ABSTRACT

Title of dissertation: SYNERGY OF ACOUSTIC-PHONETICS AND AUDITORY MODELING TOWARDS ROBUST SPEECH RECOGNITION

Om D. Deshmukh, Doctor of Philosophy, 2006

Dissertation directed by: Dr. Carol Y. Espy-Wilson
Department of Electrical and Computer Engineering

The problem addressed in this work is that of enhancing speech signals corrupted by additive noise and improving the performance of automatic speech recognizers in noisy conditions. The enhanced speech signals can also improve the intelligibility of speech in noisy conditions for human listeners with hearing impairment as well as for normal listeners.

The original Phase Opponency (PO) model, proposed to detect tones in noise, simulates the processing of the information in neural discharge times and exploits the frequency-dependent phase properties of the tuned filters in the auditory periphery along with the cross-auditory-nerve-fiber coincidence detection to extract temporal cues. The Modified Phase Opponency (MPO) proposed here alters the components of the PO model in such a way that the basic functionality of the PO model is maintained but the various properties of the model can be analyzed and modified independently of each other. This work presents a detailed mathematical formulation of the MPO model and the relation between the properties of the narrowband signal that needs to be detected and the properties of the MPO model.

The MPO speech enhancement scheme is based on the premise that speech signals are composed of a combination of narrow band signals (i.e. harmonics) with varying amplitudes.

The MPO enhancement scheme outperforms many of the other speech enhancement techniques when evaluated using different objective quality measures. Automatic speech recognition experiments show that replacing noisy speech signals by the corresponding MPO-enhanced speech signals leads to an improvement in the recognition accuracies at low SNRs. The amount of improvement varies with the type of the corrupting noise. Perceptual experiments indicate that: (a) there is little perceptual difference in the MPO-processed clean speech signals and the corresponding original clean signals and (b) the MPO-enhanced speech signals are preferred over the output of the other enhancement methods when the speech signals are corrupted by subway noise but the outputs of the other enhancement schemes are preferred when the speech signals are corrupted by car noise.

SYNERGY OF ACOUSTIC-PHONETICS AND AUDITORY MODELING TOWARDS ROBUST SPEECH RECOGNITION

by

Om D. Deshmukh

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee:
Dr. Carol Y. Espy-Wilson, Chair/Advisor
Dr. Shihab A. Shamma
Dr. Jonathan Z. Simon
Dr. Laurel H. Carney
Dr. P. S. Krishnaprasad
Dr. William J. Idsardi

DEDICATION

To the three ladies who have and will continue to shape my personal and professional life:

My Mother, My Guru[1] and My Wife.

[1] Guru (Sanskrit): The syllable ”gu” means “darkness” and ”ru” means “remover”. Guru is the person who dispels the darkness of ignorance by permeating the light of knowledge.

ACKNOWLEDGMENTS

I would like to express my immense gratitude to my thesis advisor and my Guru Dr. Carol Epsy-Wilson. Dr. Carol Espy-Wilson has been the inspiration for my research throughout my graduate student years. Her guidance and encouragement have helped me develop great enthusiasm and respect for research.

I would like to thank my other committee members professor Shihab Shamma, professor P.S. Krishnaprasad, professor Jonathan Simon, professor William Idsardi and professor Laurel Carney for their valuable suggestions on my thesis work. I am particularly thankful to professor Laurel Carney with whom we have had a fruitful collaboration over the past several years.

I would like to thank all my fellow graduate students and collegues especially Amit Juneja, Tarun Pruthi and Xinhui Zhou for both their friendship and for sharing their thoughts and ideas on technical matters.

I would like to acknowledge the timely help and support I received from the ECE, ISR and UMIACS IT staff, particularly from Peggy Jayant.

Sarah Friedman's help in the subjective evaluations is greatly appreciated. I would like to thank Ayanah George for her help in coding the spectral subtraction speech enhancement method, Esfandiar Zavarehei for making the source code of some of the other speech enhancement techniques publicly available, Dr. John Hansen and Dr. Bryan Pellom for making the source code for the objective quality

evaluations publicly available.

Finally, I would like to acknowledge the love and support of my family over the years: my parents Prabha and Dadaji Deshmukh, my brother Dr. Girish Deshmukh and my lovely wife Lavanya Deshmukh.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The problem addressed here is that of enhancing speech signals that are corrupted by different types of noise and of improving the performance of the Automatic Speech Recognition (ASR) systems when the input speech is degraded by noise.

Almost all applications of ASR systems are about interacting with humans (e.g., automatic weather updates, automatic flight status inquiry, automatic credit card inquiry, voice controlled navigation system, etc.). For such systems to become mainstream and to be used in day-to-day applications, they should be able to replicate the human speech perception performance not only in clean environments, but also in noisy environments.

There have been several studies to test the performance of human speech perception in background noise. It has been shown in [2] that humans can understand speech with less than 1% error both in quiet and at Signal to Noise Ratios (SNRs) as low as -3 dB. In a different study [3], it was found that the error rate of human speech perception on a digit recognition task was less than 1% both in quiet and at an SNR of 0 dB. The same study evaluated the performance of ASR systems with and without noise adaptation. The parameters of the HMM-based statistical back-end were modified to model the noise and the noisy speech signals. The lowest error rate for ASRs in quiet was about 2%, but the error rates increased to almost

100% without noise adaptation and to about 40% with the best noise adaptation algorithm. Performance of large vocabulary Hidden Markov Model (HMM) based continuous speech recognizers have been compared with that of humans. Speech was recorded using two high-quality microphones and with a low-quality omni-directional electret microphone. Human perception error rates were less than 1% irrespective of which microphone was used for the perceptual tests [4] . ASR error rates for the same task were about 8% when speech recorded from high-quality microphones was used for training and testing. But the error rate shot up to about 24% [5] when low-quality microphone speech was used for testing. This increase occurs despite extensive adaptation algorithms to compensate for channel variability introduced by different microphones. A detailed comparison of performance of human speech perception to that of ASR can be found in [6].

In a recent study [101, 106], the performance of human speech perception was compared with that of ASR systems when the speech signals were corrupted by speech-shaped noise at various SNRs. The speech signals consist of sentences which are simple sequences of the form:

$$< command : 4 >< color : 4 >< preposition : 4 >< letter : 25 >< number : 10 >< adverb : 4 >$$

The numbers in brackets indicate the number of choices at each point. The performance results for human perception and for the ASR system are tabulated in Table 1.1 and are also plotted in Fig. 1.1. The accuracies are computed only on the *color, letter* and *digit* keywords. As is evident from Fig. 1.1, the performance of ASR is

Table 1.1: Comparison of performance of human speech perception and ASR in speech-shaped noise conditions. Numbers indicate accuracies in percentage.

SNR	Clean	6 dB	0 dB	-6 dB	-12 dB
Human perception	99.4	98.3	95.0	79.3	37.5
ASR	99.0	54.3	18.0	11.4	12.8

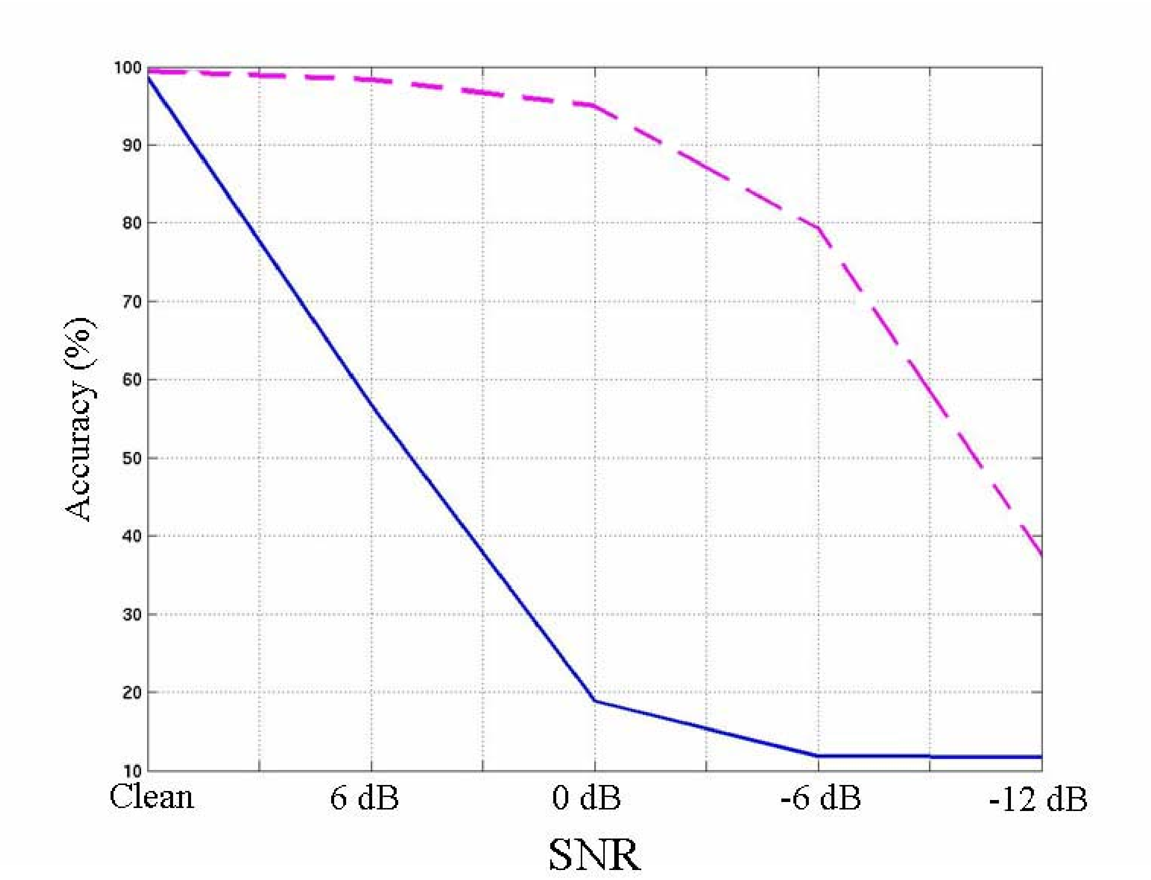

Figure 1.1: Comparison of performance of human speech perception and ASR in speech-shaped noise conditions. Pink (dashed-line): human perception; Blue (solid line): ASR.

very close to that of human perception in clean condition, but drops drastically as the SNR is reduced: At 6 dB SNR, the ASR accuracy is about 45% below that of human perception and at 0 dB SNR the ASR accuracy is about 80% below that of human perception.

The above studies show that there is a wide gap between human performance and machine performance, especially in degraded conditions. This difference in performance has fueled a variety of research to develop and implement algorithms for speech enhancement and robust speech recognition. In spite of the research, the gap in the human-machine performance hasn't been bridged. Moreover, systems that perform well in one kind of background noise typically fail to maintain the performance when tested on a different kind of disturbance.

There are several different sources that distort the speech signal. The two most important sources of distortion are: additive noise and linear filtering. Some of the everyday examples of additive noise sources are train noise, car noise, speech babble from background speakers, ambient air flow and noise emitted by surrounding machineries like fans and computers. Some of the sources of linear filtering are different configurations of vocal tracts of individual speakers and different microphones. Speech signals can also be corrupted by nonlinearities that occur due to room reverberation or due to changes in the telephone network. The problem is further compounded by the fact that speakers speaking in noisy environments make (statistically) significant changes in their articulation in an attempt to increase the communication efficiency [17, 18]. This phenomenon, referred to as the 'Lombard Effect', plays a significant role in the degradation of ASR systems when tested in ambient noise.

Various different approaches are being pursued to make the ASR systems robust in noise. The 'backend-intensive' approaches rely on the ability of a statistical backend (typically HMM) to form statistical models of different speech segments based on the set of training data. The basic premise for these methods is that the performance of a speech recognizer is optimal when there is little or no mismatch between the training and the testing conditions.These models can then be adapted to various background and speaking conditions to minimize the mismatches in training and testing environments. These methods typically are very data intensive and make minimal use of the insights gained into the functioning of the human speech production and speech perception apparatus.

A different class of approach is to develop speech features and distance measures that are invariant to distortions introduced by background noise and/or the channel characteristics. These methods typically make very little or no assumptions about the interfering noise. Many of these noise robust features find their origin in human speech perception studies.

A third approach, and the one we focus on in the proposed work, tries to enhance the speech signal by suppressing the noise as much as possible with very little distortion to the actual speech content. Many of the speech enhancement techniques were originally developed for speech quality improvement. But they can also be used as a pre-processing block for ASR systems. The enhanced speech may have a higher SNR, but the higher SNR does not necessarily translate into better quality or higher intelligibility as the improvement in the SNR could have been obtained at the expense of introducing distortions in the speech signals.

Speech enhancement techniques can also be used to improve the speech-in-noise intelligibility performance of human listeners with hearing impairment who typically have high SNR thresholds for speech reception.

A prominent class of speech enhancement and robust speech recognition techniques is based on using multiple microphone arrays. In this work, these methods are reviewed only marginally since the method proposed in this work for speech enhancement and robust speech recognition is a single-channel method.

The approach proposed in the present work can be used for both speech enhancement as well as for extraction of noise-robust parameters for speech recognition. The proposed model is based on a physiological model for detection of tones in the presence of additive noise , called the PO model, initially proposed in [1]. The model does not need an estimate of noise and makes minimal assumptions about the characteristics of the noise. The various components of the PO model are modified in such a way that the basic functionality of the PO model is maintained but the various properties of the model can be analyzed and modified independently of each other. A detailed mathematical formulation of the MPO model is developed. The relation between the properties of the narrowband signal that needs to be detected and the properties of the MPO model is also presented. The performance of the MPO speech enhancement scheme is evaluated using several different objective quality assessment measures and compared with some of the other speech enhancement techniques proposed in the literature. The subjective quality of the MPO-enhanced speech signals is also evaluated using six subjects with normal hearing. The MPO speech enhancement scheme is also used as a preprocessor for robust automatic speech recognition systems when the speech signals are corrupted by different noise types at various SNRs.

Chapter 2

Background

The various different approaches proposed for noise robustness can be classified into different broad categories based on the central premise on which the particular method is based. Methods based on speech enhancement try to extract the clean speech signal from the noisy signal either by estimating and subtracting the noise or by using the known properties of speech to predict the speech signal in a noisy observation. Some of the speech enhancement techniques are based on studies of the human auditory system while others are more signal-theoretic. Techniques based on computing noise-robust parameters rely on the ability of some of the discriminating features to maintain their discriminating abilities even in the presence of noise. Most of the noise-robust parameters are perceptually motivated. Statistical techniques for robust speech recognition usually estimate the statistical properties of corrupting noise to develop algorithms to counter the effect of noise. Techniques based on Computational Auditory Scene Analysis (CASA) develop speech separation methods based on principles of human hearing. In the following sections, some of the prominent methods in each of these categories are reviewed. A more thorough review can be found in [19].

2.1 The human auditory system

This section presents a brief overview of the various responses evoked in the peripheral and the central auditory system by different kinds of input sounds at different levels. Details can be found in [49, 50]. Some of the prominent computational models for mimicking these responses are discussed. Attempts to use some of these models for speech enhancement are also presented.

The human ear can be split into three different sections: outer, middle and inner. Outer ear, which is important in our ability to localize sounds, collects the sound pressure waves and transmits them to the middle ear. These pressure waves cause the tympanic membrane (or ear drum), a thin membrane that separates the outer ear from the middle ear, to vibrate. The middle ear consists of three ossicles that propagate the vibration to a opening in the inner ear- the oval window. The ossicles are designed to propagate the vibrational energy with minimum loss due to reflection. Hence they are also jointly referred to as the 'impedance matching device'. They are also used to attenuate sudden loud bursts in the incoming sound signal. The middle ear is usually modeled as a Band Pass Filter (BPF) with pass band between 1.5 kHz and 5.0 kHz.

The pressure waves produce mechanical movements in the Basilar Membrane (BM) in the inner ear. The location at which the maximum displacement occurs is dependent on the frequency content of the incoming signal and moves closer to the apex of the BM as the frequency of the signal reduces. Thus the BM can be thought of as a *linear* filterbank with each segment of the BM modeled as a bandpass filter with a certain center frequency and bandwidth. The bandwidth of these filters increases in proportion to the center frequency and the slope of the high frequency skirt of the filter is usually sharper on the high frequency side than on the low frequency side especially for CF $>$ 1500 Hz. The linear filterbank model for the BM is essentially a simplified model and does not adequately account for observed nonlinear phenomena (for example, the broadening of the response functions of the BM at high-amplitude levels). Some researchers have proposed a nonlinear model for the BM [53, 52].

The mechanical movement of the BM causes the inner ear fluid to flow which in turn bends small filaments called cilia. Cilia are attached to Inner Hair Cells (IHC). There are about 3500 IHCs in one cochlea in the human auditory system. Bending of the cilia results in flow of ionic currents through nonlinear channels into the IHCs. Thus the output of each IHC is a time-varying receptor potential. This stage can be modeled as a half-wave rectifier followed by a saturating nonlinearity. The form of the nonlinearity used is different in different models. Some of them have a static nonlinearity whereas some use an adaptive nonlinearity [67]. The ionic flow generates action potential across the hair cells. It is widely accepted that the potentials produced by the IHCs are proportional to the velocity of the BM vibration and not to the BM displacement itself. These potentials are transmitted to the central auditory system (cochlear nucleus) by the Auditory Nerve (AN) fibers as a train of impulses or spikes. Several different models simulating the interaction between the IHCs and the AN fibers have been proposed [68].

The AN fibers exhibit spontaneous firing rates (firing in absence of any external acoustic stimulation) that vary from close to 0 to about 100 spikes per second. About 60% of the fibers have high-spontaneous firing rates (>18 spikes/sec), 25% have medium-spontaneous firing rates (0.5-18 spikes/sec) and the rest have low-spontaneous firing rates (<0.5 spikes/sec). The sound intensity threshold at which a fiber starts responding depends on the frequency of the input signal. The frequency at which the threshold is the lowest is called the Characteristic Frequency (CF) of the fiber. As the intensity of the input signal increases the firing rate of the fiber increases and reaches a saturation firing rate. Any further increase in the input intensity level does not increase the firing rate. Several studies have shown that in response to a single, low level, pure tone there is a high level of activity in neurons with CFs close to the tone frequency, with activity dropping off for neurons with CFs on either side. However, at higher sound levels , due to neural saturation there is about uniform level of activity over a wide range of CFs around the tone frequency and the activity falls off at CFs far removed from the input tone frequency.

It is known that in response to low-frequency stimuli the fibers tend to fire at a particular phase of the stimulating waveform. A given fiber does not necessarily fire on every cycle of the stimulus, but when it does fire, it fires at the same phase of the waveform each time. This phenomenon is referred to as phase locking and can be thought of as a consequence of the transduction process: When the BM moves upwards, the IHCs are bent and a neural response is initiated. No response will occur when the BM moves downwards. Phase locking is not seen over the entire range of audible frequencies. The upper frequency limit for phase locking is known to be around 2 kHz. This lack of phase locking at higher frequencies is modeled as a Low Pass Filter (LPF).

Several different mechanisms for processing the speech-evoked physiological responses in the AN fibers to provide a coherent representation of the speech spectrum are presented in literature. Some of them are discussed below. In each of the methods discussed below, the model used for the auditory periphery is based on the modeling stages mentioned above.

2.1.1 Rate-place representation

Authors in [54] have presented a representation of the speech spectrum based on the rate-place profile of the population of AN fibers. They show that at a low Sound Pressure Level (SPL), the average firing rate of AN fibers closer to the formant frequencies is higher than that of the other AN fibers. Thus the peaks in the average rate profile of the population of AN fibers are good indicators of the formant frequencies of the input speech signal. As the input SPL is increased, high-spontaneous fibers corresponding to frequencies in the vicinity of the formants saturate and the rates of high-spontaneous fibers with CFs between the formant frequencies increase. The result is that the valleys in between the formants reduce and the peaks are less obvious. But the peaks are maintained in the profile of low and medium-spontaneous fibers. Thus appropriately weighted combinations of low, medium and high spontaneous rate fibers can form a rate-place representation which is robust over a wide range of SPLs.

