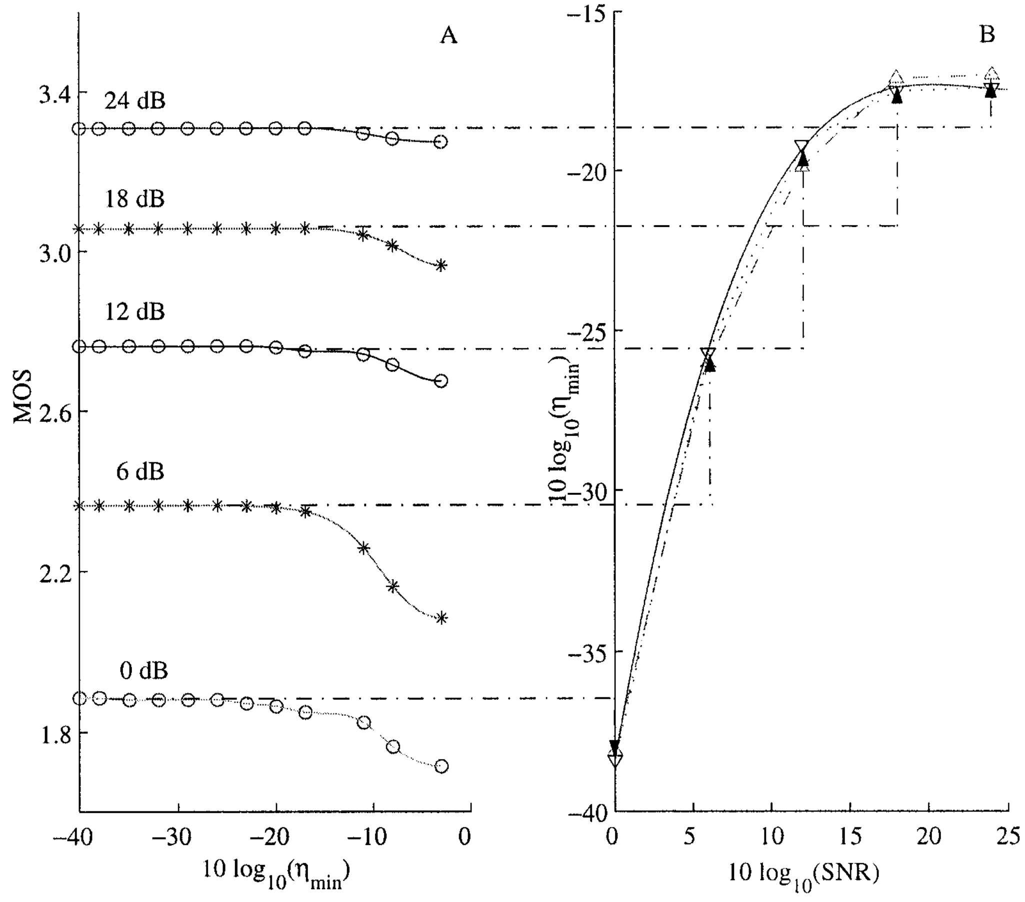


Fig. 3.8 Heuristic approach to defining MMSE-LSA with ALSP.

3.2.4 Comparison between the noise suppression rules

In order to understand the different noise suppression rules, discussed earlier, it would be informative to consider Fig. 3.9. This graph is indicative of the amount of attenuation applied to short-time noisy speech frames for a given input SNR (or, *a priori* SNR, ξ_k).

As noted earlier, EVRC noise suppressor sets up a floor at around −13 dB for noise only frames, Section 2.2. It is this very factor that assists it in reducing the annoying effects of musical noise and leaving a "natural sounding" background noise during speech/pause transitions. But, EVRC noise suppressor may not be optimal in the sense of parameter estimation. In fact, consider an input SNR of about 0 dB. At that input, MMSE algorithms offer an additional 3.5 dB attenuation over EVRC noise suppression rule. From a formant

Table 3.2 Levels of attenuation offered by various noise suppression rules vs. SNR of −20 dB.

Noise suppression rules	Gain in dB
EVRC	−13.00
MMSE-STSA	−16.22
MMSE-LSA	−16.94
MMSE-LSA with ALSE	−14.79
MMSE-LSA with ALSP	−16.94

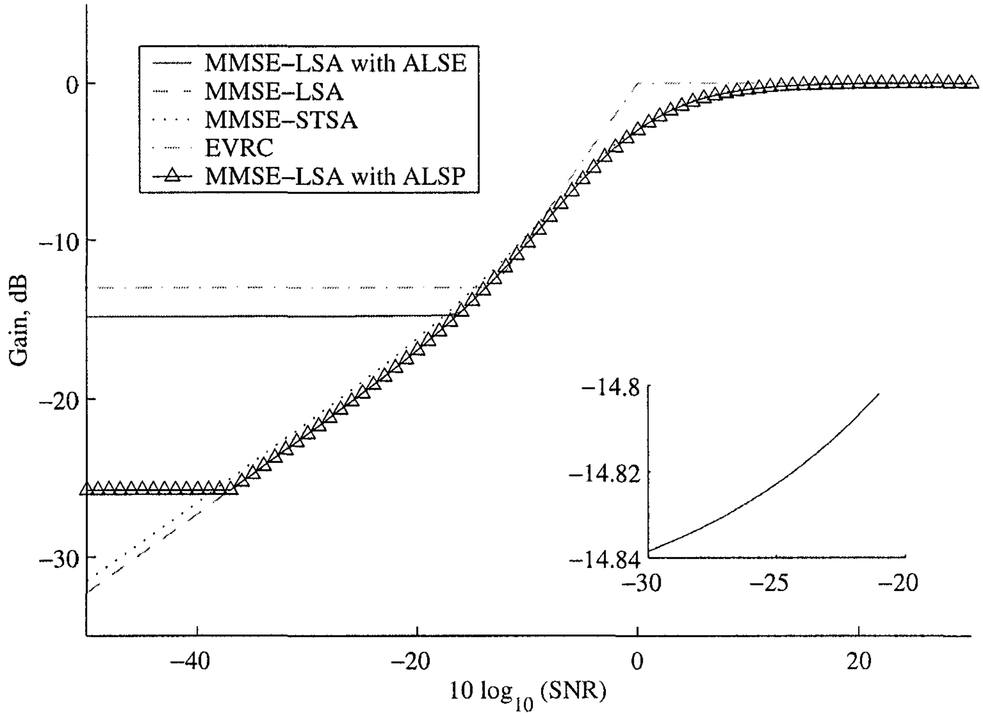
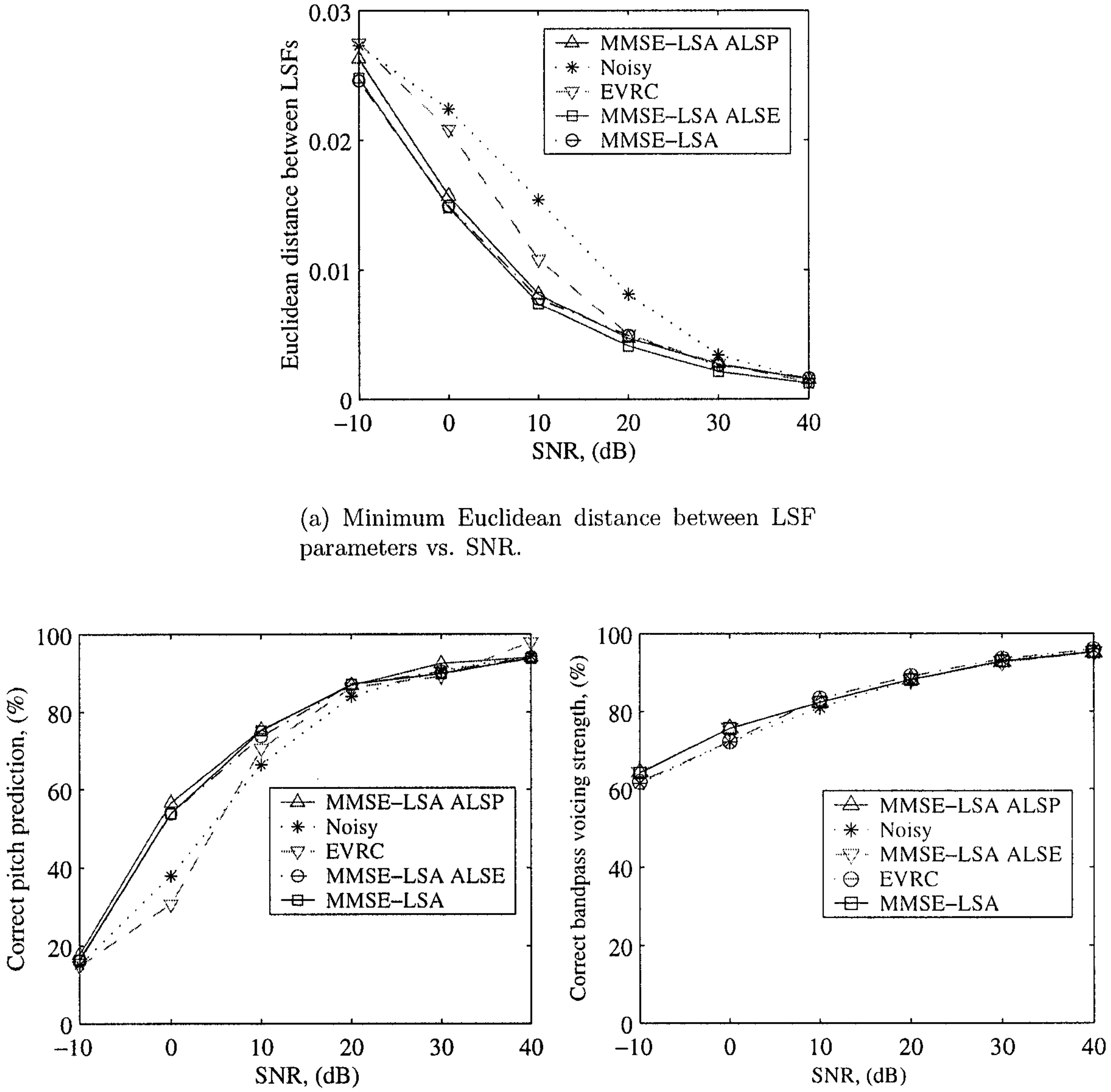