2.1.2 Average localized synchronized rate

Authors in [54] have presented a representation based on the phase-locking of the responses of AN fibers. A period histogram, which is a plot of instantaneous discharge rate of a fiber close to a formant, shows that the rate is synchronized to the formant frequency. A quantitative measure of this synchronization, called *Average Localized Synchronized Rate (ALSR)*, can be estimated by computing the Fourier transform of the histogram. ALSR, at a given frequency ω_0, is defined as the average of the Fourier transform component of the histogram at frequencies within 0.25 octaves of ω_0. ALSR plots show clear peaks in the vicinity of formants and the peaks are robust over a wide range of SPLs. ALSR plots can be thought of as the temporal-place representation of the spectrum in the auditory nerve.

2.1.3 Generalized synchrony detection

In [60], the representation of the speech spectrum based on synchrony detection is presented. The model used for auditory periphery consists of an initial stage of linear filterbank followed by a nonlinear model of the prominent transformations from the BM vibration to the response of AN fibers. The nonlinear model consists of four subcomponents: a half-wave rectifier, a short-term adaptation component, a LPF and an Automatic-Gain-Control (AGC). The parameters of these components were adjusted to match relevant physiological data. The output of this stage is used as input for the Generalized Synchrony Detector (GSD). GSD is based on the ratio of the estimated magnitude of a sum waveform to the estimated magnitude of a difference waveform. The inputs to the sum and difference computation are the GSD input signal and a delayed version of the input signal. The delay is equivalent to the CF of the channel. This ratio is followed by a saturating nonlinearity to bound the output, especially when the input to the GSD is perfectly periodic with period equal to the delay. Thus the formant peaks in speech spectrum will result in high synchrony output in channels with CFs close to the formant frequency. Speech recognition results in noise using the output of GSD are presented in Section 2.3.7.

2.1.4 Ensemble interval histogram

A spectral representation of input speech signal based on the ensemble histogram of interspike intervals generated by a simulated array of AN fibers is presented in [75]. The BM is modeled as a linear filterbank. The ensemble of nerve fibers innervating a single IHC is simulated with an array of level-crossing detectors at the output of each cochlear filter (i.e. each level-crossing detector is equivalent to a fiber of specific threshold). The value assigned to each level is a random Gaussian variable with mean values uniformly distributed on the log scale over the dynamic range of the speech sounds and variance proportional to the mean value. For each level an inverse-interval histogram is computed using 100 linearly spaced bins covering the entire frequency range. An interval is defined as the time between two adjacent positive-going level crossings. To measure the extent of coherent neural activity across the fiber array, the individual histograms are collected into one ensemble histogram by summing corresponding bins for all fibers. The resulting frequency representation is the Ensemble Interval Histogram (EIH). The extent of coherent neural activity for a given frequency region is proportional to the magnitude of the corresponding bin in the EIH spectrum. Performance of EIH-spectrum-based parameters in speech recognition in noise is discussed in Section 2.3.6.

2.1.5 Cross-channel correlation technique

Authors in [53] have presented a representation based on the cross-correlation of simulated temporal activity of AN fibers in adjacent frequency channels. The temporal activity of AN fibers in response to speech signals is simulated using a composite model of the auditory periphery. Two different models of BM are included in the composite model of the auditory periphery. In the linear model, both the damping coefficients and the stiffness coefficients of the BM are modeled as exponential functions of the CF. In the nonlinear model, the damping terms are modeled as explicit functions of the displacement of the BM partition. The BM stage is followed by the IHC stage which is modeled as a memoryless compressive nonlinearity followed by a LPF. The next stage represents the interface between the IHCs and the auditory nerve and is modeled using the Oono-Sujaku reservoir model. Output of this stage, which demonstrates AGC properties is a representation of discharge activity of single auditory-nerve fiber. It is shown that if the nonlinear model of BM is used, the output of the composite model is orderly and systematic and doesn't change much even when the input SNR is as low as 0 dB. Channels between the first formant (F1) and the second formant (F2) appear to be synchronized to a waveform with a common periodicity. Channels above F2 appear to be synchronized to a waveform with a common periodicity but different from that driving the channels between F1 and F2. (This phenomenon is called synchrony capture.) Also the transition from channels synchronized to F1 to those synchronized to F2 is very sharp. A cross-correlation of the output of adjacent channels will be high when the channels are between two formants, but when there is a formant between the two channels the cross-correlation will be low. This information can be used for formant extraction.

2.1.6 Lateral Inhibition Network

The Lateral Inhibition Network (LIN) presented in [55, 56, 57] uses a composite model of the auditory periphery followed by Neural Networks (NN) to develop robust spatial-temporal representation of speech sounds. The composite model consists of three stages: analysis, transduction and reduction. The BM is modeled as a linear filterbank where the filters are related by a simple dilation and the BM response is a wavelet transform of the input sound signal. This constitutes the analysis stage. The transduction stage is modeled using a three-step process: a temporal derivative is used to convert instantaneous membrane displacement into velocity, the nonlinear channel through the hair cell is modeled by a sigmoid-like function and the leakage of the cell membrane is accounted for by a LPF. The first set of NNs operate on the output of composite model of the auditory periphery. The output of the first set of NNs at each frequency channel is computed by running a cross-channel-subtraction operation on the composite model's output. The second set of NNs operates on the output of the first set of NNs in such a way that a large peak in the input pattern dominates the output activity in its neighborhood. These peaks are usually the high-frequency harmonics and one or two resolved harmonics near F1. The LIN representation of speech spectrum is shown [57] to be robust to noise distortions. LIN representation of speech spectrum degrades at a lower pace than the linear power spectrum as the SNR is reduced. A mechanism for reconstructing the acoustic signal from its LIN output is presented in [56]. Such a reconstructed speech signal is shown to exhibit noise suppression.

A simplified version of LIN processing was used in [58] for speech enhancement. It was shown that for speech degraded by heavy noise, the improvement in SNR is as high as 12 dB and that the algorithm works better for vowels than for consonants. Authors in [59] also used LIN processing for speech enhancement. In their framework, LIN was used only in sections that were judged to be periodic. The quality of enhanced speech was assessed using the Itakura-Saito measure and showed consistent improvement over a wide range of input SNRs.

The cross-channel-correlation technique proposed in [53] (described in Section 2.1.5) determines the locations of spectral peaks in a manner very similar to that of LIN. It is noted in [51] that the spectral edges detected by these two techniques could shift a little away from the tonotopic location of the spectral peaks as the input SPL increases.

2.1.7 Auditory model for spectral shape analysis

Authors in [77] have presented an auditory model of spectral shape analysis in the central auditory system. The cortical stage of the model implements a two-dimensional wavelet transform on the auditory spectrogram. Each two-dimensional filter is tuned to a different spectral as well as temporal modulation pattern. The temporal modulation patterns are referred to as the *rates* and the spectral modulation patterns are referred to as the *scales*. An iterative method to reconstruct the speech signal from the auditory representation is also presented. It is shown that this model is able to differentiate additive noise from the speech signal even when the two have spectrally overlapping characteristics, as long as the modulation patterns of the noise are different from that of the speech signal.

In [78], this model is used for speech enhancement. The spectro-temporal modulation patterns of noise are estimated from noise-only regions and the relative weights for every frequency, rate and scale at each time instant are computed using generalized Weiner filter. The objective perceptual evaluation of speech quality of the enhanced speech shows improvement over a minimum-statistics-based enhancement scheme [109].

2.1.8 Phase Opponency

A model for detection of tone-in-noise based on processing the information in neural discharge times is presented in [1]. This model exploits the frequency-dependent phase properties of the tuned filters in the auditory periphery and uses cross-AN-fiber coincidence detection to extract temporal cues. It is shown that responses of some of the cross-channel coincidence detectors are reduced when a tone is added to a noise. This reduction in response in the presence of the target is referred as Phase Opponency (PO).

In the present study, we use a modified version of the PO model for speech enhancement and robust speech recognition. The PO model and the proposed modifications are discussed in detail in section 3.

2.2 Speech enhancement based on signal theoretic approaches

In this section we provide an overview of some of the signal theoretic approaches used to enhance speech corrupted by additive noise or linear filtering .

Most of the practical situations of speech distortion can be modeled using the following equation:

$$y[l] = h[l] * x[l] + n[l] \tag{2.1}$$

where $y[]$ is the observed signal, $h[]$ is the model of the linear filtering (i.e. convolutive distortion), $x[]$ is the clean speech signal and $n[]$ is the additive noise.

In the absence of any information about $y[]$, $h[]$, $x[]$ and $n[]$, it is impossible to recover the clean speech. Different systems make different assumptions about one or more of the above components which lead to different speech enhancement algorithms.

One of the initial systems developed for processing speech in background noise relied on representing speech as the response of the vocal tract to a pulse-train excitation for voiced sounds and a noise-like excitation for unvoiced sounds. The vocal tract itself is represented as a quasi-stationary all-pole system. A commonly used transfer function for an all-pole model of the vocal tract is of the form:

$$V(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

When there is no background noise, the optimal values of a_k can be computed by solving a set of linear equations [27]. The set of equations that need to be solved for a_k when the effect of additive noise is considered is a nonlinear set [28]. This is generally computationally undesirable. Authors in [28] proposed a sub-optimal but computational tractable method as an alternative to solving a set of nonlinear equations when the additive noise is modeled as a Gaussian with zero mean. Instead of computing $p(a|y)$ (which leads to solving a set of nonlinear equations), their method begins with some initial assumed values of a_0 for the coefficient vector a. Based on a_0, clean speech x_0 is estimated by maximizing $p(x_0|a_0, y_0)$ where y_0 is the observed vector. A first estimate of a, $\hat{a_1}$, is then formed based on x_0. This procedure is iterated till a stopping criterion is reached. The estimates of a and x vectors are obtained by solving a set of linear equations. It is shown that non causal Wiener filtering is a limiting case of this method. Results show that the poles obtained using this method on noise-corrupted speech are very close to that of clean speech. At low SNRs the primary perceptual effect is generation of musical noise. A similar algorithm is presented in [29] where the authors have used HMMs with mixtures of Gaussian AutoRegressive (AR) output probability to model the clean speech and the additive noise. The Markovian assumption leads to dependence of the estimates on adjacent frames.

A Wiener filter is the least square error-optimal linear filter used to estimate clean speech which is corrupted by additive noise. The frequency response of such a filter is given by:

$$\begin{aligned} H(\omega) &= \frac{\Phi_x(\omega)}{\Phi_y(\omega)} \\ &\approx \frac{P_y(\omega) - P_n(\omega)}{P_y(\omega)} \end{aligned}$$

where $\Phi(\omega)$ is the power spectral density. But since the power spectral densities are rarely known before hand the filter is approximated by using the short time power spectra $P(\omega)$. Thus, the design of this filter requires that the signal and the noise be stationary and that their statistics be known *a priori*.

The assumptions of the Wiener filter rarely hold in practical scenarios. This leads to the Least Mean Square (LMS) adaptive noise cancellation techniques. In these techniques it is assumed that a reference noise signal, $n(l)$, highly correlated to the actual additive noise corrupting the speech signal, but uncorrelated with the speech signal, can be used as an input to the adaptive filter. This reference noise signal is filtered through the adaptive filter and the output is the estimate of the corrupting noise $\hat{n}(l)$. This noise estimate is then subtracted from the noisy signal, $y(l)$, to get an estimate of clean speech, $\hat{x}(l)$. The clean speech estimate is in turn used to control the parameters of the adaptive filter. The parameters of the adaptive filter are such that the mean square error, $E((x(l) - \hat{x}(l))^2)$, is minimized. The problem is that the reference noise signal, $n(l)$, is not always available. Thus, an alternative method is proposed in [30]. In this method, a reference signal of the original speech is formed instead. Speech is known to be quasi-periodic. If the pitch period of the speech signal is found to be T then $x(l)$ and $x(l-T)$ will be highly correlated but $n(l)$ and $x(l-T)$ will not be. Estimates of speech are computed using an adaptive filter as follows:

$$\hat{x}(l) = \sum_{i=0}^{L} b_i . y(l-i-T)$$

Filter coefficients b_i are computed such that $E((x(l)-\hat{x}(l))^2)$ is minimum. This is the least mean square estimate of clean speech. The filtered speech demonstrates about 7 dB improvement in SNR at 0 dB SNR.

2.2.1 Spectral Subtraction

Spectral subtraction [31] is one of the simplest yet effective methods of speech enhancement when the speech signal is corrupted by additive noise. Spectral subtraction assumes that (a) the background noise remains stationary to the degree that its spectral magnitude (expected value) just prior to speech activity equals its expected value during speech activity and (b) speech and noise are uncorrelated and stationary stochastic processes. It is also assumed that removing the effect of the noise magnitude alone will result in substantial noise reduction. (The clean speech waveform is assumed to have the same phase as that of the noisy input signal.) The model assumed is similar to one in equation (2.1) with $h(n) \equiv \delta(n)$. This leads to:

$$\Phi_y(\omega) = \Phi_x(\omega) + \Phi_n(\omega)$$

where $\Phi(\omega)$ is the power spectral density. The above identity holds only approximately for short time spectral estimates obtained through DFT:

$$P_y(\omega) \approx P_x(\omega) + P_n(\omega)$$

$$P_x(\omega) \approx P_y(\omega) - P_n(\omega)$$

Where $P(\omega)$ is short time power spectrum. The magnitude of $P_n(\omega)$ is computed by taking its average over the non-speech region, $\bar{P}_n(\omega)$. The above equation can lead to negative power spectrum when the average noise power is more than the noisy signal spectrum. This problem can be resolved by using the following modified formula:

$$P_x(\omega) = max(P_y(\omega) - \bar{P}_n(\omega), P_0) \qquad P_0 \geq 0$$

The resulting speech magnitude estimate is subject to a few simple residual noise suppression techniques. A time waveform is calculated from the modified magnitude. This waveform is then overlap added to the previous data to obtain the enhanced speech. Enhanced speech doesn't increase the intelligibility, but it is shown to increase the quality of the speech. One of the main distortions introduced by this method is the musical noise. As the value of P_0 is increased, the musical noise is replaced by less conspicuous white noise. Spectral subtraction can be extended to generalized spectral subtraction by:

$$P_x(\omega) = |max(P_y^{\gamma}(\omega) - \alpha P_n^{\gamma}(\omega), P_0^{\gamma})|^{1/\gamma}$$

It is noted in [32] that significant overestimation of noise (i.e. $\gamma >> 1$) is advantageous. From the auditory perception viewpoint it is more appropriate to minimize the mismatch in the log spectral domain than in the power spectral domain. Modeling spectral subtraction in the log domain leads to complex and unwieldy derivations and the simplicity of spectral subtraction is sacrificed. Relative performances of power spectral subtraction, generalized spectral subtraction and nonlinear spectral subtraction are compared in [33]. Improved speech enhancement algorithms based on some form of spectral subtraction continue to be proposed even to date [34, 35, 36]

2.2.2 Soft Decision Noise Suppression Filter

Authors in [37] have proposed a two-state soft-decision maximum likelihood envelope estimator. This model takes into account the fact that the speech signal is not always present in the observed noisy signal. The two state model considers the probability of speech presence in each frame and can be represented as:

$$\begin{aligned} H_0: &\quad \text{speech absent:} \quad |y_l| = |n_l| \\ H_1: &\quad \text{speech present:} \quad |y_l| = |Ae^{j\theta} + n_l| \end{aligned}$$

This algorithm applies considerably more suppression when the measurement corresponds to low speech SNR. Since this case 'most likely' corresponds to noise alone, it is seen that the effect of residual noise should be reduced considerably. When the speech SNR is large, the measured SNR will be large and it is 'most likely' that speech is present. In this case, the original maximum likelihood algorithm is applied.

2.2.3 Ephraim-Malah MMSE-STSA estimator

Ephraim and Malah have proposed a Minimum Mean Square-Error Short-Time Spectral Amplitude estimator (MMSE-STSA) [38]. This model assumes that each of the Fourier expansion coefficients of the speech and noise process can be modeled as Gaussian random variables with zero mean. Moreover, it is also assumed that these coefficients are independent of each other. This assumption is not completely accurate but greatly simplifies the algorithm. The observed signal is given by:

$$y[t] = x[t] + n[t], \qquad 0 \leq t \leq T \tag{2.2}$$

Let $X_k = A_k exp(j\alpha_k)$, N_k and $Y_k = R_k exp(j\vartheta_k)$ denote the kth spectral component of the signal $x(t)$, the noise $n(t)$ and the noisy observation $y(t)$ respectively. The spectral components $Y_0, Y_1, \ldots$ bear the same information as that of $y(t)$ for every $t \in [0, T]$. Thus the MMSE estimation of A_k can be derived based on the observation set $Y_0, Y_1, \ldots$. Moreover, since the spectral components are assumed to be statistically independent, the MMSE estimator can be derived from Y_k alone. MMSE estimate, as is well known [39], is the conditional expectation and is given by:

$$\begin{aligned}
\hat{A}_k &= E[A_k|y(t)], \quad 0 \leq t \leq T \\
&= E[A_k|Y_0, Y_1, \cdots] \\
&= E[A_k|Y_k] \\
&= \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} exp(\frac{-v_k}{2}) \left[(1+v_k)I_0(\frac{v_k}{2}) + v_k I_1(\frac{v_k}{2})\right] R_k \qquad (2.3)
\end{aligned}$$

where $\Gamma(.)$ is the gamma function; $I_0(.)$ and $I_1(.)$ are the modified Bessel functions of zero and first order respectively; v_k is defined by:

$$\begin{aligned} v_k &= \frac{\xi_k}{1+\xi_k}\gamma_k \\ where, & \\ \xi_k &\triangleq \frac{\lambda_x(k)}{\lambda_n(k)} \qquad \text{apriori SNR} \\ \gamma_k &\triangleq \frac{R_k^2}{\lambda_n(k)} \qquad \text{a posteriori SNR} \end{aligned}$$

$\lambda_x(k) \triangleq E|X_k|^2$, and $\lambda_d(k) \triangleq E|D_k|^2$ are the variances of the kth spectral components of the speech and the noise respectively.

In practical situations, the *a priori* SNR, ξ_k, and the noise variance, $\lambda_n(k)$, are unknown and need to be estimated from the observed signal. The authors show that the estimator is more sensitive to underestimates of the *a priori* SNR than to its overestimates. A 'decision-directed' approach is used to estimate the *a priori* SNR. The model is extended later to take into consideration speech presence uncertainty. The quality of the enhanced speech is better using the MMSE estimator that takes into account the speech presence uncertainty than the one that does not. The power of the musical noise is low compared to that obtained by using spectral subtraction or Wiener filtering and the residual noise is perceived more as a colorless noise than as musical noise. The reduction in musical noise is attributed to the smooth variation of a priori SNR estimates [40]. The MMSE-STSA algorithm is extended in [41] to compute the STSA estimator that minimizes the mean-square error of the log-spectral amplitude which is a more relevant criterion for perceivable distortions in speech. Authors in [42], replaced the squared-error cost function by perceptually more relevant cost functions that take into account the auditory masking effects. Authors in [43] have presented a non-causal estimator for the a priori SNR which is capable of discriminating between speech onsets and noise irregularities.

2.2.4 Some other techniques for speech enhancement

There are several other techniques for speech enhancement : Some are based on using multitaper spectrum schemes in which a certain number of spectrum estimators are computed, each using a different taper (window), and then averaged across the population to compute the multitaper estimator. The estimate is then refined using wavelet thresholding [44]. Techniques based on Signal Subspace Approach (SSA) decompose the noisy observation signal into two orthogonal subspaces called the noise subspace and the signal subspace [45]. Attempts are also being made to incorporate some aspects of human auditory system in SSA [46]. Authors in [47, 48] have explored the use of super-Gaussian priors densities for the spectral components of speech signals.

2.3 Noise-robust parameters for robust speech recognition

In one class of countering the effect of noise, features that are inherently robust to noise are extracted. Most of the noise-robust parameters are motivated from the study of the human auditory system, although there are a few noise-robust parameters that are signal-theoretically motivated. One of the advantages of these techniques is that they generally make weak or no assumptions about the noise nor is any explicit estimation of noise statistics required. In this section we first review some of the noise-robust parameters based on signal theoretic approaches followed by a review of the ones based on the study of the human auditory system.