Fig. 3.9 Gain curves for various noise suppression rules vs. input SNR.

structure viewpoint, it means that MMSE algorithms tend to remove more noise from the valleys in between the formants. As was seen earlier in Fig. 3.3, removing more noise from the valleys helps in restoring the bandwidth of the formants (allowing a better estimation of LPC parameters). Hence, EVRC noise suppression block tends to perform poorly for LPC parameter estimation, as it leaves more noise in the spectral valleys. On the other hand, both MMSE-STSA and MMSE-LSA attenuate noise rather significantly. From Fig. 3.9 it is



(a) Minimum Euclidean distance between LSF parameters vs. SNR.

(b) Correct pitch detection vs. SNR.

(c) Correct bandpass voicing strength vs. SNR.

Fig. 3.10 Effect of different noise suppression rules on parameter extraction.

seen that the MMSE-LSA attenuates noise by an additional amount of 2 dB over MMSE-STSA noise suppression rule. This phenomenon can be explained by exploiting Jensen's inequality (seen in Section 2.4.1), which states that the log estimate is always smaller or equal to the actual estimate. MMSE it is seen that the log of the magnitude estimate is always smaller than or equal to the actual magnitude estimate.

Although MMSE-LSA offers more attenuation of noise under harsh acoustic conditions it is not necessarily optimal for LPC estimation. Fig. 3.9 also shows the gain curve for MMSE-LSA with ALSE. Note, that on the bottom right of the figure there is an expanded section of this graph that shows the different levels of attenuation applied to input frames. This is in contrast to the EVRC noise suppression where the applied gain is constant for frames with input SNR less than 3 dB. Hence, the name adaptive for the rule proposed by Martin and Cox, [15]. Similarly, the gain curve associated with MMSE-LSA with ALSP is seen as dash-triangle curve and has a 'sigmoid' shape. Thus, the noise suppression rule developed in this thesis attempts to take advantage of the high maximum attenuation offered by MMSE-LSA, and at the same time minimizes signal distortion by setting a threshold (at around −26 dB) for input speech SNR less than −35 dB.

How these noise suppression rules affect the various parameter estimation (such as pitch and bandpass voicing strength) for MELP speech coder is of interest. Some of the relevant distortion measures were discussed in Chapter 2.

Fig. 3.10, indicates three graphs that show the effect of additive noise on speech parameters used by the MELP encoder. From Fig. 3.10(a) it is apparent that ALSE Eq. (3.6) is the best choice for Type L enhancement. However, Fig. 3.10(b) shows that MMSE-LSA with ALSP is about 4% superior over other STSA algorithms and about 30% over EVRC noise suppression algorithm when used to extract pitch value from speech with input SNR of 0 dB. Fig. 3.10(c) presents percentage of accurate bandpass voicing strength extraction from corrupted speech. It is noticed that, although, EVRC noise suppression rule compares to other algorithms at speech SNR higher than 10 dB, it fails to achieve high accuracy for signals under harsh acoustic conditions. Thus, in the light of pitch prediction results it seems germane to use MMSE-LSA with ALSP as Type R enhancement scheme for low-bit rate speech coders.

3.3 Summary

This chapter looked at the new Federal standard on voice coding—MELP and the various parameters it extracts from speech to encode it. The harsh effects of background noise on parameter estimation and specialized pre-processors to circumvent these issues were discussed. Further, some results were presented that show use of one speech enhancement technique being more appropriate for parameter extraction, while another technique may be better suited for improved speech quality. The next chapter focuses on the validity of such a dichotomy between estimation and perception to produce a more robust speech enhancement block for low-bit rate speech coders.

Chapter 4

System Evaluation

This chapter presents evaluations of the performance of several speech denoising algorithms discussed earlier in this thesis. As noted in Chapter 3, different speech enhancement algorithms are tailored to extract different speech parameters. For instance, MMSE-LSA with ALSE is used as a noise suppression algorithm that will assist in better LP parameter estimation, while MMSE-LSA with ALSP was geared at maximizing the MOS rating. Therefore, a *two-branch* scheme is suggested that will serve to dichotomize LP parameter estimation and speech perception. It is speculated that such a scheme will improve the overall perceptual quality of encoded speech in the presence of background noise. To verify this speculation, some of the tests listed in Chapter 2 are conducted. The preparation of the test data is discussed in detail in Section 4.1. In Section 4.2, details of the experiments carried out in the laboratory are mentioned.

In this thesis, several combinations are tested subjectively using the A–B comparison test, see Section 4.3. Based on the results from these subjective tests (i.e., listeners' opinion), it is relatively easy to select the preferred speech enhancement combination. Later in Section 4.4 the preferred system is tested under several acoustic conditions in the presence of various noisy environments.

4.1 Preparing Test Data

The test data was prepared using the procedure described in Supplement 23 to ITU-T P-series Recommendations[1], [47]. In fact, the methodology described in the recommendation was used to test the speech quality of the G.729 speech coder. Hence, it seems appropriate to follow the instructions in the recommendation to prepare test data for this study.

4.1.1 Speech Material

The speech material consists of simple and meaningful sentences (referred to as utterances). These utterances are chosen from lists of phonetically balanced[2] utterances and are easy to understand. In 1969, the IEEE committee issued a six year study on phonetically balanced short utterances, which later became available as the Harvard list, [48]. The Harvard list consists of 720 utterances that have been grouped into 72 phonetically balanced lists of 10 utterances each. These utterances were recorded under controlled conditions at a sampling frequency of 48 kHz (for testing, these were downsampled to 8 kHz and filtered with a lowpass filter designated as the modified Intermediate Reference System (mIRS)[3]. A 16-bit (i.e., dynamic range of $[-32768, +32767]$) A/D converter is used for encoding.

According to ITU-T Series P recommendations, [47], the speech files should be made into sets of two or three in such a way that there is no obvious connection of meaning between the utterances in a set. Thus, a *speech test file* is generated with an initial silence (prologue to the sentences) of 0.3 s followed by two same gender utterances, separated with a silence (or pause) of 0.5 s, and terminated (epilogue to the sentences) with a silence of 0.2 s, [47]. For details on the speech test files used in this thesis see Table 4.1.

Once, the clean speech files are created, they are 'corrupted' with additive noise.

[1] ITU-T P-series are Recommendations on telephone transmission quality, telephone installations, local line networks.

[2] In phonetically balanced utterances, phonemes appear with a similar frequency to their occurrence in normal speech.

[3] The filter response is specified in ITU-T Recommendation G.712. The ITU-T models the overall telephone transmission system (including the effects of the transmitting and the receiving end as the Intermediate Reference System (IRS)). However, for speech coding purposes only part of this overall response is seen by the speech coder. Telecommunications Industry Association (TIA) and ITU-T for speech coders use mIRS that models only the transmitting side, [49].

Table 4.1 Speech files. Numbers in square brackets indicate lengths in seconds.

Speaker	[prologue]Utterance1[pause]Utterance2[epilogue]
Male1	[0.3]The lure is used to catch trout and flounder.[0.5]
	The hostess taught the new maid to serve.[0.2]
Male2	[0.3]A siege will crack the strong defense.[0.5]
	The thaw came early and freed the stream.[0.2]
Female1	[0.3]The birch canoe slid on the smooth planks.[0.5]
	Wipe the grease off his dirty face.[0.2]
Female2	[0.3]The nozzle of the fire hose was bright brass.[0.5]
	The beach is dry and shallow at low tide.[0.2]

4.1.2 Noise Material

The noise files suggested in the ITU-T Recommendation on subjective test plan, [47], have been used in this thesis. The noise types listed in the recommendation are: office babble, car noise, street noise, Hoth and music. For the purposes of calculating the level at which noise must be added, ITU-T Recommendation P.56 is used, [50]. Essentially, the SNR is determined as the ratio of the active level for speech to the RMS level of the recorded noise. Specifications for the measurement of the active level of speech signals are given as Method B in [50].

4.2 The Experiment

The aim of the experiments is to evaluate the performance of the MELP speech coder (in the presence of background noise) when its input is conditioned with the following four noise suppression algorithms: MMSE-LSA, MMSE-LSA with ALSE, MMSE-LSA with ALSP and the EVRC noise suppression rule. For the purposes of evaluation, noisy speech (without any enhancement) is also encoded.

Apart from these noise suppression rules, the evaluation also aims at testing the two-branch schemes (for instance, MMSE-LSA for LP parameter estimation, and the EVRC noise suppression for residual signal computation). It must be noted that a total of p^2 schemes can be produced, where p is the number of speech enhancement algorithm. In this thesis, 5 "enhancement algorithms" (4 noise suppression rules and 1 without any enhancement) are considered. Thus, there are 25 potential pre-processing schemes, see Table 4.3.

Table 4.2 Nomenclature for noise suppression rules.

Algorithm Number	Noise Suppression Rule
1	Noisy (no enhancement)
2	MMSE-LSA
3	MMSE-ALSE
4	MMSE-ALSP
5	EVRC noise suppression

To understand the nomenclature better consider the underlined combination "R4-L3" in Table 4.3. Here 'R' stands for speech enhancement for Residual signal computation (i.e., Type R enhancement), while 4 stands for the fourth algorithm (i.e., MMSE-LSA with ALSP, Table 4.2). Similarly, 'L' stands for speech enhancement for LP parameter estimation, while 3 stands for the third algorithm (MMSE-LSA with ALSE, Table 4.2). This example of nomenclature description is seen in Fig. 4.1.

Table 4.3 The combinations used to test the system of Fig. 3.4. The nomenclature of $Rn - Lm$ refers to n being used for Type R (i.e., Residual signal computation) while m for Type L enhancement (i.e., LP parameter estimation).

Type L Enhancement	Type R Enhancement				
	Noisy (R1)	LSA (R2)	ALSE (R3)	ALSP (R4)	EVRC (R5)
Noisy (L1)	**R1-L1**	R1-L2	**R1-L3**	R1-L4	R1-L5
LSA (L2)	R2-L1	R2-L2	**R2-L3**	R2-L4	R2-L5
ALSE (L3)	R3-L1	R3-L2	R3-L3	R3-L4	R3-L5
ALSP (L4)	R4-L1	R4-L2	R4-L3	R4-L4	R4-L5
EVRC (L5)	R5-L1	R5-L2	R5-L3	R5-L4	**R5-L5**

In order to decrease the complexity of the experimentation, initially only one noise type—car noise with an SNR of 10 dB is considered. Preliminary subjective results are obtained with car noise (see Section 4.3.2) and are used to identify the preferred speech enhancement scheme from the 25 schemes listed in Table 4.3. Followed by this, objective measures are used to qualify the reliability of the system in the presence of other background noises at varying SNR.

For the preliminary evaluation, two male and two female speech files (Table 4.1) are contaminated with car noise at 10 dB. These files are then high-pass filtered (as done by

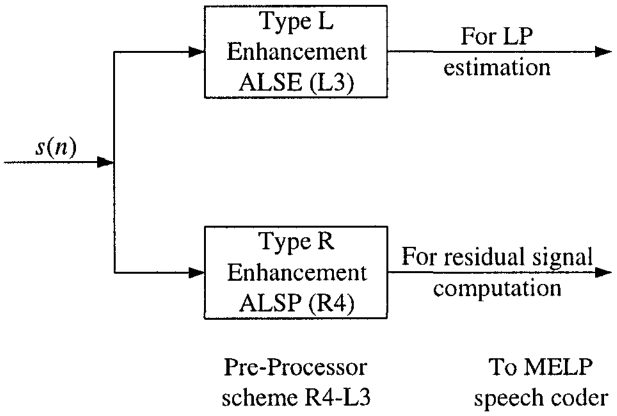


Fig. 4.1 Basic system under test.

the EVRC standard, Section 2.1.3). The files are then level-aligned with a clean reference signal, prior to sending them through the four noise suppression algorithms, see Table 4.2. After denoising the speech files they encoded using the MELP speech coder. Subjective tests are performed on the encoded speech, Section 4.3. The files are once again level-aligned so that they have the same playback volume. These files have also been time-aligned, and its importance will be seen in Section 4.4.2.