One of the initial methods of computing noise-robust parameters is based on cepstral mean subtraction [79]. In this method, the short-term average of the cepstral vectors of the input speech signal is subtracted from each cepstral vector. This method is known to compensate for the effect of unknown linear filtering.

2.3.1 Harmonic demodulation

It is shown [81] that additive noise affects frequencies with low energies more adversely than the frequencies with higher energies. Authors in [26] have presented a nonlinear envelope detection technique that is less susceptible to variations in energy valleys. In this method, speech production is viewed as a result of amplitude modulation in the frequency domain with the harmonic excitation as the carrier and the vocal tract transfer function as the modulating signal. The vocal tract transfer function can thus be obtained using demodulation techniques. The linear envelope detection technique for frequency-domain demodulation can be represented as:

$$Y(k) = X(k) * h(k) = \sum_{i} [X(i)h(k-i)]$$

where $X(k)$ is the discrete speech spectrum, and $h(k)$ is the discrete characteristic of a low-pass filter in the frequency domain. This linear envelope detection is susceptible to spectral valleys and any change in valleys will affect the resulting envelope. The nonlinear envelope detection technique focuses only on spectral peaks by replacing the summation in the above equation by a *max* operation:

$$Y(k) = max_i[X(i)h(k-i)]$$

To further reduce the mismatch in spectral valley regions between clean and noisy frames, the spectral regions that fall below a threshold after nonlinear envelope detection are set to that threshold. This threshold is empirically determined for each database.

2.3.2 Peak isolation

Authors in [82] have proposed noise-robust parameters based on raised-sine cepstral liftering followed by explicit peak normalization. The resulting parameters isolate local spectral peaks. Raised-sine cepstral liftering is equivalent to weighting the cepstral vector by the first half-period of a raised-sine function. A raised-sine lifter deemphasizes slow changes with frequency, often associated with overall level, as well as fast changes that may reflect numerical artifacts. The valleys are explicitly removed by half-wave rectification.

This method is extended in [26] where the peak-to-valley ratio is locked by normalizing the highest peak to a fixed value and scaling the rest of the cepstrum proportionately.

2.3.3 Phase autocorrelation

A class of noise-robust features called Phase AutoCorrelation (PAC) is presented in [83]. In PAC the angle between two vectors is used as a measure of correlation instead of the dot product. Consider two N dimensional speech frames that are spaced at an interval of k:

$$\begin{aligned}\boldsymbol{x_0} &= s_t[0], s_t[1], \ldots s_t[N-1] \\ \boldsymbol{x_k} &= s_t[k], s_t[k+1], \ldots, s_t[N-1], s_t[0], \ldots, s_t[k-1]\end{aligned}$$

The autocorrelation of these two vectors can be written as the dot product:

$$\boldsymbol{R}[k] = \boldsymbol{x_0^T x_k}$$

The magnitude of the two vectors $\mathbf{x_o}$ and $\mathbf{x_k}$ is the same since the individual components are the same. Let the magnitude be denoted by $||\mathbf{x}||$ and let θ_k denote the angle between the two vectors. The above autocorrelation equation can be rewritten as:

$$\begin{aligned}\boldsymbol{R}[k] &= ||\boldsymbol{x}||^2 cos(\theta_k) \\ \boldsymbol{P}[k] &= \theta_k = cos^{-1}(\frac{\boldsymbol{R}[k]}{||\boldsymbol{x}||^2})\end{aligned}$$

The new set of autocorrelation coefficients, $\mathbf{P}[\mathbf{k}]$, is referred to as the PAC. DFT performed on PAC will result in a PAC spectrum. From the PAC spectrum other features like the filter-banked PAC spectrum, PAC MFCCs can be computed.

Some of the noise-robust parameters based on a study of the human auditory system are reviewed below. One of the first efforts to incorporate perceptually motivated features was the inclusion of the Mel frequency scale [27] or the Bark scale which simulate the human ear's frequency resolution. It has been shown [84] that MFCCs are more robust to noise than LPCs.

2.3.4 Perceptual linear prediction

A technique for speech analysis based on the concepts of psychophysics of hearing is presented in [85]. It is known that for amplitude levels typically encountered in conversational speech, hearing is more sensitive in the middle frequency range of the audible spectrum. Consequently, spectral details extracted in linear prediction are not always in accordance with their auditory prominence. The method presented in [85] modifies the power spectrum prior to its approximation by the AR model. The first step in computing the Perceptual Linear Prediction (PLP) features is convolving the power spectrum with a simulated critical-band masking pattern. The resulting spectrum has significantly reduced spectral resolution as compared to the original power spectrum. The next step is to resample the resulting spectrum at equal Bark intervals. This is followed by emphasizing the bark-spectrum according to the sensitivity of the human ear to different frequencies. The net result is a slight increase in amplitude of frequencies between 3 to 5 kHz. The last operation before all-pole modeling is amplitude compression using a cube-root function. This approximates the power law of hearing. The resulting spectrum is then modeled as a AR process. Coefficients of the AR process are referred to as the PLP features. It is shown that PLP analysis is consistent with human hearing to changes in several important speech parameters like relative changes in frequencies and bandwidths of the formants, spectral tilt and fundamental frequency. PLP analysis is also consistent with the effective-second-formant theory [110] and the 3.5-Bark spectral integration theory of vowel perception [111].

2.3.5 RASTA Processing

A class of robust representations that exploit the differences between the temporal properties of speech and that of the environmental effects is presented in [86, 87]. Such representations are called Relative Spectra (RASTA). Human hearing is known to have greater sensitivity to modulation frequencies around 4 Hz than to lower or higher modulation frequencies. In RASTA processing, spectral estimate in each frequency channel is band-pass filtered by a filter with 'sharp' zeros at zero frequency and at 28.9 Hz and 50 Hz to emphasize the frequency region around 4 Hz. Prior to RASTA filtering the spectral amplitude is usually transformed through a compressing static nonlinearity and then transformed back after the RASTA filtering using an expanding nonlinearity. The type of nonlinearity depends on the type of distortion that is prominent. If the distortion is mainly convolutive then a logarithmic nonlinearity is used and the resulting processing is log-RASTA. If the distortion is both additive and convolutive then the nonlinearity is of the form $ln(1+Jx)$; where J is a signal-dependent positive constant. This processing is called lin-log RASTA or J-RASTA. Use of RASTA-PLP as speech parameters is shown to be robust to additive noise as well as linear convolutional distortions.

2.3.6 EIH parameters

EIH spectrum was presented in [75] and is summarized in section 2.1.4. In the earliest experiments using EIH as a noise-robust feature set, a Dynamic Time Warping (DTW) based speech recognizer was used as a back-end [76]. In this work, the EIH spectrum was modeled using the LPC technique and its performance was compared with power spectrum based LPC. It was shown that as the SNR reduces, the decline in performance using the EIH parameters was not as sharp as was found using power spectrum based parameters. A modified version of EIH spectrum is presented in [88] where the EIH spectrum of noise is determined apriori and all the EIH magnitudes below the noise floor are discarded from the final calculation of the EIH spectrum of speech. A different set of features that are motivated from EIH representation are the Zero Crossings with Peak Amplitudes (ZCPA) [90]. ZCPA are computed by passing a speech frame through a subband filterbank and finding all the positive-going zero crossings for each subband. For each pair of successive zero crossings the inverse interval length between the zero crossings is computed and a histogram of these inverse interval lengths is formed. The histogram count of each inverse interval length is weighted by the logarithm of the peak value of the signal between the two zero crossings. A detailed analysis of influence of different parameter choices on the ZCPA performance is presented in [89].

2.3.7 GSD parameters

GSD parameters were presented in [60] and are reviewed in section 2.1.3. Initial experiments with GSD parameters have highlighted their ability to estimate pitch frequency and to detect formant frequencies in clean speech [61, 62]. A modified form of GSD called the Average Localized Synchrony Detection (ALSD) is presented in [63]. Output of each ALSD is the average of several GSDs tuned to the same frequency but applied to several filters in the neighborhood of the filter corresponding to that frequency. ALSD based parameters are shown to give superior performance in detecting formants in noisy speech. In [74], authors have shown that combining the GSD model with normalized cepstral processing results in improved performance in noisy environments.

Perceptually motivated noise-robust parameters continue to be developed and used in robust speech recognition [64, 65]. One of the important things that is pointed out [66, 67] is that some of the perceptually motivated parameters perform better when used in conjunction with neural networks than with the traditional HMMs.

2.4 Statistical techniques for robust speech recognition

The statistical techniques are based on the objective of developing statistical models of clean or noisy speech and then adapting these models to accommodate for the noisy test environment. The statistical model used in current speech recognition systems is predominantly HMM with the output probability distribution of each state modeled as a mixture of Gaussians [7]. When the testing environment is not adequately represented in the training data, the model parameters can potentially be optimized to better represent the current environment and thus to obtain high recognition results. The best approach in terms of additive noise is to add the noise to the speech signal and train the models using this noisy speech. This approach, however, is not feasible in all cases especially when the database is large. Model-based compensation schemes assume that the model trained on clean speech contains sufficient information about the statistics of the clean speech signal and can thus be used for model compensation (instead of using the entire clean speech database) along with some of the noisy speech data.

In the following sections, X represents the clean speech signal, N represents the noise and Y represents the noisy speech signal. Different models or functions of these signals are represented with a corresponding subscript.

2.4.1 Stochastic Matching

The mismatch between the test speech utterance Y and the speech model trained on clean speech Λ_X can be reduced either in the feature domain by transforming the features of the test utterance to better match the features of the training data (i.e. $\hat{X} = \mathcal{F}_\nu(Y)$) or in the model domain by transforming the model to better match the estimated distribution of the features of the observed signal (i.e. $\Lambda_Y = \mathcal{G}_\eta(\Lambda_X)$). The unknown parameters ν and η can be estimated recursively to maximize the likelihood of the observed speech Y given the model Λ_X. In [15, 16], a Expectation-Maximization (EM) algorithm was formulated based on Maximum Likelihood (ML) approach for computing the parameters ν and η when the function $\mathcal{F}_\nu$ and $\mathcal{G}_\eta$ are assumed to be $Y_t - b_t$ where Y_t is the cepstrum of the observed noisy speech signal and b_t is the additive cepstral bias at time t. Thus, it is assumed that the distortion undergone by the speech signal can be modeled as linear filtering. Here the only parameter that needs to be estimated is the bias

b_t. If b_t is assumed to be unknown but non-random, then it is more appropriate to modify the features to compensate for the noise. If b_t is modeled stochastically, then it is more convenient to modify the speech model. The cases considered here are when the bias b_t is unknown but state-dependent and unknown but fixed for the given test utterance. When the bias is assumed to be random it is modeled as a single Gaussian density with diagonal covariance matrix. In a feature compensation technique, the bias is initialized to zero. In the case of model compensation, the bias mean is initialized to zero and the variance is initialized to a small positive number. The input string is then recognized based on the initial estimate of the bias. The bias is then re-estimated conditioned on this recognized string using the two-step EM iterative procedure. Thus, the performance depends heavily on the initial hypothesis. The performance of this algorithm was evaluated on recordings of 300 utterances of ARPA 91 RM database [20] spoken by two non-native male speakers using a close talking microphone and a telephone handset. The Word Error Rate (WER) was reduced from 14.1% with no compensation to 4.6% and 4.1% when feature compensation and model compensation were used respectively.

2.4.2 Parallel Model Compensation (PMC)

This technique assumes that the noise can be modeled using standard HMMs with Gaussian output probability distribution. The parameters of the corrupted speech model are estimated from the model of clean speech and the model of noise. No speech from the new acoustic environment is used. A HMM for the background noise is generated using some of the samples from the test data. The next step is to find a method of combining the parameters of the clean speech model and the noise model to estimate the noisy-speech models. In its simplest form, PMC assumes that each speech and noise state pairing can be modeled by a single Gaussian component. This approximation is very crude but greatly simplifies the formulation. The likelihood of the corrupted-speech observation being produced by the corrupted-speech model is given by:

$$\mathcal{L}(O^c(\tau)|q_j(\tau), q_v^n(\tau), \mathcal{M}_x, \mathcal{M}_n) \approx \mathcal{N}(O^c(\tau); \hat{\mu}^c, \hat{\sigma}^c)$$

where $\mathcal{N}()$ is the Gaussian pdf, $q_j(\tau)$ and $q_v^n(\tau)$ denote occupation of speech state j and noise state v at time τ, $\mathcal{M}_x$ is the clean-speech model and $\mathcal{M}_n$ is the noise model. A compensation scheme is now required to estimate the new means, $\hat{\mu}^c$, and the new variance, $\hat{\sigma}^c$, based on the clean speech model and noise model. The function used to capture the effect of noise on the speech parameters is called the 'mismatch function'. For additive noise, the mismatch function, F, for static parameters is given by:

$$\begin{aligned} O(\tau) &= \mathcal{F}(X((\tau), N(\tau)) \\ &= log(g.exp(X((\tau)) + exp(N(\tau)))) \\ \hat{\mu}_i &= \varepsilon log(g.exp(X((\tau)) + exp(N(\tau)))) \end{aligned} \tag{2.4}$$

where $X(\tau)$ and $N(\tau)$ are the log spectra of the speech signal and noise at time τ respectively. This equation has no simple closed-form solution and various approximations like log-normal approximation, log-add approximation, numerical integration have been tried. Details can be found in [8] .

2.4.3 Vector Taylor Series

Vector Taylor Series (VTS) is very similar to the PMC technique. The main difference between VTS and PMC is that VTS approximates the mismatch function by a finite length Taylor series and the statistics over this truncated Taylor series approximate the statistics of the corrupted-speech parameters. The computational cost of using VTS depends on the number of terms in the truncated Taylor series and as the number of terms is increased the approximation becomes more and more accurate. Details of VTS can be found in [11]. The mismatch function used for VTS is given by:

$$y \equiv f(n, x) = x + log(1 + exp(n - x))$$

where y, x and n represent the log-spectrum of noisy speech, clean speech and noise respectively. VTS was tested on the 1993 WSJ0 database with white noise added at various SNRs. At 25dB SNR, WER is about 15% using first order VTS. A further extension to VTS is proposed in [12] where the i^{th} order Taylor series is approximated by a minimum mean square error first order polynomial. The noise estimates can be refined using either sequential estimation with constant forgetting [13] or using sequential estimation with optimal filtering [14].

2.4.4 Iterative PMC

The previous two methods assume that the corrupted-speech distribution can be modeled using a single Gaussian component. But this is a rather crude approximation. In Iterative PMC (IPMC), several Gaussian mixtures are used to model the corrupted speech. The likelihood of an observation is then given by:

$$\mathcal{L}(O^c(\tau)|q_j(\tau), q_v^n(\tau), \mathcal{M}_x, \mathcal{M}_n) \approx \sum_i \hat{\omega}_i \mathcal{N}(O^c(\tau); \hat{\mu}^{(i)c}, \hat{\sigma}^{(i)c})$$

The next step is to estimate the statistics of the parameters of corrupted speech. Generally, in such cases a method called Data-driven PMC (DPMC) is used. In DPMC speech and noise observations are generated from their respective models and then combined using the chosen mismatch function to obtain corrupted-speech observations. Statistics of corrupted-speech are approximated as the statistics of these observations of corrupted-speech [10]. For IPMC, the system was tested on the Resource Management database. When clean speech was used for training and testing the WER is 4.6%. The performance of the system without any adaptation when the test data is corrupted with helicopter additive noise at 18dB SNR is 34.7%. Using the IPMC method the WER goes down to 7.6%. At 10dB SNR the lowest WER obtained is about 15.6% [9]. The WER increases gradually as the SNR goes down and is about 25% at 10 dB SNR [11]. PMC was tested on NOISEX92 digits database. In clean environment there were no errors but using additive noise the WER goes up to 83% at 0 dB SNR with no adaptation. Using PMC the WER goes down to 2%.

2.4.5 SPLICE

Authors in [21] have proposed an algorithm for noise reduction in the cepstral domain. The algorithm is called Stereo-based Piecewise LInear Compensation for Environment (SPLICE). As the name implies, the algorithm needs stereo clean/noisy speech data. The cepstral vector, $\mathbf{y}$, of noisy speech is modeled by a mixture of Gaussians, and the aposteriori probability of clean speech vector $\mathbf{x}$ given the noisy speech $\mathbf{y}$ and given the mixture component k is modeled using an additive correction vector $\mathbf{r_k}$.

$$p(\boldsymbol{x}|\boldsymbol{y}, k) = N(\boldsymbol{x}; \boldsymbol{y} + \boldsymbol{r_k}, \boldsymbol{\Gamma_k})$$

Thus a fundamental assumption made in the SPLICE algorithm is that the conditional mean of the a posteriori probability $p(\boldsymbol{x}|\boldsymbol{y})$ is a shifted version of the noisy data $\boldsymbol{y}$. The correction vectors, $\mathbf{r_k}$, are trained using the stereo data based on the ML principle:

$$\boldsymbol{r_k} = \frac{\sum_{t=0}^{T-1} p(k|\boldsymbol{y}_t)(\boldsymbol{x}_t - \boldsymbol{y}_t)}{\sum_{t=0}^{T-1} p(k|\boldsymbol{y}_t)}$$

Given a test vector, $\mathbf{y}$, the optimal noise-reduced speech vector is found using the MAP principle. A version of SPLICE based on MMSE decision is proposed in [22]. It is shown that HMMs trained on clean speech processed through SPLICE and tested on noisy speech processed through SPLICE performs better than HMMs trained on clean speech.

2.4.6 Estimating non-stationary additive noise

Authors in [23] have proposed a recursive algorithm for estimating additive noise in the cepstral domain. The model for additive distortion can be represented as:

$$\begin{aligned} y[l] &= x[l] + n[l] \\ |Y[k]|^2 &= |X[k]|^2 + |N[k]|^2 \\ \boldsymbol{y} &= \boldsymbol{x} + \boldsymbol{g}(\boldsymbol{n} - \boldsymbol{x}) \end{aligned}$$

where $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$ are the cepstral vectors of distorted speech, clean speech and additive noise respectively and $\mathbf{g}(\mathbf{n} - \mathbf{x})$ is given by:

$$\boldsymbol{g}(\boldsymbol{n} - \boldsymbol{x}) = \boldsymbol{C} ln[\boldsymbol{I} + exp[\boldsymbol{C}^{\boldsymbol{T}}(\boldsymbol{n} - \boldsymbol{x})]]$$

where $\mathbf{C}$ is the discrete cosine transformation matrix. $\mathbf{g}$ is thus a nonlinear function of $\mathbf{n}$ and $\mathbf{x}$. A linear approximation is made by truncating the Taylor series expansion of the nonlinearity, around a frequently updated operating point, up to the linear term. It is assumed that the noise cepstrum is unknown but non-random and is time varying. The noise cepstrum is estimated for every time frame t using recursive-EM algorithm. The updated noise estimate becomes the new Taylor series expansion point. This method, when used in conjunction with a noise-normalized version of a front-end denoising algorithm, SPLICE (sec 2.4.5), results in relative WER reduction of 27.9% on the Aurora database.

2.4.7 Phase-sensitive model of the acoustic environment

Consider the following model of the acoustic environment:

$$y[l] = h[l] * x[l] + n[l] \tag{2.5}$$

If it is assumed that $h[l] = \delta[l]$, (i.e. there is no distortion due to linear filtering) then the only distortion is because of the additive noise. The power spectrum of the noisy speech can then be obtained as:

$$\begin{aligned} |Y[k]|^2 &= |X[k] + N[k]|^2 \\ &= |X[k]|^2 + |N[k]|^2 + 2|X[k]||N[k]|cos\theta_k \end{aligned}$$

In most cases, it is assumed that the last term in the above equation is zero. A model that assumes non-zero value for the last term in the above equation is presented in [24]. In terms of the log spectra the above equation can be expressed as:

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{y} + log[1 + e^{\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h}} + 2\boldsymbol{\alpha} \bullet e^{(\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h})/2}]$$

where α is the phase information. It is assumed that the phase factor can be modeled as a zero-mean Gaussian with diagonal covariance matrix. The log spectral vector of the noise $\mathbf{n}$ is assumed to be non-stationary, unknown but non-random and is estimated using the method mentioned in section (2.4.6). The clean speech log spectral vector $\mathbf{x}$ is estimated using a MMSE estimator. The phase-sensitive estimator results in about 6% reduction in error rate when tested on the Aurora database. This model is extended in [25] by proposing a prior speech model based on static and dynamic features.

A technique that utilizes relative phase information of clean speech and additive noise to compute robust parameters for speech recognition is presented in [26].

2.5 Missing Data Techniques for robust speech recognition

When a speech signal is affected by interfering noise it is expected that some spatio-temporal regions will be more affected than others. The approach pursued by missing data techniques is to compute a time-frequency reliability mask to identify regions that retain reliable speech information. These masks can either be binary with a value of unity representing reliable region and a value of zero representing unreliable region [69] or real valued [70]. These masks can be computed using computational models of Auditory Scene Analysis (ASA) [71] or using statistical approaches. A binaural processing model for missing data robust speech recognition is presented in [69]. This model utilizes interaural time difference and interaural level difference cues along with precedence effect [71] to compute a binary reliability mask. Once the mask is computed the unreliable components can be dealt with in several different ways. They can be estimated based on the values of the reliable components and the covariance structure of each recognition category. This method is called data imputation [72]. An alternative approach is to integrate over the unreliable components and classify based solely on the reliable components. This method is called marginalization[72].

The binaural model in [69] uses bounded marginalization and is shown to give substantial improvements in recognition accuracy as the spatial separation of speech and noise sources increase from 10° to 40°. The performance is consistently better than that of a baseline MFCC system as the T_{60} reverberation time is increased from 0 (i.e. anechoic room) to 0.3 (mildly reverberant room) to 0.45 ('live' offices) and as SNRs as low as 0 dB.

Authors in [73] have proposed a statistical approach for robust speech recognition in missing data scenarios. Their model, probabilistic union model, does not require the identity of the corrupted bands. Features are combined based on probability theory for union of random events. The basic idea of the model is that in a recognition system with N subbands if M subbands are corrupted by noise then there exists one subset of $(N - M)$ features which represent clean speech. The performance of the system depends on the estimate of M.

Chapter 3

Modified Phase Opponency

The original PO model proposed in [1] is a physiological model for the detection of tones in the presence of additive noise and relies on the temporal information contained in the discharge patterns of auditory neurons. The PO model detects the presence of narrowband signals by cross-correlating outputs of two gammatone filters (GTFs) of equal bandwidths but with slightly different Center Frequencies (CFs). The GTFs are chosen such that there is a frequency region in the common passbands of the two filters where the phase responses of the two filters are out-of-phase. This frequency region is referred to as the *out-of-phase* region. The rest of the frequency region is referred to as the *in-phase* region. Thus, cross-correlation of the outputs of the two filters will lead to a negative value when a narrowband signal is present in the *out-of-phase* region and to a positive output when only wideband noise is present which covers the *out-of-phase* as well as the *in-phase* regions. Fig. 3.1 shows the magnitude and phase response of the two GTFs used to detect a tone at 900 Hz. Notice that the magnitude response of the two GTFs is about the same at 900 Hz, but their phase responses are exactly out-of-phase. As a result, the cross-correlation of the outputs of the GTFs will lead to a strongly negative value. On the other hand, if the input is a broadband noise, the outputs of the two GTFs will be partially correlated leading to a positive or a slightly negative output. The output of each GTF is subject to a hard-saturating non-linearity. The non-linearity minimizes the magnitude information in the filter outputs and thus the final outputs depend largely on the relative temporal information.

3.1 From Phase Opponency to Modified Phase Opponency

The transfer function of a typical GTF is given by:

$$G(\omega) = \frac{\tau^{\gamma}(\gamma-1)!}{2[1+j\tau(\omega-\omega_{CF})]^{\gamma}} \tag{3.1}$$

where τ is the time constant, ω_{CF} is the radian frequency corresponding to the CF and γ is the order of the filter. The phase response of the GTF in equation (3.1) is given by:

$$\Phi(\omega_{CF}) = -\gamma tan^{-1}[\tau(\omega-\omega_{CF})] \tag{3.2}$$

If the two GTFs used in a PO model have CFs at ω_1 and ω_2, then the difference in the phase response of the two filters is given by:

$$\begin{aligned}
\Delta\Phi &= \Phi(\omega_2) - \Phi(\omega_1) \\
&= -\gamma tan^{-1}[\tau(\omega-\omega_2)] + \gamma tan^{-1}[\tau(\omega-\omega_2)] \\
&= \gamma\left[tan^{-1}[\tau(\omega-\omega_1)] - tan^{-1}[\tau(\omega-\omega_1)]\right] \\
&= \gamma tan^{-1}\left[\frac{\tau(\omega-\omega_1)-\tau(\omega-\omega_2)}{1+\tau^2(\omega-\omega_1)(\omega-\omega_2)}\right] \\
&= \gamma tan^{-1}\left[\frac{(\omega_2-\omega_1)\tau}{1+\tau^2(\omega-\omega_1)(\omega-\omega_2)}\right] \qquad (3.3)
\end{aligned}$$

The frequency ω_0, where the phase difference is equal to $-\pi$, can now be computed by equating equation (3.3) to $-\pi$ and solving for $\omega = \omega_0$:

$$tan(\pi/\gamma) = \frac{(\omega_2-\omega_1)\gamma}{[1+\tau^2(\omega_0-\omega_1)(\omega_0-\omega_2)]} \tag{3.4}$$

Thus, for GTFs with fixed bandwidth and fixed order, the frequency location where the two GTFs have out-of-phase phase responses can be controlled by varying the CFs of the two GTFs. But as the CFs of the GTFs are varied, the magnitude response of the GTFs also vary making it difficult to manipulate the relative phase response or the relative magnitude response of the two filters independent of the other. Moreover, it is difficult to predict the relation between the parameters of the GTFs and the width and the location of the *out-of-phase* region.

Fig. 3.2 shows the schematic of the proposed MPO model. In the Modified Phase Opponency (MPO) model, a Band Pass Filter (BPF) replaces the GTF in one of the paths and the GTF in the other path is replaced by a combination of the same BPF and an All Pass Filter (APF). The relative phase response of the two paths can be manipulated by changing the parameters of the APF which does not introduce any changes in the relative magnitude response. The magnitude response of the two paths can be manipulated by changing the parameters of the BPF which does not introduce any changes in the relative phase response. Thus, the MPO model allows for manipulation of the relative magnitude response and the relative phase response independently of the other. The characteristics of the BPF are mainly decided by the range of the target frequency that is to be detected. The characteristics of the APF are mainly decided by the expected bandwidths of the target signal. The location and the width of the *out-of-phase* region can be controlled by varying the parameters of the APF.

In the next section we develop the mathematical basis of the MPO model.

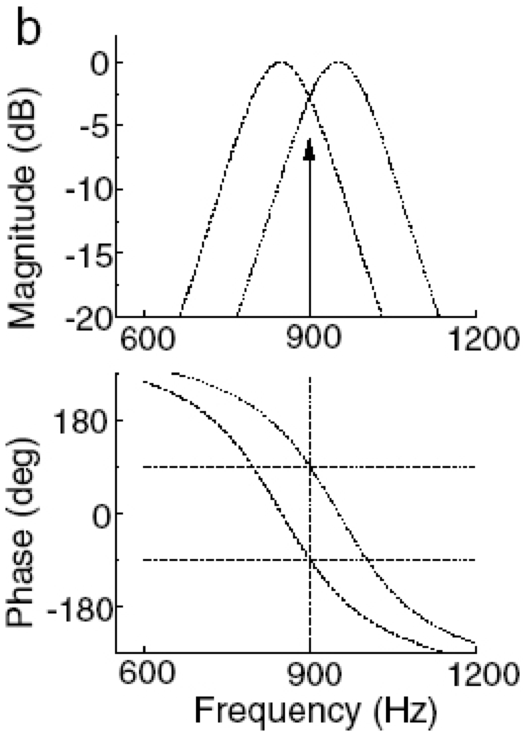

Figure 3.1: PO filter pair to detect a tone at 900 Hz

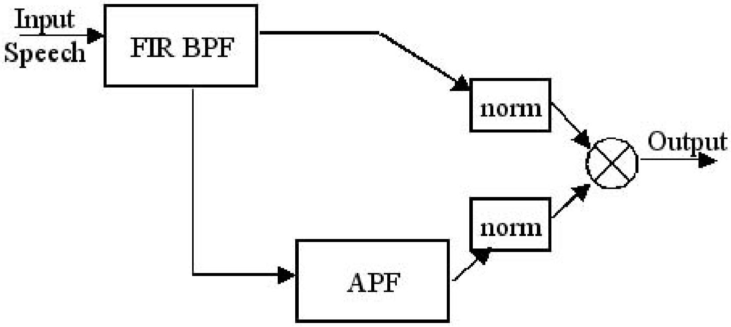

Figure 3.2: Modified PO filter pair

3.2 Mathematical formulation of the MPO model

Consider a pure tone of unit amplitude at frequency $\omega_0 : cos(\omega_0 n+\theta)$. Assume that ω_0 is in the passband of the filters in both the paths of the PO model. Also assume that the phase difference of the two filters at ω_0 is π, the response of first filter is $H_1(\omega_0)e^{j\phi}$ and that of the other filter is $H_2(\omega_0)e^{j(\phi+\pi)}$. The Fourier transform of the input signal can be written as:

$$cos(\omega_0 n + \theta) ==> \frac{1}{2}(\delta(\omega - \omega_0)e^{(j\theta)} + \delta(\omega + \omega_0)e^{(-j\theta)})$$

The frequency response of the filter in the first path is:

$$O_1(\omega) = \frac{1}{2}H_1(\omega_0)e^{j\phi}.(\delta(\omega - \omega_0)e^{(j\theta)} + \delta(\omega + \omega_0)e^{(-j\theta)})$$

giving the time-domain signal:

$$o_1(n) = A_1 cos(\omega_0 n + \theta + \phi)$$

The frequency response of the second path is:

$$O_2(\omega) = \frac{1}{2}H_2(\omega_0)e^{j\phi+\pi}.(\delta(\omega - \omega_0)e^{(j\theta)} + \delta(\omega + \omega_0)e^{(-j\theta)})$$

giving the time-domain signal:

$$o_2(n) = -A_2 cos(\omega_0 n + \theta + \phi)$$

Assume that the magnitude of the responses of both the filters at ω_0 is unity: $|H_1(\omega_0)| = A_1 = 1$ and $|H_2(\omega_0)| = A_2 = 1$. The correlation output then looks like:

$$o_1(n).o_2(n) = -cos^2(\omega_0 n + \theta + \phi)$$

Thus, if the phase difference between the two paths is π at a frequency corresponding to that of the input, then the output is negative (or zero). At the other end of the spectrum, if the phase difference is zero, the output is positive (or zero). This is the basic idea of the PO model.

The MPO model has the same BPF in both the parallel paths. One of the parallel paths has an APF that introduces phase differences in the outputs of the two paths. The parameters controlling the behavior of the BPF and the APF are governed by the expected frequency locations and bandwidths of the signals that need to be detected.

Consider an APF, $H(z)$, with one pair of complex conjugate poles.

$$H(z) = \frac{(z^{-1} - a^*)(z^{-1} - a)}{(1 - a^* z^{-1})(1 - a z^{-1})}$$

where $a = re^{j\theta}$ is the complex pole and a^* is its complex conjugate. Fig. 3.3 shows the magnitude and phase response of a typical APF with one pair of complex conjugate poles. The magnitude response is 1 for all values of ω and the phase response, $\Phi(\omega)$, is given by:

$$\begin{aligned}
\Phi(\omega) &= -\omega - 2tan^{-1}\left[\frac{rsin(\omega - \theta)}{1 - rcos(\omega - \theta)}\right] \\
&\quad -\omega - 2tan^{-1}\left[\frac{rsin(\omega + \theta)}{1 - rcos(\omega + \theta)}\right] \\
&= -2\omega - 2tan^{-1}\left[\frac{\frac{rsin(\omega - \theta)}{1 - rcos(\omega - \theta)} + \frac{rsin(\omega + \theta)}{1 - rcos(\omega + \theta)}}{1 - \frac{rsin(\omega - \theta)rsin(\omega + \theta}{(1 - rcos(\omega - \theta))(1 - rcos(\omega + \theta))}}\right] \\
&= -2\omega - 2tan^{-1}\left[\frac{2rsin(\omega)cos(\theta) - r^2 sin(2\omega)}{1 - 2rcos(\omega)cos(\theta) + r^2 cos(2\omega)}\right]
\end{aligned} \tag{3.5}$$

We are interested in deriving the relation between r and θ and the location and the width of the *out-of-phase* region. Notice from Fig. 3.3 that locating the *out-of-phase*

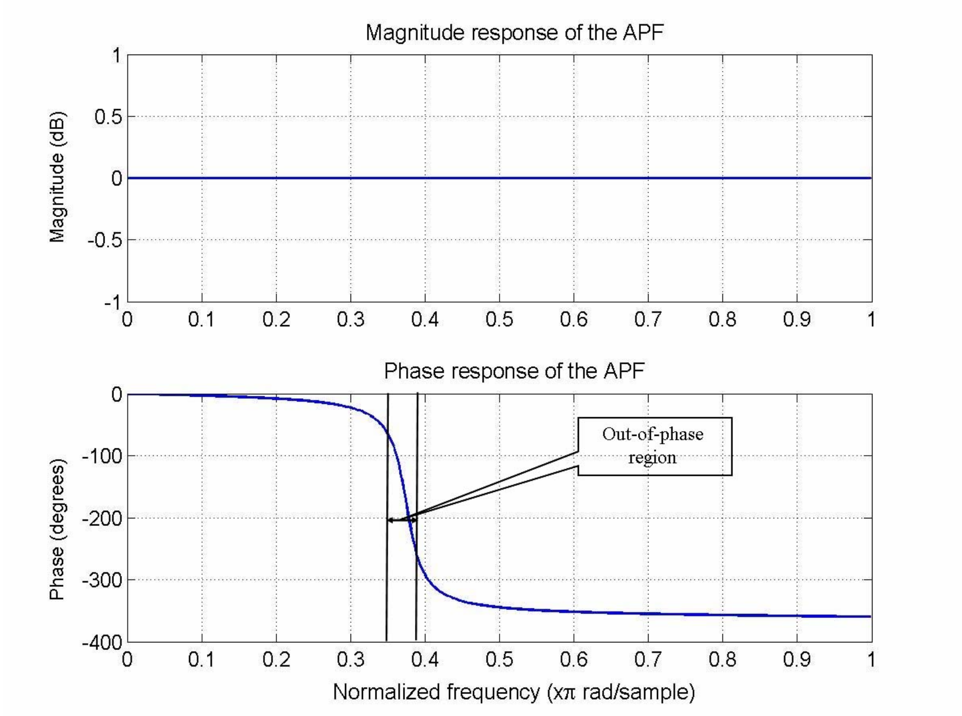

Figure 3.3: Magnitude and phase response of a typical all pass filter with one pair of complex conjugate poles

region is equivalent to locating the frequency region where the phase response is the steepest. This region can be located by finding the frequency where the slope of the phase response has an inflexion point, i.e. find ω for which $d^2(\Phi(\omega))/d\omega^2 = 0$. For simplicity, let us first compute $d^2(\Phi(\omega))/d\omega^2$ for just one pole, a, and then account for the complex conjugate pole a^*. Taking the derivative w.r.t to ω on both sides of equation (3.5) (but with only the first two terms from the r.h.s. which correspond to the pole a), we have:

$$\begin{aligned}
\frac{d(\Phi(\omega))}{d\omega} &= -1-2\left[\frac{1}{1+\left[\frac{rsin(\omega-\theta)}{1-rcos(\omega-\theta)}\right]^2}\right. \\
&\quad \left.\frac{rcos(\omega-\theta)(1-rcos(\omega-\theta))-rsin(\omega-\theta)rsin(\omega-\theta)}{(1-rcos(\omega-\theta))^2}\right] \\
&= -1-2\left[\frac{rcos(\omega-\theta)-r^2cos^2(\omega-\theta)-r^2sin^2(\omega-\theta)}{1-2rcos(\omega-\theta)+r^2cos^2(\omega-\theta)+r^2sin^2(\omega-\theta)}\right] \\
&= -1-2\left[\frac{rcos(\omega-\theta)-r^2}{1-2rcos(\omega-\theta)+r^2}\right] \qquad (3.6)
\end{aligned}$$

The second order derivative is then given by:

$$\begin{aligned}
\frac{d^2(\Phi(\omega))}{d\omega^2} &= -2\left[\frac{-rsin(\omega-\theta)(1-2rcos(\omega-\theta)+r^2)}{(1-2rcos(\omega-\theta)+r^2)^2}\right] \\
&\quad -2\left[\frac{-(rcos(\omega-\theta)-r^2)(2rsin(\omega-\theta))}{(1-2rcos(\omega-\theta)+r^2)^2}\right] \\
&= -2\left[\frac{(r^3-r)sin(\omega-\theta)}{(1-2rcos(\omega-\theta)+r^2)^2}\right] \qquad (3.7)
\end{aligned}$$

If we factor in the effect of the complex conjugate pole in equation (3.7), we have:

$$\frac{d^2(\Phi(\omega))}{d\omega^2} = -2\left[\frac{(r^3-r)sin(\omega-\theta)}{(1-2rcos(\omega-\theta)+r^2)^2}+\frac{(r^3-r)sin(\omega+\theta)}{(1-2rcos(\omega+\theta)+r^2)^2}\right] \qquad (3.8)$$

Since we are only interested in finding the value of ω for which the above equation (3.8) becomes zero, we can conveniently ignore the denominator. The numerator can be written as:

$$\begin{aligned} N(\omega) &= -2(r^3-r)\Big[2sin\omega cos\theta - 8rsin\omega cos\omega + 4r^2sin\omega cos\theta \\ &\quad -8r^3sin\omega cos\omega + 2r^4sin\omega cos\theta \\ &\quad +4r^2\left[sin(\omega-\theta)cos^2(\omega+\theta)+sin(\omega+\theta)cos^2(\omega-\theta)\right]\Big] \end{aligned} \tag{3.9}$$

The term

$$4r^2\left[sin(\omega-\theta)cos^2(\omega+\theta)+sin(\omega+\theta)cos^2(\omega-\theta)\right]$$

in equation (3.9) can be simplified as follows:

$$\begin{aligned} &= 4r^2[sin(\omega-\theta)-sin(\omega-\theta)sin^2(\omega+\theta) \\ &\quad +sin(\omega+\theta)-sin(\omega+\theta)cos^2(\omega-\theta)] \\ &= 8r^2sin\omega cos\theta[1-sin(\omega-\theta)sin(\omega+\theta)] \\ &= 8r^2sin\omega cos\theta[cos^2\omega+sin^2\theta] \end{aligned} \tag{3.10}$$

The numerator in equation (3.8) can then be rewritten as:

$$\begin{aligned} N(\omega) &= -2(r^3-r)[(2+4r^2+2*r^4)sin\omega cos\theta - \\ &\quad -(8r+8r^3)sin\omega cos\omega + 8r^2(cos^2\omega+sin^2\theta)sin\omega cos\theta] \end{aligned}$$

Equating $N(\omega)$ to zero implies,

$$[1 + 2r^2 + 4r^2(cos^2\omega + sin^2\theta) + r^4]cos\theta = (4r + 4r^3)cos\omega$$

$$i.e. \quad \left[\frac{1 + 2r^2 + 4r^2(cos^2\omega + sin^2\theta) + r^4}{4r(1 + r^2)}\right] cos\theta = cos\omega \tag{3.11}$$

$$i.e. \quad D(r, \omega, \theta)cos\theta = cos\omega \tag{3.12}$$

where,

$$D(r, \omega, \theta) \quad = \quad \left[\frac{1 + 2r^2 + 4r^2(cos^2\omega + sin^2\theta) + r^4}{4r(1 + r^2)}\right]$$

Notice that the sum of the coefficients in the numerator of $D(r, \omega, \theta)$ in equation (3.11) $(1 + 2 + 4 + 1 \;==\; 8)$ is exactly equal to that of the coefficients in the denominator $(4 * (1 + 1) == 8)$. Also notice that the $cos\theta$ term on the l.h.s. is balanced by the $cos\omega$ term on the r.h.s. Thus, the equality in equation (3.11) holds for $\theta = \omega$ and $r = 1$. But stability of the APF dictates that the magnitude of r be less than 1. Table 3.1 shows that $D(r, \omega, \theta)$ is very close to one for various values of r less than 1. Thus it is reasonably accurate to assume that the slope of the phase response, $\Phi(\omega)$, of a stable APF with a pair of complex conjugate poles at $a = re^{j\theta}$ and a^* is steepest at $\omega = \theta$. The following theorem makes it clearer.