4.3 Subjective Results

In Chapter 2, several subjective measures were listed. Most of those methods require rigid testing environments. A simpler and more amenable subjective test is the A–B comparison test and is described in Section 4.3.1.

4.3.1 A–B Comparison Test

The A–B comparison test involves presenting listeners with a sequence of two speech test files[4] (A and B). After listening to them once, they have to decide whether they preferred file A or file B. If they have no preference they are allowed to choose as "don't care". It may happen that the listener is slightly biased to the second file (B in this case), since that is the last (s)he heard. This will be referred to as the *order bias*. To reduce this bias, listeners are provided with both A–B and B–A orders at different times. Consider the

[4] Recall that a test file has the structure defined in Table 4.1.

playlist in Table 4.4. In row 2 and 4 the order of the files is reversed (i.e., the listeners are first presented combination R4-L3–R5-L5 and then the order is reversed to R5-L5–R4-L3). However, the underlying speaker (female in this case) is the same while the utterances change (Female1 in A–B case, while Female2 in B–A case). Further, the reversed order is played after at least one other combination (Female1 R2-L3–R4-L3, in this instance) is played. Such a spacing between reversed playback, allows listeners to forget about the first combination (i.e., R4-L3–R5-R5 in this example).

Table 4.4 Sample playlist for the A–B comparison test.

Speaker	A	B
Male1	R1-L1	R2-L3
Female1	R4-L3	R5-L5
Female1	R2-L3	R4-L3
Female2	R5-L5	R4-L3
⋮	⋮	⋮

Complexity in A–B Comparison tests

Assume that there are p algorithms to be tested against each other. This would require at least $2(p^2 - p)$ speech test files, since a given algorithm is *not* compared with itself and the files are duplicated to perform both A–B and B–A tests. In this thesis, there are 25 combinations to be tested against each other. This would translate to a total of 1200 test files to be played to listeners. To reduce this number of test files, preliminary informal tests were conducted (three subjects were used) to eliminate some of the 25 combinations. The three listeners eliminated combinations that were terribly annoying, either due to muffled speech output or the presence of tonal thumps. The five combinations that were selected for exhaustive A–B comparison tests appear in boldface in Table 4.3. For convenience these are listed again: R1-L1, R1-L3, R2-L3, R4-L3 and R5-L5. A detailed description of these schemes is shown in Fig. 4.2. Note that prior to any sort of testing, subjective or objective, the output from all schemes is coded with the MELP speech coder. For the rest of this thesis, when referring to the output of any scheme, it is implied that it is coded as well.

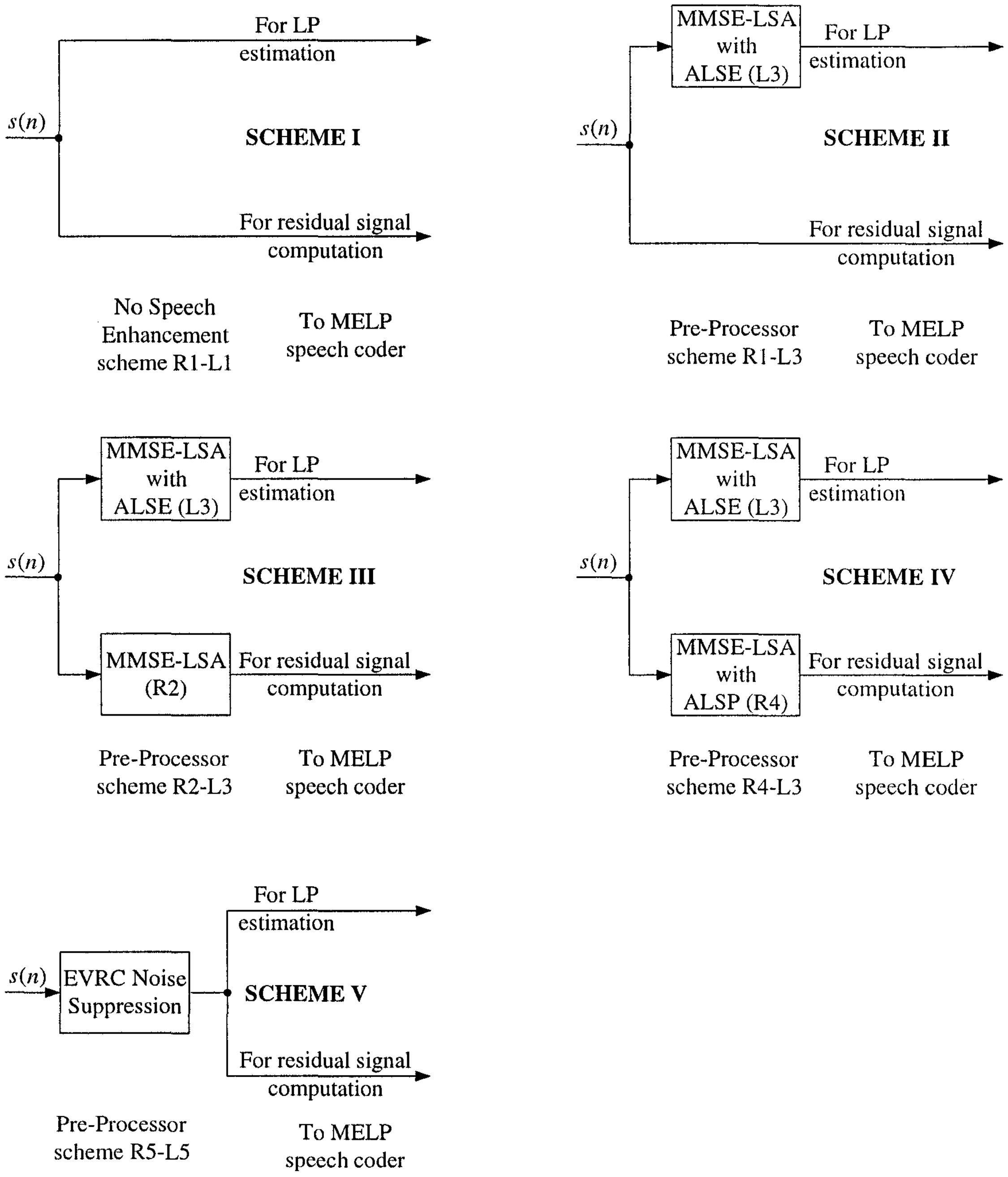


Fig. 4.2 Schemes on which the A–B was conducted.

4.3.2 A–B Test Results

Male1, Male2, Female1 and Female2[5] (from Table 4.1) were used 10 times each, while making sure that the same speaker with different utterances is used. A team of 15 listeners (2 female and 13 male), aged between 19–26 were taken from the faculty of Engineering at McGill University.

Table 4.5 Lower Triangle (A–B) and Upper Triangle (B–A). The format of xA - d - yB, refers to x number of listeners preferring A, d denotes the "don't care" responses and y number preferring B.