Theorem 3.2.1. *Consider a stable allpass filter with a pair of complex conjugate poles at $a = re^{j\theta}$ and a^*. The frequency ω, at which the slope of the phase response, $\Phi(\omega)$, is the steepest is given by $\omega = \theta$. Moreover, this frequency value is independent of r, the magnitude of the pole.*

It is worth evaluating the phase response of the APF at $\theta = \omega$. The phase

Table 3.1: Dependence of $D(r, \omega, \theta)$ on r

r	$D(r, \omega, \theta)$
0.750	1.0008
0.775	1.0005
0.800	1.0003
0.825	1.0002
0.850	1.0001
0.875	1.0000
0.900	1.0000
0.925	1.0000
0.950	1.0000
0.975	1.0000
1.000	1.0000

response, $\Phi(\omega)$ is given by:

$$\begin{aligned}
\Phi(\omega) &= -\omega - 2tan^{-1}\left[\frac{rsin(\omega-\theta)}{1-rcos(\omega-\theta)}\right] \\
&\quad -\omega - 2tan^{-1}\left[\frac{rsin(\omega+\theta)}{1-rcos(\omega+\theta)}\right]
\end{aligned}$$

$$\theta = \omega \implies$$

$$\begin{aligned}
\Phi(\omega) &= -2\theta - 2tan^{-1}\left[\frac{rsin(2\theta)}{1-rcos(2\theta)}\right] \\
&= -2\theta - 2tan^{-1}\left[\frac{2rsin\theta cos\theta}{1-r+2rsin^2\theta}\right]
\end{aligned}$$

If $r \approx 1$, then $1 - r \approx 0$ and the above equation is further simplified to:

$$\begin{aligned}
\Phi(\omega) &= -2\theta - 2tan^{-1}\left[\frac{2rsin\theta cos\theta}{2rsin^2\theta}\right] \\
&\approx -2\theta - 2tan^{-1}(cot\theta)
\end{aligned}$$

$$\Phi(\omega) \approx \begin{cases} -2\theta - 2[-\frac{1}{2}\pi - cot^{-1}(cot\theta)] & \text{if } cot\theta < 0 \\ -2\theta - 2[\frac{1}{2}\pi - cot^{-1}(cot\theta)] & \text{if } cot\theta > 0 \end{cases}$$

$$\Phi(\omega) \approx \begin{cases} \pi & \text{if } cot\theta < 0 \\ -\pi & \text{if } cot\theta > 0 \end{cases} \tag{3.13}$$

The phase response at $\theta = \omega$ can thus be approximated as $\pm\pi$. The closer the value of r to 1, the more accurate the approximation is. Table 3.2 shows the exact phase response at $\theta = \omega$ for values of r below 1. Also, note that the frequency where the phase response is exactly out-of-phase (i.e. $\Phi(\omega) = -\pi$) is only about 4 Hz away from the CF when the pole magnitude value is greater than or equal to 0.9 and is within 50 Hz of the CF for pole magnitude values as low as 0.75.

The next step is to express the slope of $\Phi(\omega)$ at $\omega = \theta$ in terms of r and θ.

Table 3.2: Change in the phase response at $\theta = \omega$ as the value of r is varied. The frequency location where the phase response is exactly out-of-phase (ω_{op} Hz) is tabulated in the third column. The CF is 1000 Hz corresponding to $\theta = 0.25 * \pi$.

r	$\Phi(\omega)\|_{\omega=\theta}$	ω_{op} Hz
0.750	$-0.910 * \pi$	1050.78
0.775	$-0.920 * \pi$	1042.96
0.800	$-0.930 * \pi$	1027.34
0.825	$-0.939 * \pi$	1019.53
0.850	$-0.948 * \pi$	1019.53
0.875	$-0.958 * \pi$	1011.71
0.900	$-0.967 * \pi$	1003.90
0.925	$-0.975 * \pi$	1003.90
0.950	$-0.984 * \pi$	1003.90
0.975	$-0.992 * \pi$	1003.90
1.000	$-1.000 * \pi$	1000.00

From equation (3.6), we know that the derivative of the $\Phi(\omega)$ w.r.t ω is given by:

$$\begin{aligned}
\frac{d(\Phi(\omega))}{d\omega} &= -1-2\left[\frac{rcos(\omega-\theta)-r^2}{1-2rcos(\omega-\theta)+r^2}\right] \\
&\quad -1-2\left[\frac{rcos(\omega+\theta)-r^2}{1-2rcos(\omega+\theta)+r^2}\right] \\
&= -2\left[1+r\left[\frac{cos(\omega-\theta)-r}{1-2rcos(\omega-\theta)+r^2}+\frac{cos(\omega+\theta)-r}{1-2rcos(\omega+\theta)+r^2}\right]\right] \\
&= -2\left[\frac{1+(2r^3-2r)cos\omega cos\theta-r^4}{1-(4r+4r^3)cos\omega cos\theta+2r^2+r^4+4r^2(cos^2\omega-sin^2\theta)}\right]
\end{aligned} \tag{3.14}$$

From equation (3.11), we can infer that the following equality holds for ω corresponding to the steepest slope.

$$(4r+4r^3)cos\omega cos\theta = (1+2r^2+4r^2(cos^2\omega+sin^2\theta)+r^4)cos^2\theta \tag{3.15}$$

In light of the above equality, Equation (3.14) can be rewritten as (only for ω corresponding to the steepest slope):

$$\frac{d(\Phi(\omega))}{d\omega} = -2\left[\frac{1+(2r^3-2r)cos\omega cos\theta-r^4}{sin^2\theta(1-2r^2+r^4+4r^2cos^2\omega-4r^2cos^2\theta)}\right] \tag{3.16}$$

Applying Theorem 3.2.1, (i.e. $\omega=\theta$), the above equation can be simplified to:

$$\begin{aligned}
\frac{d(\Phi(\omega))}{d\omega} &= -2\left[\frac{1+(2r^3-2r)cos^2\theta-r^4}{sin^2\theta(1-2r^2+r^4)}\right] \\
&= -2\left[\frac{(1-r^2)(1+r^2)-2(1-r^2)rcos^2\theta}{(1-r^2)(1-r^2)sin^2\theta}\right] \\
&= -2\left[\frac{1-2rcos^2\theta+r^2}{(1-r^2)sin^2\theta}\right]
\end{aligned} \tag{3.17}$$

The above equation is evaluated for various values of θ and ω and the results are tabulated in Table 3.3. Notice that, for a given value of r, the value of $d(\Phi(\omega))/d\omega$ is not very sensitive to the value of θ. On the other hand, it is very sensitive to the

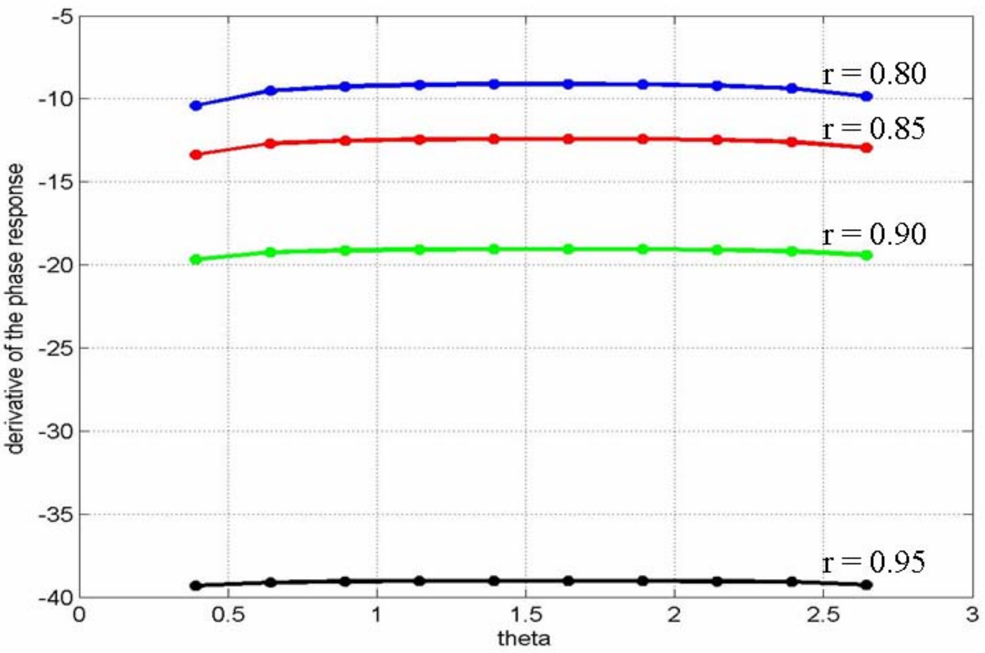

Figure 3.4: For a given value of r the derivative of the phase response evaluated at $\omega = \theta$ is approximately independent of the value of θ.

choice of r. It can thus be assumed that $d(\Phi(\omega))/d\omega$ is independent of θ. Fig. 3.4 makes it clear. This is stated more formally in the following theorem.

Theorem 3.2.2. *Consider a stable allpass filter with a pair of complex conjugate poles at $a = re^{j\theta}$ and a^*. Let ω_0 be the frequency at which the slope of the phase response, $\Phi(\omega)$, is the steepest. Then it is a relatively accurate assumption that $d(\Phi(\omega))/d\omega$ evaluated at ω_0 is independent of θ and is only dependent on the value of r.*

Thus the width of the *out-of-phase* region depends only on r and is relatively insensitive to the changes in θ. A simple closed form relation between the width of the *out-of-phase* region and the value of r cannot be derived using the derivative in

Table 3.3: Dependence of $d(\Phi(\omega))/d\omega$ w.r.t ω on r and θ.

	r=0.80	r=0.85	r=0.90	r=0.95
θ	$d(\Phi(\omega))/d\omega$	$d(\Phi(\omega))/d\omega$	$d(\Phi(\omega))/d\omega$	$d(\Phi(\omega))/d\omega$
0.393	-10.41	-13.36	-19.67	-39.32
0.643	-9.51	-12.70	-19.24	-39.12
0.893	-9.26	-12.52	-19.12	-39.06
1.143	-9.16	-12.45	-19.07	-39.04
1.393	-9.12	-12.42	-19.06	-39.03
1.643	-9.11	-12.42	-19.05	-39.03
1.893	-9.14	-12.43	-19.06	-39.03
2.143	-9.20	-12.48	-19.10	-39.05
2.393	-9.37	-12.60	-19.17	-39.08
2.643	-9.86	-12.96	-19.41	-39.20

equation 3.17 as tan^{-1} is a highly compressing nonlinearity and the *actual* value of this derivative is of little practical significance.

The findings of the above mathematical analysis of the phase response of an APF with poles at $a = re^{j\theta}$ and a^* can be summarized as:

1. The *out-of-phase* frequency region of the APF is centered near $\omega = \theta$, irrespective of the value of r.

2. The phase response at $\omega = \theta$ is approximately equal to $\pm\pi$.

3. The width of the *out-of-phase* frequency region is controlled only by the value of r, irrespective of the value of θ.

3.3 Detection of narrowband signals using the MPO structure

Consider a situation where we have to design a MPO structure to detect narrow band signals centered at ω_c and of bandwidths less than or equal to $\Delta\omega$. Let us first compute the parameters of the APF and then decide the parameters of the BPF. The parameters of the APF have to be chosen such that: (a) the phase response of the APF is about $-\pi$ at ω_c and (b) the *out-of-phase* region has a bandwidth of about $\Delta\omega$ centered at ω_c. From theorem 3.2.1, we know that the first condition is satisfied by choosing the pole, $a = re^{j\theta}$, of the APF such that $\theta = \omega_c$. Note that this value of θ will guarantee that the phase response of the APF is about $-\pi$ at ω_c irrespective of the value of r.

The bandwidth of the *out-of-phase* region is controlled by, $d(\Phi(\omega))/d\omega$, the derivative of the phase response of the APF. Theorem 3.2.2 states that the derivative of the phase response of the APF is controlled only by r and is independent of θ. Our aim is to use a value of r such that the phase response, $\Phi(\omega)$, of the APF spans $-\pi/2$ to $-3\pi/2$ (i.e. out-of-phase region) in $\Delta\omega$ radians centered around ω_c. This is feasible because, as equation (3.13) showed, the phase response at $\theta = \omega_c$ is approximately equal to $-\pi$ and the phase response is a continuous and monotonic function of ω. Equation (3.17) can be thought of as a linear approximation to the relation between r and the phase response at ω_c. But this relation cannot be extended for ω values far from ω_c as the phase response function (i.e. tan^{-1}) is highly nonlinear. The value of r satisfying the above conditions needs to be found using empirical experiments. Assume that the optimal value of r was found to be $r = r_c$ and the frequencies at which the phase response is $-\pi/2$ and $-3\pi/2$ to be ω_1 and ω_2, respectively. Fig. 3.5(a) shows the phase response of the APF corresponding to $\omega_c = 1000$ Hz and $\Delta\omega = 235$ Hz. The corresponding value of r_c is 0.91 and ω_1 and ω_2 are 895 and 1129 Hz, respectively.

We need to now decide the parameters of the BPF. The BPF has to satisfy two constraints: (a) The passband should include the *out-of-phase* frequency range (i.e. ω_1 to ω_2), and (b) the passband should also include some of the *in-phase* region. The second condition to ensure that the output of the MPO structure will be zero (or positive) when the input is a wideband signal (noise). Several BPFs can be designed that satisfy the above two constraints. Fig. 3.5(b-d) show three different choices of the BPF. The passband of the BPF in Fig. 3.5(b) is symmetric about the CF and the corresponding MPO structure is referred to as the symmetric MPO. The passbands of the BPFs in Fig. 3.5(c) and (d) are skewed upward and downward in

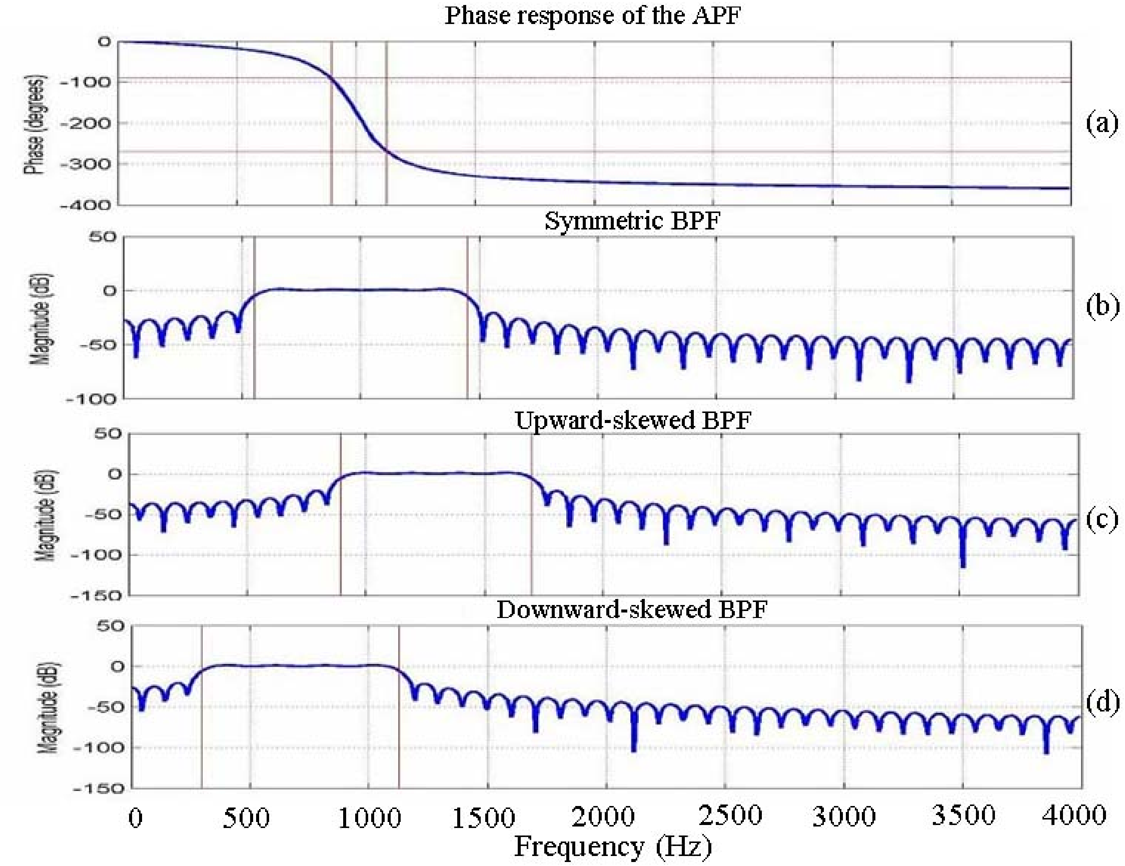

Figure 3.5: (a) Phase response of the APF corresponding to the MPO structure with CF=1000 Hz. Magnitude response of the corresponding (b) symmetric, (c) upward-skewed and (d) downward-skewed BPF.

frequency with respect to the CF and the corresponding MPO structures are referred to as the upward-skewed and downward-skewed MPO structures, respectively. In the initial version of MPO-based speech enhancement [92, 93], only the symmetric MPO structures were used. The downward and upward-skewed BPFs offer some advantages over the symmetric BPF and will be discussed in section 4.1. The next step is to decide on the bandwidth of the symmetric BPF. The optimal bandwidth of the BPF is computed by calculating the two-class (narrowband-signal-in-noise vs. noise-only) classification error for different choices of bandwidths and choosing the one that gives the least error. For low values of bandwidth the output for presence-of-signal situations as well as for absence-of-signal situations will be negative leading to many false-positive errors (Type I errors), whereas for high values of bandwidth the output for absence-of-signal situations as well as for presence-of-signal situations will be positive leading to many correct-miss errors (Type II errors). Fig. 3.6 plots the total classification error for a MPO structure that uses the APF shown in Fig. 3.5(a) and for different bandwidths of the corresponding symmetric BPF. The optimal BPF is $450 * 2 = 900$ Hz.

Fig. 3.7 shows the distribution of the output of the MPO model shown in Fig. 3.5(a) and 3.5(b) for 5000 frames each of white noise and a bandlimited signal centered at 1000 Hz and of bandwidth 235 Hz corrupted with white noise at ∞, 20, 10 and 0 dB SNR. Notice that the distribution of the output for white noise is well separated from that for the bandlimited signal at ∞ dB SNR. Moreover, the distribution of the bandlimited signal corrupted by white noise remains quite similar over the wide range of SNRs used in this study (∞ to 0 dB). The threshold

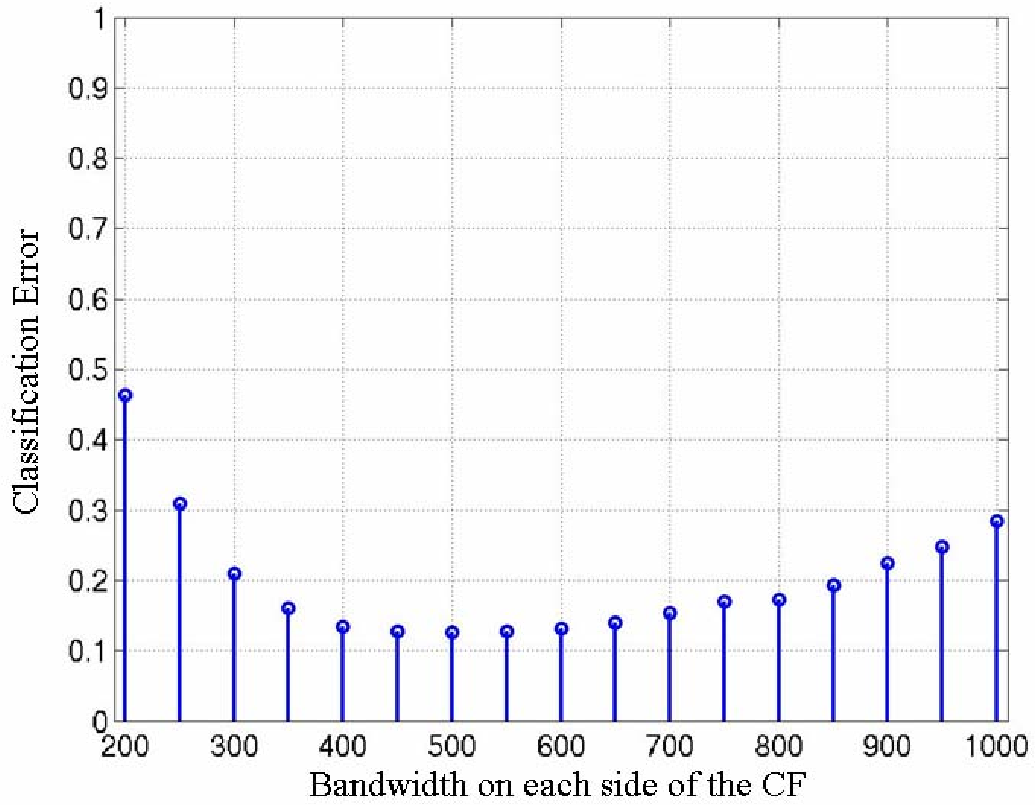

Figure 3.6: Variation in the binary classification error as the bandwidth of the BPF is varied. The two classes are: (a) presence of narrowband signal in broadband noise at 0 dB SNR and (b) broadband noise.

to discriminate the presence of signal from the absence of signal was computed using the Maximum Likelihood (ML)-based Likelihood Ratio Test (LRT) under the assumption that each of the distributions can be modeled as a Gaussian. The optimal threshold in this case is -0.0215 which, as expected, is very close to zero. Fig. 3.8 shows the Receiver Operating Characteristic (ROC) curve for MPO detectors at three different CFs: 950 Hz (red-dash curve), 1000 Hz (green-dotted curve) and 1050 Hz (blue-solid curve). The optimal threshold values are: -0.0183, -0.0215 and -0.0197, respectively. The ROC curves in the figure were obtained by varying the threshold over the range: $[opt\_thresh - 0.05 : -0.005]$ where opt_thresh is the optimal threshold for the corresponding MPO detector. In general, it is observed that the probability of false alarm is below 3% for threshold values below 0 and the probability of detection remains above 96% for threshold values above '$opt\_thresh - 0.05$' indicating that the exact value of the threshold is not critical for the overall operation of the MPO detectors. Note that the thresholds for the MPO detectors at different CFs are computed using the two extremes of (a) narrow-band signals centered at the CF and (b) white noise and are not retrained when the background conditions change. It is shown in Chapter 5 that the MPO speech enhancement scheme is robust to various noise types at different levels with no additional noise-specific training.

The next chapter describes how the MPO model can be used for enhancing speech signals corrupted by additive noise.

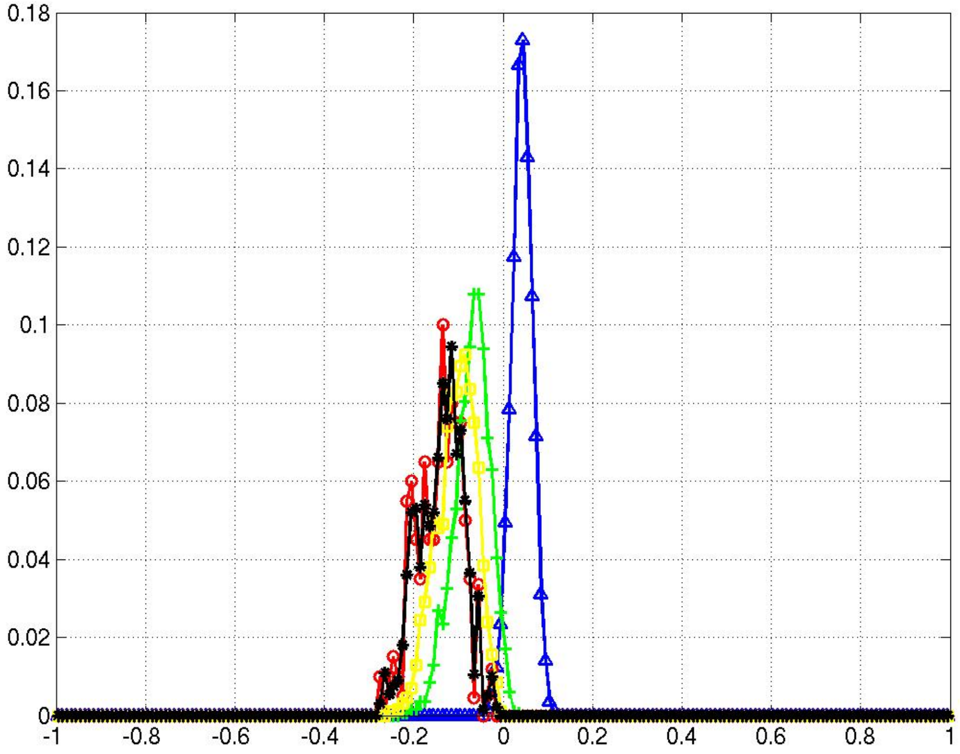

Figure 3.7: Distribution of the output of MPO model when the input is white noise (blue curve:△); bandlimited signal at ∞ dB SNR (red curve:o); at 20 dB SNR (black curve: *); at 5 dB SNR (yellow curve:□) and at 0 dB SNR (green curve: +).

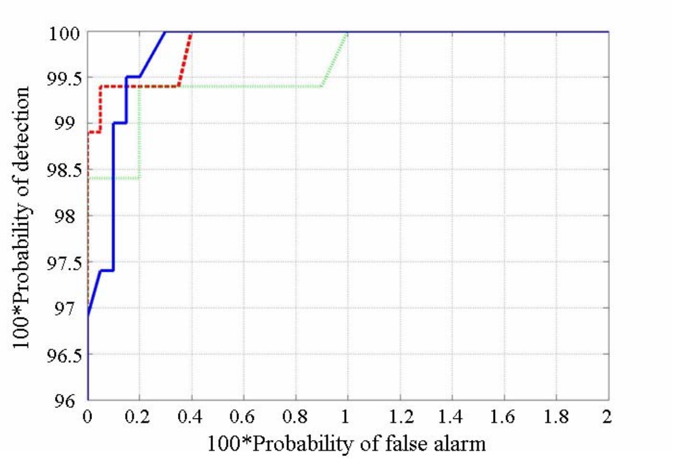

Figure 3.8: ROC curves for MPO detectors at three different CFs: 950 Hz (red-dash curve); 1000 Hz (green-dotted curve); 1050 Hz (blue-solid curve)

Chapter 4

MPO-based speech enhancement

Speech signals, for the most part, are composed of narrowband signals (i.e. harmonics) with varying amplitudes. The MPO-based speech enhancement scheme attempts to detect and maintain these time varying narrowband signals while attenuating the other spectro-temporal regions. Fig. 4.1 shows the schematic of the MPO-based speech enhancement scheme. The analysis-synthesis filterbank can be any near-Perfect Reconstruction (PR) filterbank. The overall performance of the MPO enhancement scheme is insensitive to the choice of the analysis-synthesis filterbank. In the present work, a DFT based PR filterbank is used. The input speech signal is split into overlapping frames of length 30 ms at a frame rate of 5 ms. Each MPO_i in the figure is a MPO structure (Fig. 3.2) with a different CF. The CFs are spaced every 50 Hz from 100 Hz to just below the maximum frequency. The threshold, x_i, to discriminate the presence of signal from the absence of signal is trained separately for each of the MPO structures as described in Section 3.3. The MPO structures act as switches allowing the spectro-temporal speech region to either pass as it is for reconstruction if the corresponding MPO output is less than the threshold (indicating presence of signal) or be greatly attenuated if the output is greater than or equal to the threshold (indicating absence of signal). The speech enhancement scheme can thus be thought of as applying a time-frequency two-dimensional

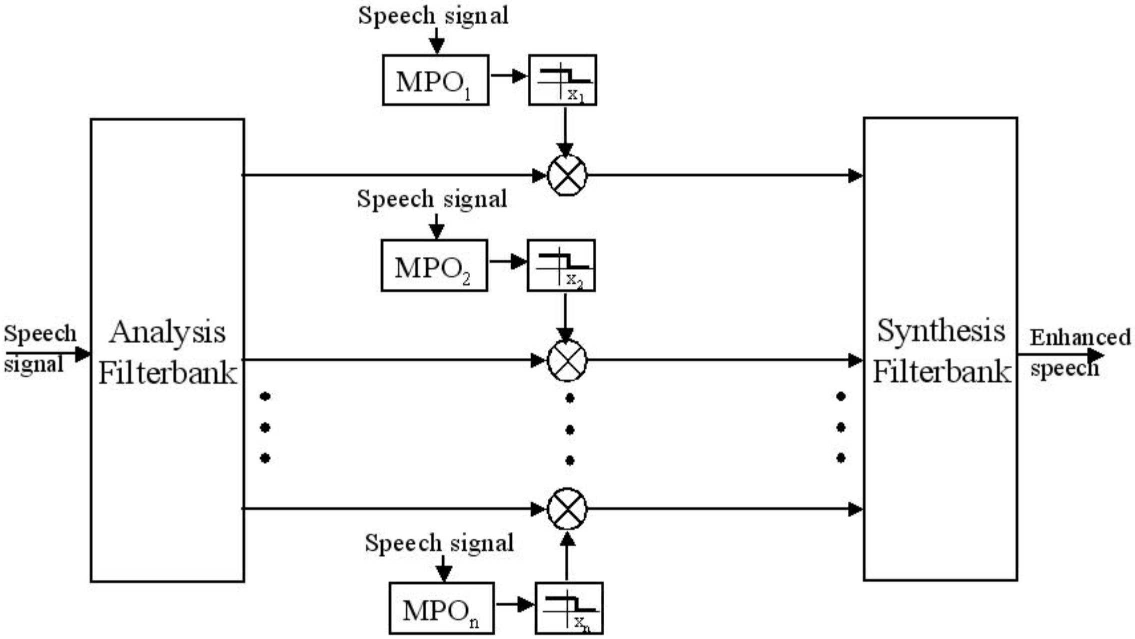

Figure 4.1: Schematic of the MPO-based speech enhancement scheme. The threshold, x_i, is trained using the ML-LRT technique and all the regions with output above this threshold are suppressed.

binary mask to the input speech signal. The binary mask has a value of one in spectro-temporal regions where the speech signal is dominant and has a value of zero where the noise signal is more dominant. The binary mask is referred to as the *MPO-profile*. Fig. 4.2(a,b) shows the spectrogram of a speech signal in clean and when it is corrupted by additive white noise at 10 dB SNR respectively. Fig. 4.2(c) shows the spectogram of the noisy speech signal overlaid with the *MPO-profile*. The *MPO profile* is 1 in the blue/dark regions and zero elsewhere. The use of binary masks is fairly common in auditory scene analysis based speech enhancement and robust speech recognition techniques [98]. The use of binary masks is motivated by the phenomenon of masking in human hearing, in which a strong signal masks all the weaker signals in its critical frequency band[49]. In the present speech enhancement method, spectro-temporal regions corresponding to a mask of zero are not completely eliminated before reconstruction. Instead, they are greatly attenuated.

In the initial version of the MPO-based speech enhancement scheme, each of the MPO_i in Fig. 4.1 consisted of a symmetric BPF and the APF was configured so that signals centered at the CF of the MPO and with bandwidths less than or equal to 235 Hz would lead to negative outputs. Such a scheme performs well when the input speech signal is corrupted by additive white noise which has a relatively flat spectrum with minimal level fluctuations over time. Fig. 4.2(c) shows the *MPO profile* for a speech signal corrupted by additive white noise at 10 dB SNR. Notice that the *MPO profile* is 1 in most of the speech-dominant regions and is zero in most of the noise-dominant regions. Contrast this with Fig. 4.2(e) which shows the spectrogram of the same speech signal corrupted by additive subway noise at 10 dB

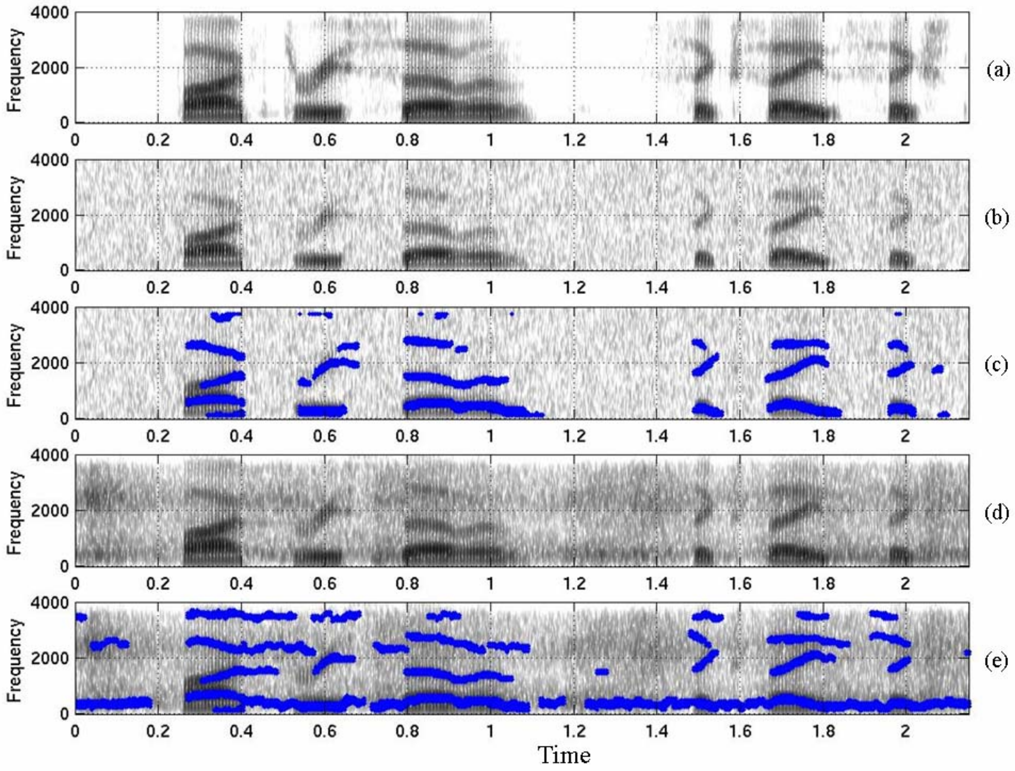

Figure 4.2: (a) Spectrogram of clean speech utterance 'Five three seven six eight six'. (b) Spectrogram of the speech signal corrupted by additive white noise at 10 dB SNR. (c) Spectrogram of the noisy speech signal overlaid with the corresponding *MPO profile*. (d) Spectrogram of the speech signal corrupted by subway noise at 10 dB SNR. (e)Spectrogram of the noisy speech signal overlaid with the corresponding *MPO profile*.

SNR overlaid with the corresponding *MPO profile*. The *MPO profile* is 1 not just in most of the speech-dominant regions but also in a lot of the noise-only regions. Two general shortcomings of the MPO speech enhancement scheme stand out: (a) Some of the important speech information is missed even at a relatively high SNR of 10 dB. For example, in Fig. 4.2(c) the F2 information (near 1000 Hz) at the beginning of the word 'five' (around 0.26 sec) is missed. (b) The *MPO profile* is 1 in many noise-only regions. This will retain a lot of noise in the reconstructed speech signal. Efficient removal of such colored noise while maintaining most of the speech information calls for a closer look at the MPO structures used at each CF.

4.1 Choosing the BPF

Consider the spectral slice shown in Fig. 4.3(e). Our aim is to detect the F2 region (second formant around 1050 Hz). For the right choice of the BPF, a MPO strcuture with CF=1000 Hz and APF as shown in Fig. 4.3(a) will be able to detect the F2. The harmonics close to F2 fall in the out-of-phase frequency region of the APF. The harmonics close to F1 (around 550 Hz) fall in the in-phase frequency region and are also in the passband of the symmetric BPF. The amplitude of F1 (and hence that of the harmonics close to F1) is greater than that of F2 due to the known spectral tilt in sonorant regions of speech signals. As a result, although there is a strong narrow band signal at the CF of the MPO, the output of this MPO structure will be positive and therefore the speech information present in that frequency region will be missed. The upward skewed BPF shown in Fig. 4.3(b), on the other hand, will attenuate the F1 region and thus the output of the upward-skewed MPO structure will be driven only by the frequency content near and above the CF. Most of the time such upward skewed MPO structures are able to correctly detect the speech information as they inherently take advantage of the spectral tilt present in sonorant speech regions. The F2 information in Fig. 4.3(e) that was missed by the symmetric MPO structure will be detected by the upward-skewed MPO structure.

Consider the spectral slice shown in Fig. 4.4(e). The spectrum is typical of front vowels which have second formant well above 1500 Hz (e.g. /iy/ in 'three'). In this case, F2 and F3 are of comparable amplitudes and are in close proximity in frequency. Hence, the harmonics near these formant frequencies also have comparable amplitudes. Our aim is to detect the F2 region (around 2300 Hz). For the right choice of the BPF, a MPO structure with CF=2300 Hz and APF as shown in Fig. 4.4(d) will be able to detect the F2. The downward-skewed filter shown in Fig. 4.4(c) is the exact opposite of the upward-skewed filter shown in Fig. 4.4(b). Its passband extends downwards in frequency with respect to the CF of the MPO structure. The upward-skewed MPO structures will detect the higher frequency harmonics corresponding to F3 but will fail to detect the lower ones corresponding to F2.The downward-skewed MPO structure centered on the lower frequency harmonics can successfully detect such instances as its passband extends only on the lower frequency side attenuating the high-amplitude high-frequency harmonics.

Thus, for robust detection of speech information, each CF needs to be analyzed using an upward MPO structure as well as a downward MPO structure. Fig. 4.5(d)

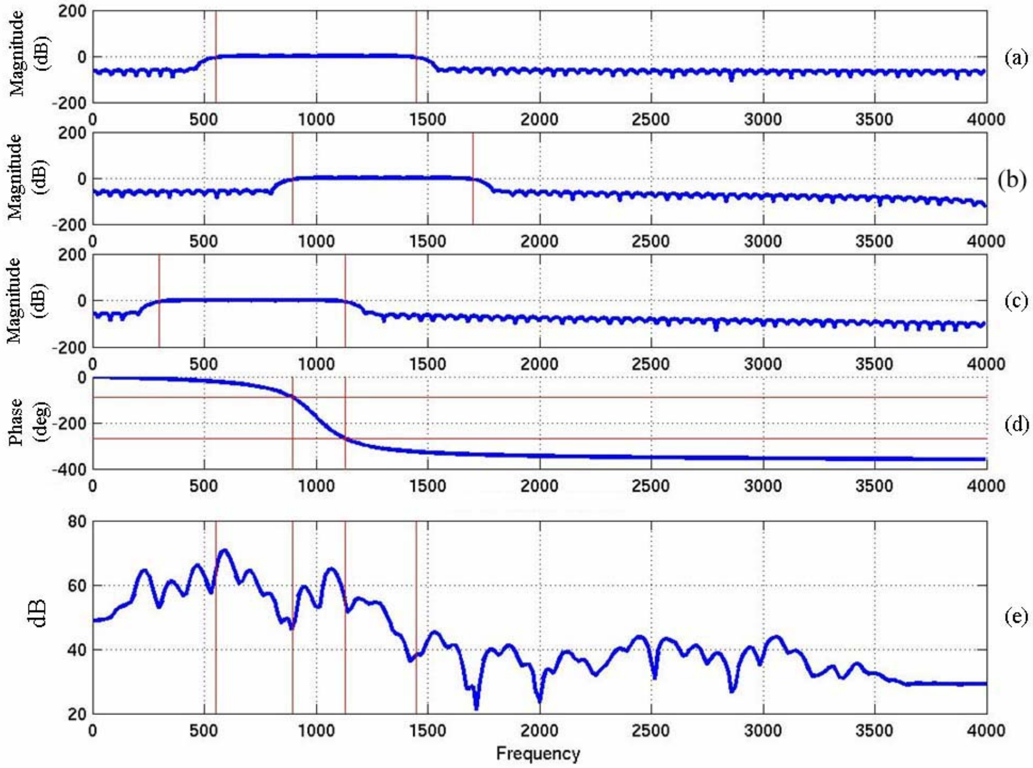

Figure 4.3: This figure shows a case where the upward-skewed MPO structure is better suited. Magnitude response of (a) Symmetric BPF (b) Upward-skewed BPF (c) Downward-skewed BPF (d) Phase response of an APF with CF = 1000 Hz. (e) Spectral slice of a speech signal.