A	B				
	R1-L1	R1-L3	R2-L3	R4-L3	R5-L5
R1-L1	×	0A - 1 -14B	8A - 1 - 6B	8A - 0 - 7B	2A - 9 - 4B
R1-L3	8A - 4 - 3B	×	11A - 2 - 2B	12A - 0 - 3B	*11A - 2 - 2B*
R2-L3	13A - 1 - 1B	13A - 0 - 2B	×	2A - 3 -10B	7A - 0 - 8B
R4-L3	13A - 1 - 1B	14A - 0 - 1B	4A - 6 - 5B	×	1A - 0 -14B
R5-L5	5A - 2 - 8B	*2A - 0 -13B*	14A - 0 - 1B	0A - 2 -13B	×

Table 4.5 shows the A–B and B–A comparison test results. In this table the lower triangle is representative of the A–B tests, while the upper triangle is an indication of the B–A tests. For instance consider the underlined entry in the lower triangle (i.e., A–B test) of the table (13A - 1 - 1B). This implies that the listeners are presented with R2-L3 (i.e., Scheme III) and then R1-L1 (i.e., Scheme I) in that order. From the raw data it is seen that 13 listeners preferred A, 1 did not have any preference over the two files and 1 listener preferred B. To circumvent the order bias, the B–A test is performed. The results are seen in the upper triangle of the table (8A - 1 - 6B). Thus, when the speech test files are played back in the reverse order (R1-L1 first and R2-L3 second), it is seen that only 8 (as opposed to 13 in the A–B test) preferred A, 1 did not have any preference and 6 preferred B (as opposed to 1, when the files were played in the A–B order).

Playing the files in reverse order does may not always discount the order bias. For instance, consider the italicized entries in Table 4.5. In the A–B test (lower triangle) 13 listeners prefer B over A (B being the second file played), while from its counterpart during B–A testing (upper triangle), 11 listeners prefer A over B (where A is now played second). This implies that the

[5] Each speech test file was corrupted with 10 dB car noise and denoised using different algorithms. It would have been ideal to try 5 and 20 dB and other noise types as well, but to avoid listener strain the tests were limited to 10 dB car noise only.

listeners were biased as they preferred the second speech file regardless of the playback order being reversed.

4.3.3 Summary of A–B Test Results

Table 4.6 shows the aggregate of the A–B and the B–A tests. These are then computed as percentage. Readers are pointed to Fig. 4.2, to better understand the inferences drawn from these A–B listening tests.

Table 4.6 Combined A–B and B–A results. The format of xA - d - yB, refers to x percentage of listeners preferring A, d denotes the "don't care" responses and y percentage preferring B and .

A	B				
	R1-L1	R1-L3	R2-L3	R4-L3	R5-L5
R1-L1	×				
R1-L3	27A -16 -57B	×			
R2-L3	70A - 7 -23B	80A - 7 -13B	×		
R4-L3	70A - 3 -27B	87A - 0 -13B	20A -30 -50B	×	
R5-L5	23A -37 -40B	43A - 7 -50B	70A - 0 -30B	3A - 7 -90B	×

- The majority of the listeners preferred R4-L3 (i.e., Scheme IV) over R5-L5 (i.e., Scheme V, which is the EVRC noise suppression block only as noise suppression pre-processor for each of the two branches) scheme (90%). This implies that the EVRC noise suppression block is not an ideal pre-processing stage for low bit-rate speech coding. In fact, listeners preferred no enhancement of noisy speech as input to vocoders (R1-L1 or Scheme I) rather than pre-enhancing it with EVRC noise suppression rule (40% prefer no enhancement over 23% preferring EVRC noise suppression and 37% have no preference). As seen earlier in Section 3.2.4, EVRC noise suppression tends to floor attenuation to −13 dB, thus filling spectral valleys with noise (and hence improving auditory impression) and disrupting accurate LP estimation.

- Both R4-L3 and R2-L3 perform equally well as pre-processors. This is seen in their comparable performance against R1-L1 and R1-L3. Against R5-L5, 90% listeners prefer R4-L3, while only 70% preferred R2-L3. However, when R4-L3 is compared against R2-L3, it is seems that 50% preferred R2-L3, while 30% did not care.

From these results, it seems that R2-L3 and R4-L3 schemes are the most preferred preprocessors to low bit-rate speech coding. These will be referred to as Scheme III and Scheme IV, respectively, in the remainder of the thesis. In the following sections, these schemes are tested against different noise types at various SNR values using some of the objective distortion measure described in Chapter 2.

4.4 Testing the System for other Noise Types

From the A–B tests it is seen that there are at least two schemes that listeners prefer over the others. To avoid listener fatigue and exhaustive data collection, the selected schemes (III and IV) are tested for performance under other acoustic conditions (varying SNR) in the presence of different noise environments. The purpose of this section is to look at some objective measures that were discussed in Chapter 2. Segmental SNR and SNR improvements are meaningful when noise is additive. However, noise introduced by vocoders (e.g., MELP speech coder) has correlated noise components, [34, 37], and hence its computation is avoided in this thesis.

4.4.1 Mean Opinion Score

In Section 2.6.1, a detailed description of MOS ratings was presented. That section also introduced the ITU-T standard P.861 on Perceptual Evaluation of Speech Quality (PESQ). Although it has not been validated for speech enhancement schemes, it is used in this thesis to obtain MOS ratings. Both Scheme III and Scheme IV are tested against four different noise types—Hoth noise, car noise, office babble and music. Supplement 23 on ITU-T P series recommends use of these noise types at SNR values of 10 dB and 20 dB. In addition, in this thesis, a more noisy condition (SNR value of 5 dB) is tested. Fig. 4.3 shows plots of MOS ratings for speech (Male1 speech file, Table 4.1) corrupted with the aforementioned noise types. As required by PESQ, the *reference file* is the uncoded noisy file (with the specified SNR), while the *test file* is the enhanced speech file obtained by using Scheme III (dashed curve) and Scheme IV (solid curve) and are each encoded using the MELP speech coder.

It must be noted that the MOS ratings seen in this figure are in the order expected for low bit-rate speech coders, Fig. 1.2.

Two points can be inferred from these results:

- In all cases, for low input SNR (< 10 dB) Scheme IV outperforms Scheme III with a MOS improvement in the range of 0.04–0.1. This result can be explained by considering Fig. 3.9, which shows the gain curves for the various noise suppression rules. In that graph MMSE-

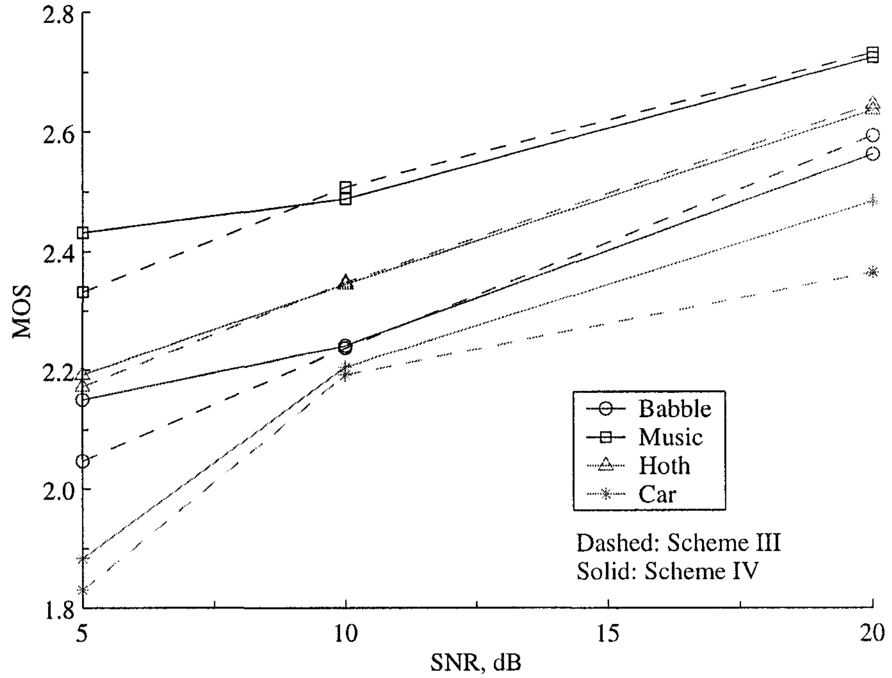


Fig. 4.3 MOS ratings for several SNR in the presence of different background noise.