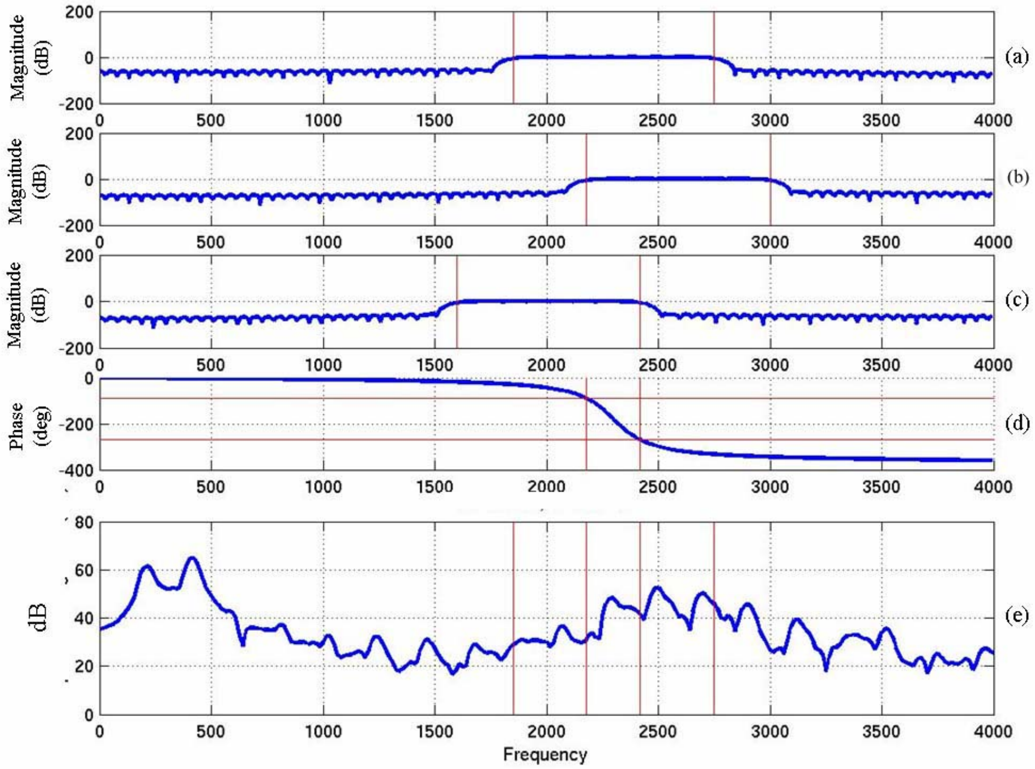

Figure 4.4: This figure shows a case where the downward-skewed MPO structure is better suited. Magnitude response of (a) Symmetric BPF (b) Upward-skewed BPF (c) Downward-skewed BPF (d) Phase response of an APF with CF = 2300 Hz. (e) Spectral slice of a speech signal.

shows the *MPO profile* obtained when each CF was analyzed using an upward-skewed and a downward-skewed MPO structure. The *MPO profile* has a value of one if either the output of the upward-skewed MPO structure is below the corresponding threshold or if the output of the downward-skewed MPO strucutre is below the corresponding threshold. Comparing this with the *MPO profile* obtained using the symmetric MPO structures at each CF (plotted in Fig. 4.5(c)) shows that the use of skewed MPO structures retains all the speech information but passes a lot of noise.

4.2 Noise removal

To reduce the number of occurances where the *MPO profile* is 1 in noise-only regions, the MPO speech enhancement scheme uses a set of downward-skewed and upward-skewed MPO structures at each CF. Each set has MPO structures with a different *out-of-phase* region ranging from 120 Hz to 250 Hz. Noise can be wrongly seen as speech signal by one or more of the different MPO structures in the set, but it is rarely seen as a narrowband speech signal by all the structures. Similarly, narrowband speech signals are almost always seen as speech signals by *all* the MPO structures. For a given spectro-temporal region, the *MPO profile* is set to zero if the output of even one of the MPO structures is above the corresponding threshold.

The overall speech enhancement scheme can now be summarized in the following two steps: In the first step, the temporal regions where speech is present are computed. For a temporal region to be voted as *speech present*, it has to satisfy

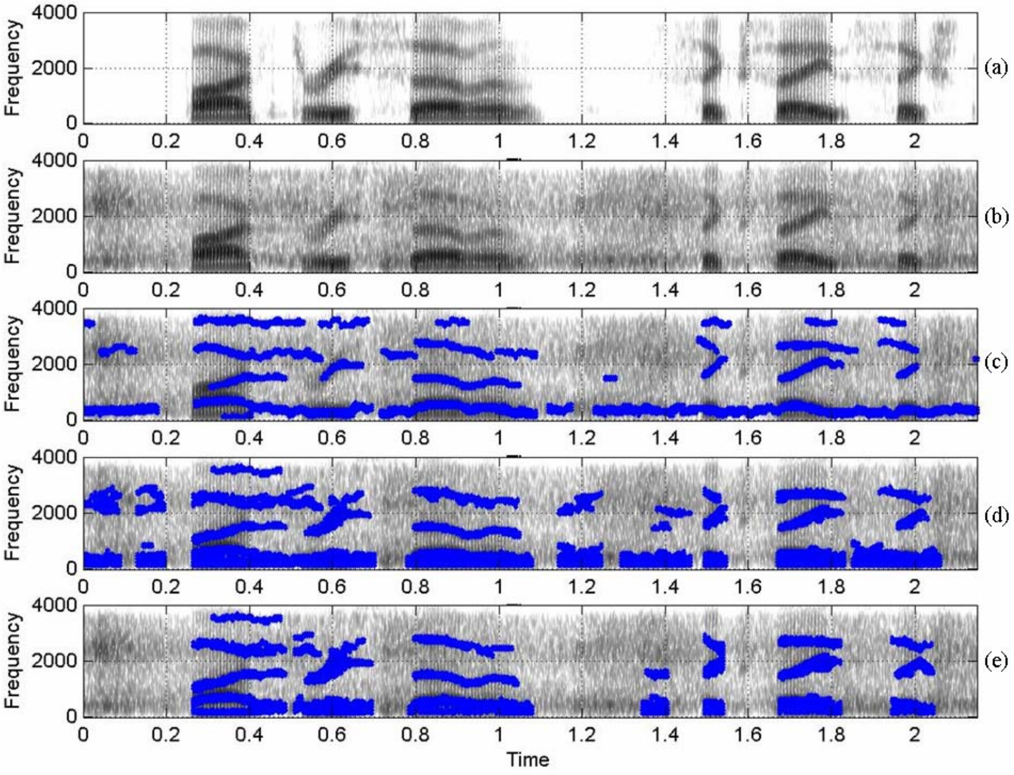

Figure 4.5: Spectrograms of (a) clean speech (b) speech signal corrupted by subway noise at 10 dB SNR. *MPO profiles* of the noisy speech signal computed by the (c) MPO scheme with one symmetric MPO structure at each CF (d) MPO scheme with an upward-skewed and downward-skewed MPO structures at each CF. (e) MPO scheme with a set of upward-skewed and downward-skewed MPO structures at each CF.

two conditions: (a) The MPO output of at least one frequency channel from all the different upward-skewed or all the different downward-skewed MPO structures should be at least four times more negative than the threshold for that particular channel, and (b) The temporal region should be at least 50 ms long. A duration of 50ms was chosen to retain most of the /I/ sounds in 'six' while removing the short-duration noise that is wrongly seen as speech.

In the second step, the frequency channels within the *speech-present* temporal regions where speech information is present are computed by finding the channels where the MPO output from all the five upward skewed or all the five downward skewed MPO structures is below the corresponding threshold. The noisy speech signal from only these channels is used for reconstruction.

Fig. 4.5(e) shows the *MPO profile* obtained when each CF is analyzed using a set of downward-skewed and upward-skewed MPO structures. Notice that all of the sonorant speech information is maintained while a lot more noise is suppressed compared to the case where only one downward-skewed and upward-skewed MPO structure was used at each CF (Fig. 4.5(d)).

4.3 Attenuating the *speech-absent* regions

As mentioned earlier, the MPO-processing leads to a binary mask, called the *MPO profile* that classifies each spectro-temporal region as either *speech-present* or *speech-absent*. The signal in the speech-present regions is used 'as-is' to construct the enhanced speech signal. The signal in the rest of the regions is greatly attenuated before being used for reconstruction. Attenuating all the *speech-absent* regions is suboptimal as it lends an unnatural characteristic to the enhanced speech. In the MPO enhancement scheme, the weighing scheme for the *speech-absent* channels in *speech-present* temporal regions is based on the transfer function associated with a conjugate pair of poles corresponding to the centroid of the frequencies of the contiguous speech-present channels. The transfer function is similar to the general form of the vocal tract transfer function derived in [97]:

$$T_n(s) = \frac{s_n s_n^*}{(s - s_n)(s - s_n^*)} \tag{4.1}$$

where $s = j2\pi f, s_n$ is the complex frequency of the pole, and $s_n = \sigma_n + j2\pi F_n$. The value of σ_n is chosen such that the bandwidth of the pole is 100 Hz. Such an attenuation scheme reduces the perceptual artifacts introduced by the enhancement technique. The weighing scheme corresponding to the frame centered at 825 ms of the utterance shown in Fig. 4.5 is displayed in Fig. 4.6. The F_n values for this frame are: 550, 1500 and 2750 Hz. The signal in the *speech absent* temporal regions, the temporal frames where the *MPO profile* has a value of zero for all the frequency channels, is uniformly attenuated by 20 dB.

Fig. 4.7(d) shows the spectrogram of the MPO-enhanced speech signal that was corrupted by subway noise at 10 dB SNR. Notice that all of the sonorant speech information is maintained while most of the noise is removed.

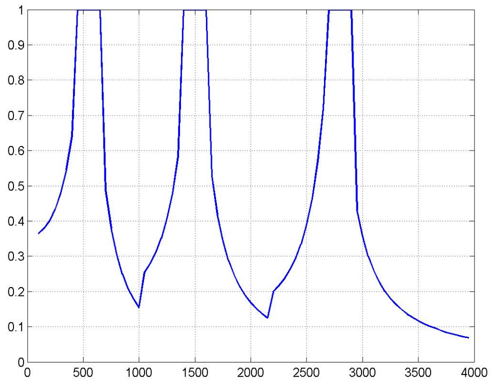

Figure 4.6: Spectral weighing scheme. X-axis is frequency in Hz. Y-axis is the value of the weight.

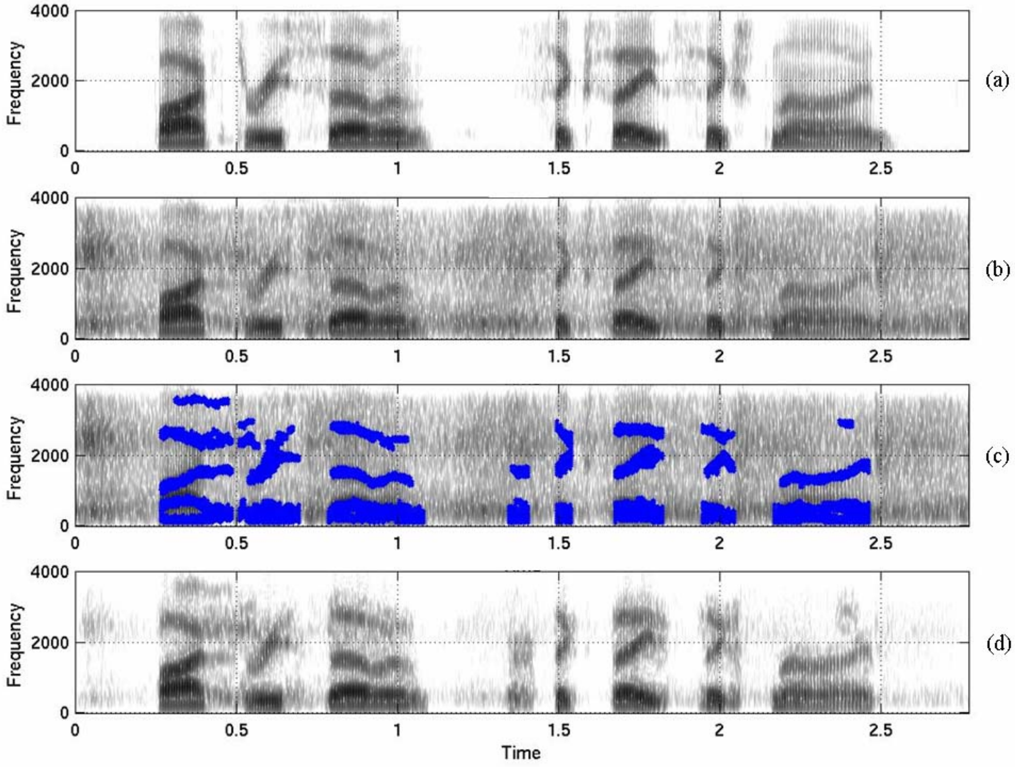

Figure 4.7: Spectrograms of (a) clean speech signal (b) speech signal corrupted by subway noise at 10 dB SNR (d) MPO-enhanced speech signal (c) spectrogram of the noise speech signal overlaid with the *MPO profile*.

Chapter 5

Results

The quality of the speech signals enhanced using the proposed MPO speech enhancement scheme was evaluated using several different objective quality assessment measures as well as subjective evaluations on human listeners with normal hearing. The performance of the MPO enhancement scheme on these tasks was also compared with that of some of the other enhancement schemes proposed in the literature. The MPO speech enhancement scheme was also used as a preprocessor for a robust automatic speech recognition system.

5.1 Databases

Two databases were used to evaluate and compare the performance of the different speech enhancement schemes. The noisy data in both the databases was obtained by artificially adding the noise signals and thus does not account for the Lombard effect described earlier in Chapter 1. The default HMM-based recognizers provided by the two databases were used for the robust automatic speech recognition experiments so that only the effect of the speech enhancement preprocessing block will be evaluated.

1. *Aurora Database :* The Aurora database [99] is formed from the TIDigits database [100] that consists of recordings of 111 male and 114 female Amer-

ican adults speaking English digits (*zero* to *nine* and *oh*) in sequences of one to seven digits in a quiet acoustic enclosure and digitized at 20 kHz. The Aurora database is constructed by first downsampling the TIDigits data to 8 kHz. The speech signals are then filtered through one of the two standard filters that simulate the frequency characteristics of equipments used in the telecommunication area. The transfer functions of the two filters (G.712 and Modified Intermediate Reference Systen (MIRS)) is shown in Fig. 5.1. Different types of noise are digitally added to these utterances at varying SNRs. The different noise types are: (1) subway noise (2) babble noise (3) car noise (4) exhibition hall noise (5) restaurant noise (6) street noise (7) airport noise and (8) train station noise. Fig. 5.2 shows the long-term spectra of these noise types. The temporal variability of these noise types is not captured by these long-term spectra. The different SNRs considered are ∞, 20, 15, 10, 5, 0, -5 dB. SNR is defined as the global ratio of the energy of the speech signal and the noise signal. The database is partitioned into a training subset and three test subsets. The training subset consists of a set of speech signals filtered by the G.712 filter either in clean or corrupted by either of the noise types (1) to (4) at either 20, 15, 10 or 5 dB SNRs. The three test subsets consist of: subset a which consits of speech signals filtered by the G.712 filter and corrupted by either of the noise types (1) to (4) at seven different SNRs: ∞, 20, 15, 10, 5, 0, -5 dB; subset b which consits of the same set of speech signals as in (a) but corrupted by either of the noise types (5) to (8) at seven different SNRs: ∞, 20, 15, 10, 5, 0, -5 dB and subset c which consists of the same set of speech

Table 5.1: The noise composition of each of the three test subsets of the Aurora database

subset	Noise types			
a	A1: subway	A2: babble	A3: Car	A4: exhibition hall
b	B1: restaurant	B2: street	B3: airport	B4: train station
c	C1: subway	C2: street	-	-

signals as in subset a but corrupted by noisy type (1) or (6) and filtered by the MIRS filter. The same set of utterances is used in all the three subsets. Table 5.1 shows the different types of noise used in the three subsets. For a given column in the table, the set of utterances used were the same. For example, the set of utterances corrupted by babble noise to form a part of the 'subset a' was the same set corrupted by street noise to form a part of the 'subset b' as well as a part of the 'subset c'. (The difference is that the signals in 'subset c' are filtered by the MIRS filter.)

In the present work, the performance of the different enhancement techniques in terms of increase in the objective distortion measures, subjective quality of the enhanced speech signals and improvement in the accuracy of robust automatic speech recognition systems was evaluated using the Aurora database.

2. *GRID database[101]:* The GRID database consists of recordings of 16 female and 18 male speaking structured sentences in a quiet acoustically-isolated booth. The recordings were digitized at 25 kHz. The sentences are of the form:

$$< command : 4 >< color : 4 >< preposition : 4 >< letter : 25 >< number : 10 >< adverb : 4 >$$

The numbers in brackets indicate the number of choices at each point. Only the 'color', 'letter' and 'number' were designated as the key words to be recognized by the automatic speech recognition systems. Each subject produced all the combinations of these three key words leading to a total of 1000 ($4 * 25 * 10$) sentences per subject. The training set consists of the 17,000 sentences (500 from each of the 34 speakers) in clean. The test set consists of two subsets: (a) The clean speech corrupted by speech-shaped noise with a spectrum similar to the long-term spectrum of the GRID database at 6, 0, -6, -12 dB SNRs. The corresponding clean utterances also form a part of this subset. (b) Pairs of utterances were acoustically mixed at 6 different Target-to-Masker Ratios (TMRs) (6, 3, 0, -3, -6, -9 dB) to simulate the two-talker condition. Target speech signal is the one that needs to detected and masker speech signal is the interfering speech signal. All the target utterances in clean are also included in this subset. All the target utterances contain the word 'white'. In one third of the utterances, the masker utterance and the target utterance are spoken by the same speaker. In one third of the utterances, the masker utterance and the target utterance are spoken by different subjects of the same gender and in the remaining utterances, the masker and the target utterances are spoken by subjects of different genders.

In the present work, the GRID database was used to evaluate the performance of the MPO speech enhancement scheme as a preprocessing block for the robust speech recognition system.

5.2 Binary mask based evaluations

As mentioned in Chapter 4, the output of the MPO speech enhancement scheme can be thought of as a binary spectro-temporal mask that has a value of 1 if the speech energy in the particular time-frequency channel is more dominant than the energy of the corrupting noise (*speech-present* regions) and a value of 0 if the energy of the noise signal is more dominant than the speech energy (*speech-absent* regions). In the actual MPO enhancement scheme, the *speech-absent* regions are not completely removed but are greatly attenuated as mentioned in Section 4.3. The binary mask is referred to as the *MPO-profile* and is a convenient tool to analyze the performance of the enhancement scheme.