LSA with ALSP (or algorithm number 4) tracks the MMSE-LSA (or algorithm number 3) until very low *a priori* SNR values (about −35 dB). At lower than −35 dB MMSE-LSA with ALSP, tends to attenuate the signal less thus trading off between signal distortion and amount of noise attenuation better than MMSE-LSA. Hence, for low SNR values MMSE-LSA with ALSP tends to show higher MOS ratings, as it tends to leave a little more noise in the spectral valleys. For larger SNR, the MOS ratings are almost similar with the exception when background disturbance is car noise (20 dB point on the star-solid/dashed curve).

- The steepness of the curves is the same, except for the car noise case. Between 5–10 dB, car noise case MOS rating is the steepest (almost a 0.4 improvement in MOS over 10 dB) after which it tends to follow the steepness of other curves. Unlike other noise types, car noise has high energy content in low frequencies. Thus, at low SNR of 5 dB, the high frequency content of speech is preserved.

4.4.2 Objective Results

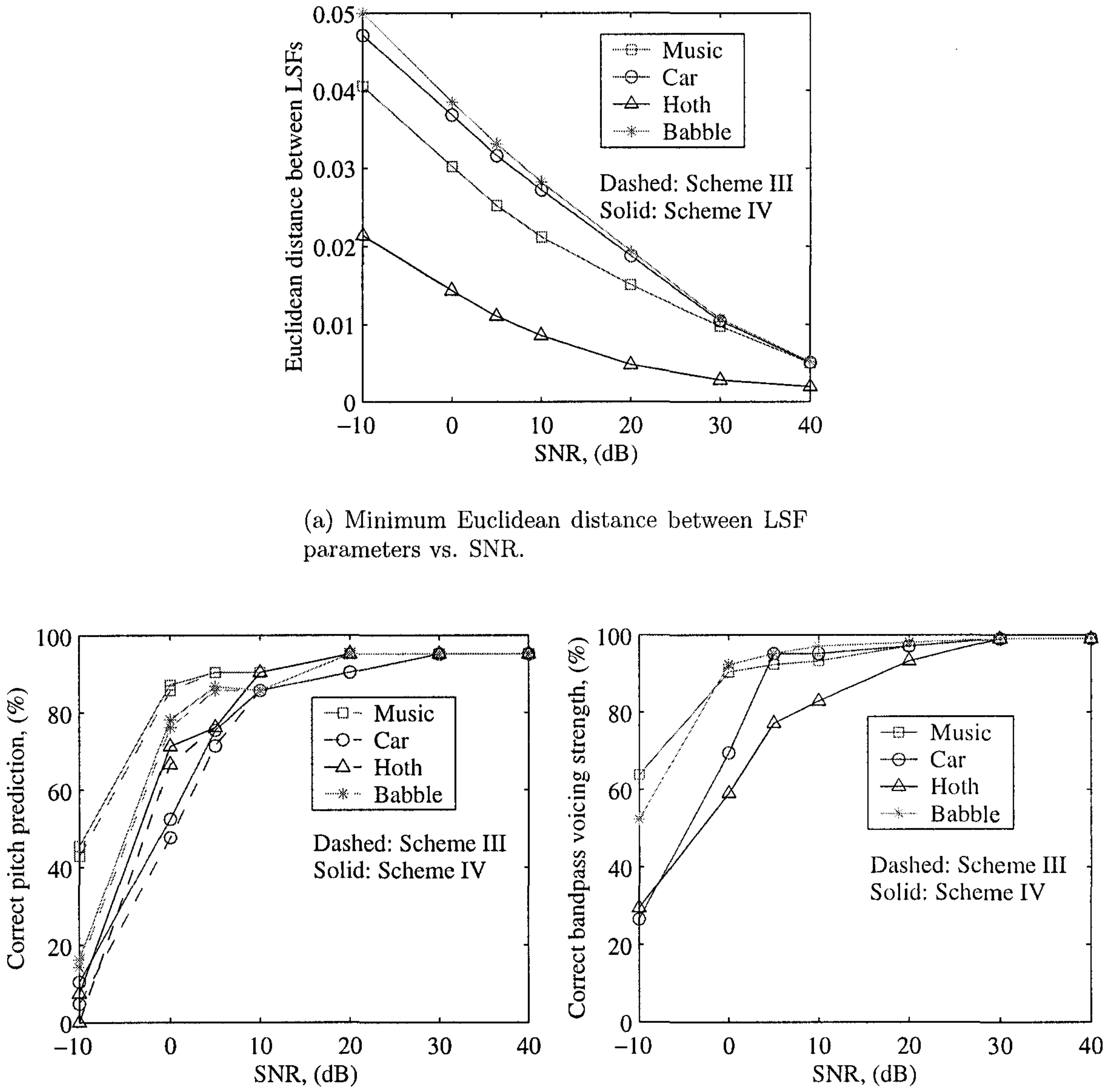
This subsection shows how the two schemes perform against each other in extracting speech parameters (in the presence of several noise types) required by the MELP speech coder. In Fig. 4.4, three plots are presented.

Fig. 4.4(a) shows similar results for either Scheme III (MMSE-LSA with ALSE for LP estimation and MMSE-LSA for residual signal computation) and Scheme IV (MMSE-LSA with ALSE for LP estimation and MMSE-LSA with ALSP for residual signal computation), Fig. 4.2. This does not come as a surprise, since in either scheme the same speech enhancement (MMSE-LSA with ALSE) is used to estimate the LPC parameters. Fig. 4.4(b) presents results on percentage correct pitch prediction. In this graph it is seen that Scheme III performs slightly better (about 2–5%) when the SNR is in the range -10 to 10 dB. However, for higher SNR, enhancing speech with either scheme produces almost identical results. In Fig. 4.4(c), results of percentage correct bandpass voicing strength are plotted. It is noted that for either scheme the results are almost identical.

4.5 Summary

This chapter looked at the methodology adopted to evaluate a speech enhancement system as a pre-processor to low bit-rate speech coding. Such a speech enhancement system is one that uses different noise suppression schemes to extract speech parameters (in the presence of background noise) used by the MELP speech coder.

Speech files are corrupted with additive car noise (such that the SNR is 10 dB) and presented to listeners, to obtain their preference (using the A–B comparison test) on the various combinations studied in Chapter 2 and Chapter 3. From the listeners choices two schemes were selected—Scheme III and Scheme IV, see Fig. 4.2. Followed by this, the two schemes are tested for various acoustic conditions (varying SNR values) in the presence of office babble, music, Hoth and car noise. From the results obtained in Fig. 4.3 and Fig. 4.4, it is noted that Scheme IV outperforms Scheme III slightly under harsh acoustic conditions (ranging from -10 to 10 dB). However for benign environments ($>$ 10 dB), either scheme can be used, as the results obtained are very similar. Therefore, for low bit-rate coding it is seen that Scheme IV (MMSE-LSA with ALSE for LP estimation and MMSE-LSA with ALSP for residual signal computation) can be used for better auditory impression of MELP encoded speech even under harsh background noise.



(a) Minimum Euclidean distance between LSF parameters vs. SNR.

(b) Correct pitch detection vs. SNR.

(c) Correct bandpass voicing strength vs. SNR.

Fig. 4.4 Effect of Scheme III and Scheme IV on parameter extraction.

Chapter 5

Conclusion

Improving the perceptual quality of degraded speech that is encoded with a 2.4 kbps voice coder is the main goal of this thesis. In this work, a modification to the existing Ephraim and Malah Minimum Mean Square Error of the Log Spectral Amplitude (MMSE-LSA) estimator is introduced. The motive is to maximize the Mean Opinion Score (MOS) ratings of the reconstructed signal. It is shown that by using two different speech enhancement algorithms (referred to as the *two-branch* pre-processor in this thesis), improvements in perceptual quality can be achieved. The purpose of a two-branch scheme is to have one denoising algorithm that is tailored for LP estimation, while the other for improved perception. Such a pre-processing system has demonstrated good robustness properties with respect to the other noise types under varying acoustic levels (or SNR conditions).

5.1 Thesis Summary

Chapter 2, starts off by introducing the fundamentals of noise suppression algorithms. It considers the assumptions, constraints and noise variance estimation issues that are mandatory to set up a framework for building a noise suppression algorithm. This chapter focuses on the importance of Short Time Spectral Amplitude (STSA) estimators and the venerable noise suppression algorithm described in the Enhanced Variable Rate Coder (EVRC). Two STSA algorithms that were developed by Ephraim and Malah in 1984 (MMSE-STSA) and in 1985 (MMSE-LSA) are studied in further detail. Followed by this, the chapter looks at the limitations of noise suppression algorithms; such as signal distortion, musical noise, and temporal smearing. This chapter is concluded with a discussion of several distortion measures (both objective and subjective) that are used to qualify effectiveness of speech enhancement algorithms.

The purpose of Chapter 3 is two-fold. Section 3.1 starts off by arguing why the Mixed Excited Linear Prediction (MELP) speech coder is used as testbed for the derived speech enhancement pre-processor. It then describes the details of the relevant speech parameters (such as LP parameters, pitch, bandpass voicing strength etc.) used by the encoder. This is followed by showing how these parameters are disturbed in the presence of noise.

Section 3.2 gets into a specialized pre-processor that is tailored to estimate LP parameters accurately. This pre-processor is referred to as the MMSE-LSA with Adaptive Limiting Scheme for Estimation (ALSE), as it is built on the MMSE-LSA framework. The heuristic approach taken by Martin and Cox in 1999 to tune this denoising algorithm is explored extensively, and a similar procedure is adopted in this thesis to derive another noise suppression rule that aims at maximizing MOS ratings. This newly derived algorithm is referred to as MMSE-LSA with Adaptive Limiting Scheme for Perception (ALSP). Later in this chapter, a comparison and discussion of the various noise suppression rules (discussed in Chapter 2 and Chapter 3) is made. This chapter ends by showing improvements of these specialized pre-processors insofar as LPC parameter extraction, percentage correct pitch estimation and percentage correct bandpass voicing strengths are concerned. From these preliminary objective plots it is seen that MMSE-LSA with ALSE estimates LPC parameters more accurately than the others rules. On the other hand, MMSE-LSA with ALSP has a higher percentage of correct pitch prediction.

From Chapter 3, it is concluded that different denoising algorithms are beneficial for speech parameter extraction (such as LPC, pitch prediction etc.) required by the MELP speech coder. Thus, it seems appropriate to define a speech pre-processor with two branches (each with a specific denoising algorithm) so as to extract different speech parameters for the purposes of encoding. The purpose of Chapter 4, is to evaluate the performance of the two-branch speech pre-processor. The experimental protocol of Supplement 23 on ITU-T for P-series is adopted to create speech test files. Owing to the complexity and the time required in performing subjective testing (the A–B/B–A comparison test), 5 different combinations of the two-branch pre-processor schemes are tested with one noise type—car noise with an initial Signal-to-Noise-Ratio (SNR) of 10 dB. Results from the A–B comparison test allow for the selection of 2 most preferred speech enhancement schemes. These are referred to as Scheme III (MMSE-LSA with ALSE for LP estimation and MMSE-LSA for residual signal computation) and Scheme IV (MMSE-LSA with ALSE for LP estimation and MMSE-LSA with ALSP for residual signal computation). In fact, analysis of these subjective results show that Scheme IV is preferred by 90% listeners over using the EVRC noise suppression rule in each of the two-branches as a pre-processor to the MELP speech coder.

After having selected Scheme III and Scheme IV as potential candidates for the two-branch pre-processor, they are tested for different noise types—Hoth, office babble, music and car noise at varying initial SNR. ITU-T recommendation P.862 on Perceptual Evaluation of Speech Quality (PESQ)[1] is used to obtain MOS ratings for the two schemes under varying acoustic conditions (i.e., various initial SNR values ranging from 5–20 dB). From the results it is seen that Scheme IV outperforms Scheme III at very low SNR ($<$ 10 dB) by MOS ratings in the range 0.04–0.1. The chapter concludes by testing the two schemes against each other based on their accuracy of parameter extraction from speech corrupted with the other noise types and SNR values ranging from -10 to 40 dB. These results confirm the supremacy of Scheme IV over Scheme III under harsh acoustic conditions ($<$ 10 dB). Thus, it is shown that Scheme IV can be used to process speech prior to coding even under harsh noisy environments.

5.2 Future Research Directions

The purpose of this section is to look at some issues that were not considered in this thesis. The complexity of the two-branch pre-processor is almost doubled. Therefore it would be interesting to merge the benefits of MMSE-LSA with ALSE and MMSE-LSA with ALSP into one algorithm. Although, Scheme IV showed perceptual improvement, it would be useful to see the effects of using the masking properties of the human ear. Inclusion of a post-filter (that is perceptually weighted) will allow for some masking of low level distortions and perceptible tonal thumps, that occur in low bit-rate coded speech.

As seen in Chapter 2, the accuracy of a Voice Activity Detector (VAD) decision is essential in denoising corrupted speech. Its inaccuracy can either lead to poor noise variance estimation or insufficient noise removal or aggressive signal attenuation. Lack of robust VADs (that which produce accurate decisions on noisy speech), will lead to poor denoising. Presently, the algorithms derived in this thesis, takes *two* inputs: (1) the corrupted speech file, and (2) pre-estimated VAD decisions. In order to make the algorithm Single-Input-Single-Output (SISO), it would be desirable to incorporate a VAD in the algorithm. Such a VAD would have to be robust under varying noisy conditions. Besides, most industry VADs are hard-decision rules. This is to say that they return either a '1' or a '0' if the speech segment is active or pause/silence, respectively. Currently, noise variance is estimated from speech pauses/silence. If a "soft-decision" VAD were available, better knowledge of noise would be obtained from corrupted speech data.

In Section 2.5.3, it was seen that the accurate phase information becomes necessary for speech corrupted at SNR $<$ 2.5 dB. Thus far, only the magnitude of the clean speech is estimated from the noisy observation. The estimate is appended with the phase of the noisy signal to obtain the

[1] Note that PESQ has neither been validated for low bit-rate coded speech not for speech resulting from speech enhancement schemes, see Section 2.6.1.

'enhanced' speech frame. It would be worthwhile to incorporate a phase estimation algorithm to see if perceptual quality of coded signal can be improved under such low SNR conditions.

Unfortunately, owing to the complexity of subjective tests, the proposed two-branch schemes could not be subjectively tested for other noise types and varying acoustic conditions. Thus, only car noise with 10 dB SNR was tested in the laboratory using the A–B comparison test. Results drawn from this, were used to speculate performance of the scheme in the presence of other noise types using some objective measures. Therefore, it would be interesting to perform more A–B tests to see if the trends observed hold. In this thesis the MELP speech coder is used as a testbed. How well Scheme IV performs for other speech coders, such as G.729, Adaptive Multi-Rate (AMR) speech coder, would be worth analyzing.

References

[1] W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis.* Amsterdam: Elsevier, 1995.

[2] A. S. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, pp. 1541–1582, Oct. 1994.

[3] A. V. McCree and T. P. Barnwell III, "Implementation and evaluation of a 2400 bps mixed excitation LPC vocoder," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 2, (Minneapolis, Minnesota), pp. 159–162, Apr. 1993.

[4] C. F. Teacher and D. Coulter, "Performance of LPC vocoders in a noisy environment," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Washington DC, USA), pp. 216–219, Apr. 1979.

[5] J. S. Lim, *Speech Enhancement.* Englewood Cliffs, New Jersey: Prentice Hall, 1983.

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

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

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[9] M. C. Recchione, "The enhanced variable rate coder: Toll quality speech for CDMA," *Int. J. of Speech Tech.*, pp. 305–315, 1999.

[10] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Speech Processing*, vol. 39, pp. 1732–1742, Aug. 1991.

[11] A. V. McCree and T. P. Barnwell, "Improving the performance of a mixed excitation LPC vocoder in acoustic noise," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 2, (San Francisco, USA), pp. 137–140, Mar. 1992.

[12] L. Arslan, A. McCree, and V. Viswanathan, "New methods for adaptive noise suppression," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 1, (Detroit, Michigan), pp. 812–815, May 1995.

[13] J. S. Collura, D. F. Brandt, and D. J. Rahikka, "The 1.2kbps/2.4kbps MELP speech coding suite with integrated noise pre-processing," in *Proc. Conf. on Military Comm.*, vol. 2, (Atlantic City, USA), pp. 1449–1453, Oct. 1999.

[14] J. S. Collura, "Speech enhancement and coding in harsh acoustic noise environments," in *Proc. IEEE Workshop on Speech Coding*, (Porvoo, Finland), pp. 162–164, May 1999.

[15] R. Martin and R. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Proc. IEEE Workshop on Speech Coding*, (Porvoo, Finland), pp. 165–167, May 1999.

[16] A. J. Accardi and R. V. Cox, "A modular approach to speech enhancement with an application to speech coding," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 1, (Phoenix, Arizona), pp. 201–204, Mar. 1999.

[17] D. F. Hoth, "Room noise spectra at subscribers' telephone locations," *J. Acoustical Society America*, vol. 12, pp. 499–504, Apr. 1941.

[18] S. Haykin, *Communication Systems.* New York: John Wiley & Sons, Inc., third ed., 1994.

[19] A. L. Garcia, *Probability and Random Processes for Electrical Engineering.* Reading: Addison-Wesley Publishing Company, second ed., 1993.

[20] D. O'Shaughnessy, *Speech Communications: Human and Machine.* New York: IEEE Press, second ed., 2000.

[21] *Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems*, Jan. 1996. TR–45, PN–3292 {to be published as IS-127}.

[22] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals.* New York: IEEE Press, 2000.

[23] K. El-Maleh and P. Kabal, "Comparison of voice activity detection algorithms for wireless personal communications systems," in *Canadian Conf. on Elect. and Comp. Eng.*, vol. 2, (St. Johns, Canada), pp. 470–473, May 1997.

[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[25] H. L. Van Trees, *Detection, Estimation and Modulation Theory.* New York: John Wiley & Sons, Inc., 1968.

[26] D. Malah, V. C. Richard, and J. A. Anthony, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, pp. 789–792, Mar. 1999.

[27] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 434–444, May 1968.

[28] T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach.* New York: Academic Press Inc., 1967.

[29] R. M. Gray, A. Buzo, J. Gray, A. H., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-28, pp. 367–376, Aug. 1980.

[30] O. Cappé, "Elimination of musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 345–349, Apr. 1994.

[31] G. A. Soulodre, "Adaptive methods for removing camera noise from film soundtracks," Master's thesis, McGill University, Montreal, Canada, Nov. 1998.

[32] P. Vary, "Noise suppression by spectral magnitude estimation-mechanism and theoretical limits," *Signal Processing*, vol. 8, pp. 387–400, July 1985.

[33] S. Dimolitsas, "Objective speech distortion measures and their relevance to speech quality assessments," *IEEE Proc. I Communications, Speech and Vision*, vol. 136, pp. 317–324, Oct. 1989.

[34] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality.* Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[35] A. W. Rix, J. G. Beerends, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs.," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 2, (Salt Lake, UT), pp. 749–752, May 2001.

[36] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, Feb. 2001. P.862.

[37] D. Jamieson, L. Deng, M. Price, V. Parsa, and J. Till, "Interaction of speech disorders with speech coders: effects on speech intelligibility," in *Conf. on Spoken Lang. Proc.*, vol. 2, (Philadelphia, PA), pp. 737–740, Oct. 1996.

[38] G. Guilmin, R. Le Bouquin-Jeanns, and P. Gournay, "Study of the influence of noise preprocessing on the performance of a low bit rate parametric speech coder," *Proc. Europ. Conf. on Speech Comm. and Tech.*, vol. 5, pp. 2367–2370, Sept. 1999.

[39] M. Kohler, "A comparison of the new 2400 BPS MELP federal standard with other standard coders," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 2, (Munich, Germany), pp. 1587–1590, Apr. 1997.

[40] Department of Defence Digital Voice Processing Consortium, *Specifications for the Analog to Digital Conversion of Voice by 2,400 Bits/Second Mixed Excited Linear Prediction*, May 1998. Draft.

[41] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using chebyshev polynomials," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-34, pp. 1419–1426, Dec. 1986.

[42] T. Honkanen, J. Vainio, K. Jarvinen, P. Haavisto, R. Salami, C. Laflamme, and J.-P. Abdoul, "Enhanced full rate speech codec for IS-136 digital cellular system," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 2, (Munich, Germany), pp. 731–734, Apr. 1997.

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

[44] A. J. Accardi, "A modular approach to speech enhancement with an application to speech coding," Master's thesis, Massachusetts Institute of Technology, Boston, USA, May 1998.

[45] R. Martin, I. Wittke, and P. Jax, "Optimized estimation of spectral parameters for the coding of noisy speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 3, (Istanbul, Turkey), pp. 1479–1482, June 2000.

[46] ITU-T, *Coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, Mar. 1996. ITU-T Recommendation G.729.

[47] ITU-T, *ITU-T coded-speech database*, Feb. 1998. Supplement 23 to ITU-T P-series Recommendations.

[48] "IEEE recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–247, 1969.

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

[50] ITU-T, *Objective measurement of active speech level*, Mar. 1993. ITU-T Recommendation P.56.