For a given utterance, the ground truth about the *speech-present* and *speech-absent* regions is computed using the energy-based *maximal mask*. The *maximal mask* has a value of 1 if the following three conditions are satisfied: (a) the overall energy of the time frame is no less than 7.5% of the maximum frame energy over the entire utterance, (b) the channel energy is no less than 2% of the maximum channel energy in the given frame and (c) the temporal region is sonorant as detected by the Aperiodicity, Periodicity, Pitch (APP) detector [95]. The third condition ensures that the evaluation is restricted only to the sonorant regions as the MPO processing scheme does not retain the obstruents, especially at low SNRs. The energy thresh-

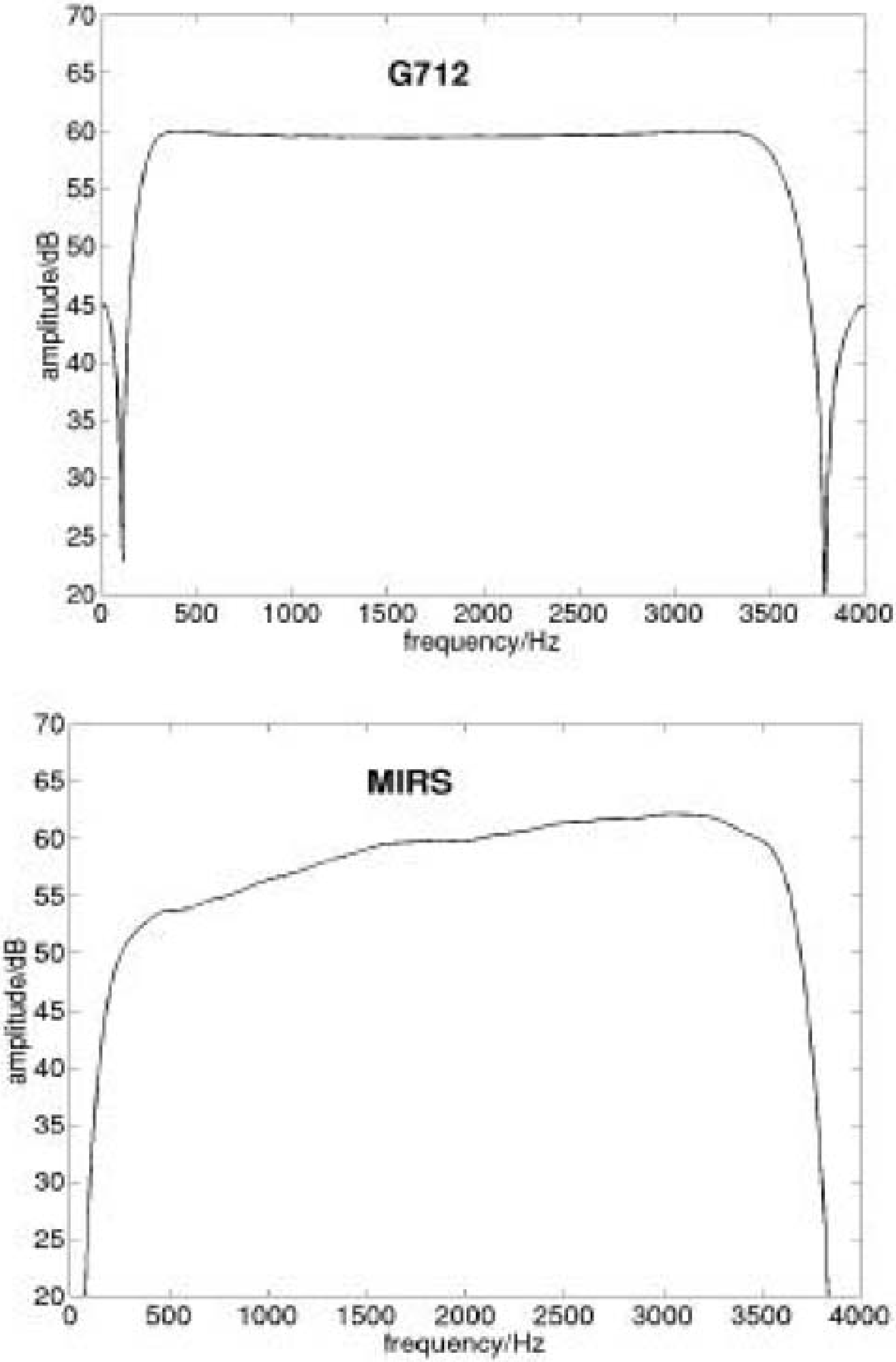

Figure 5.1: Frequency responses of the G.712 and MIRS filters. Figure adopted from [99].

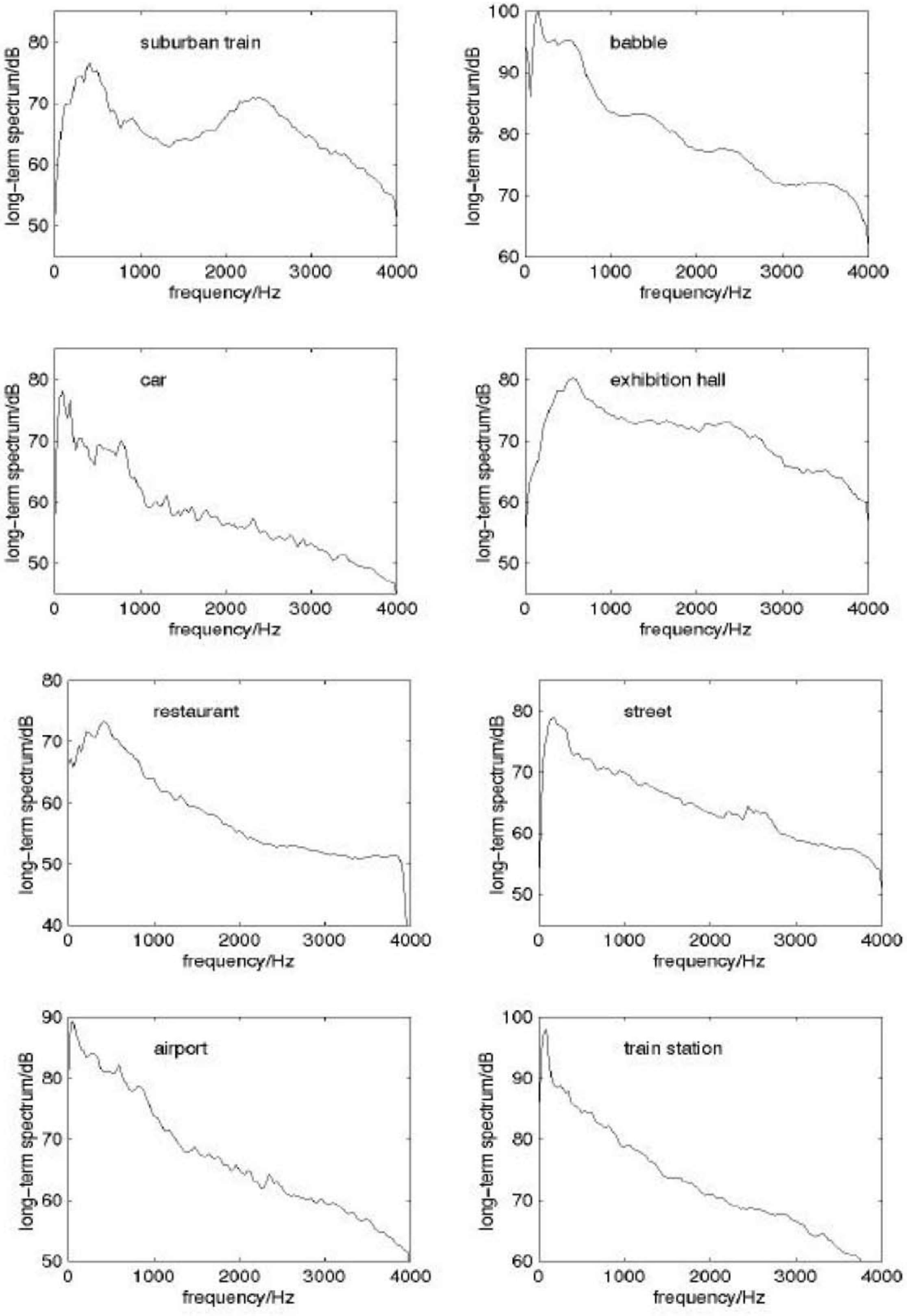

Figure 5.2: Long-term spectra of the different types of noise used in the Aurora database. Figure adopted from [99].

olds were chosen such that all of the speech information was retained. Informal hearing tests of some of the randomly chosen clean speech signals reconstructed using their corresponding *maximal mask* confirm that the reconstructed clean speech signals are very similar to the original clean speech signals. Fig. 5.3(b) shows the *maximal mask* for the utterance shown in 5.3(a). The regions where the *maximal mask* has a value of 1 (i.e. *speech-present* regions) are indicated by the blue (dark) regions.

Two other kinds of masks proposed by other researchers are:

1. *Ideal mask [98]:* An ideal mask is a binary spectro-temporal mask where a value of 1 indicates that the target energy is stronger than the noise energy within the corresponding spectro-temporal channel and a value of 0 indicates otherwise.

2. *A-priori mask [72]:* An a-priori mask is a binary spectro-temporal mask which has a value of 1 if the mixture energy in a given spectro-temporal region is within 3 dB of the target energy and a value of 0 otherwise.

It can easily be shown that the a-priori mask is identical to the ideal mask in situations where the speech signals are corrupted by additive noise:

Let S be the energy of the speech signal in a given spectro-temporal channel and Y be the energy of the mixture signal when the speech signal is corrupted by some additive noise implying that the energy of the corrupting noise is $Y - S$. The ideal mask is 1 if:

$$\begin{aligned} & (Y - S) < S \\ i.e. \quad & \frac{Y - S}{S} < 1 \\ i.e. \quad & \frac{Y}{S} < 2 \\ i.e. \quad & 10log_{10}\left[\frac{Y}{S}\right] < 3dB \end{aligned}$$

Where the last condition is the same as that used by the a-priori mask.

The energy-based *maximal mask* was preferred to the ideal mask for two main reasons: (1) The ideal mask does not have a high spectral resolution and cannot distinguish the spectral peaks from the spectral valleys even at low SNR. This lack of discrimination can be seen in Fig. 5.3(c-e) which shows the ideal mask at 20, 10 and 0 dB SNRs respectively for the utterance shown in Fig. 5.3(a). (2) For a given utterance, the ideal mask changes as the level of the corrupting noise changes. This can also be seen in Fig. 5.3(c-e).

The *MPO profile* at different SNRs is compared with the corresponding *maximal mask* to compute the percentage of correctness and percentage of insertions of spectro-temporal channels. The percentage of correctness is defined as the ratio of the number of spectro-temporal channels where both the *maximal mask* and the noisy *MPO profile* have an output of 1 to the total number of spectro-temporal channels where the *maximal mask* has an output of 1. The percentage of insertion is defined as the ratio of the number of spectro-temporal channels where the *maximal mask* has an output of 0 and the noisy *MPO profile* has an output of 1 to the total number of spectro-temporal channels where the *maximal mask* has an output of 1.

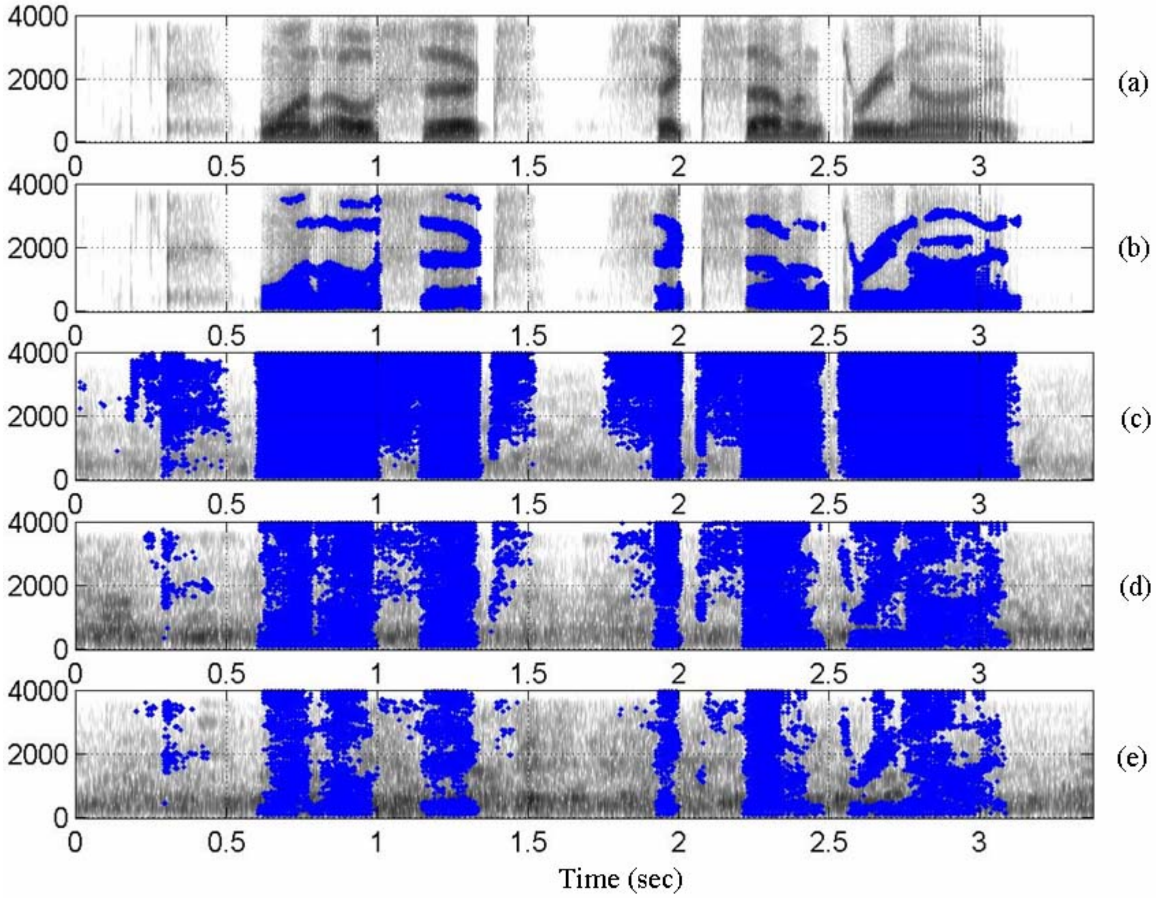

Figure 5.3: (a) Spectrogram of the utterance ' one oh six six seven three nine' (b) the energy based *maximal mask*, (c-e) ideal mask when the utterance is corrupted at 20, 5 and 0 dB SNR respectively.

Table. 5.2 shows the percentage correctness and insertion for 80 7-digit long utterances randomly chosen from the 'subset a' where the corrupting noise types were either (a) subway noise (b) babble noise (c) car noise or (d) exhibition hall noise at various SNRs. Each of the utterances is corrupted at seven different SNRs: ∞, 20, 15, 10, 5, 0 and -5 dB. Table. 5.3 shows the percentage correctness and insertion for the same 80 utterances chosen from the 'subset b' where the corrupting noise types were either (a) restaurant (b) street (c) airport or (d) train station noise at various SNRs. Table. 5.4 shows the percentage correctness and insertion for 40 of the above utterances that were also found in the 'subset c' where the corrupting noise types were either (a) subway or (b) street noise at various SNRs. The percentage of insertions in babble noise are much higher than in any of the other noise types. This is because a considerable amount of narrowband babble noise is seen as speech by the MPO analysis.

Fig. 5.4 compares the energy-based *maximal mask* of the utterance 'five three seven six eight six nine' with the *MPO profiles* computed in clean and when the utterance is corrupted by subway noise at 20, 10 and 5 dB SNRs. Fig. 5.5 compares the energy-based *maximal mask* of the utterance 'six six five four five nine nine' with the *MPO profiles* computed in clean and when the utterance is corrupted by street noise at 20, 10 and 5 dB SNRs. The percentage correctness (and insertion) values for these utterances are: (1) ∞: 79.0(43.8); 20 dB: 68.6(19.0); 10 dB: 62.6(19.2); 5 dB: 52.6(17.0) and (2) ∞: 74.3(18.7); 20 dB: 63.2(13.7); 10 dB: 51.2(10.1); 5 dB: 40.5 (5.0), respectively. Several inferences can be drawn from these two figures:

Table 5.2: Average percentage correctness (and insertion) for 80 7-digit long utterances corrupted by one of the four noise types in 'subset a' at various SNRs

SNR	subway	babble	car	exhibition hall
∞	83.7(44.7)	83.2(38.5)	83.4(39.6)	83.1(39.0)
20	70.1(20.0)	75.8(61.5)	71.2(18.9)	67.5(21.7)
15	64.4(18.1)	73.2(66.0)	66.1(16.8)	60.2(20.5)
10	57.5(16.7)	68.5(62.2)	58.8(13.9)	52.6(22.0)
5	47.7(14.8)	63.3(66.9)	46.6(12.5)	40.6(21.6)
0	34.4(12.8)	57.0(67.2)	32.7(9.5)	26.4(19.5)
-5	19.0(9.1)	49.5(68.8)	17.5(7.4)	14.1(17.1)

Table 5.3: Average percentage correctness (and insertion) for the same 80 7-digit long utterances used in Table 5.2 but here the corrupting noise types are different and are chosen from the 'subset b'

SNR	restaurant	street	airport	train station
∞	83.7(44.7)	83.2(38.5)	83.4(39.6)	83.1(39.0)
20	75.9(61.2)	71.4(31.3)	76.7(55.8)	72.0(29.7)
15	72.0(60.3)	65.3(23.4)	72.8(56.9)	66.6(28.9)
10	67.8(58.2)	60.5(32.0)	67.5(55.4)	61.6(28.7)
5	61.8(61.6)	51.2(30.7)	64.2(63.8)	55.1(38.2)
0	55.7(58.8)	38.8(22.9)	54.6(60.9)	39.6(31.6)
-5	46.1(59.2)	21.1(22.4)	47.9(64.3)	26.2(25.2)

Table 5.4: Average percentage correctness (and insertion) for 40 out of the 80 7-digit used in Table 5.2 that were also found in the 'subset c'

SNR	subway	street
∞	74.1(27.6)	74.7(23.
20	59.6(11.2)	61.4(14.3)
15	54.0(10.4)	54.2(12.7)
10	45.8(8.5)	48.6(15.3)
5	37.3(9.5)	36.3(11.2)
0	22.6(5.8)	23.5(11.2)
-5	9.7(3.7)	12.9(11.7)

1. The *maximal mask* captures not just the strong-amplitude formant regions but also captures most of the not-so-weak spectral valleys in between the formants. Increasing the energy threshold while computing the *maximal mask* will remove most of these spectral valleys but will also remove the low amplitude high frequency formant information. The *MPO profile* computed on the clean utterance captures all of the perceptually significant high-amplitude spectral information in the sonorant regions as well as some of the frequency of onset of frication in the fricative regions. The (perceptually less significant) valleys between the formants are not captured by the *MPO profile.* As a result, the percentage of correctness for ∞ SNR is not very high. The same effect is propagated to lower SNRs also. It will be shown in Section 5.6 that human listeners prefer the MPO-processed clean speech signals about as many times

as they prefer the original clean speech signals.

2. As the SNR is reduced, the *MPO profile* retains most of the spectral peaks while very little extra noise is passed. At low SNRs of 5 dB and below some of the relatively weak formant information is not detected. The spurious noise regions passed by the MPO enhancement scheme are mainly narrow bandwidths and are for short intervals leading to the well known musical-noise phenomenon.

Fig. 5.6 compares the *MPO profiles* for the utterance 'five three seven six eight six nine' when it is filtered by G.712 filter and corrupted by subway noise at 10 dB SNR (Fig. 5.6(c)), filtered by G.712 filter and corrupted by restaurant noise at 10 dB SNR (Fig. 5.6(e)) and filtered by MIRS filter and corrupted by subway noise at 10 dB SNR (Fig. 5.6(g)). Fig. 5.7 compares the *MPO profiles* for the utterance 'eight zero one one two four three' when it is filtered by G.712 filter and corrupted by babble noise at 10 dB SNR (Fig. 5.7(c)), filtered by G.712 filter and corrupted by street noise at 10 dB SNR (Fig. 5.7(e)) and filtered by MIRS filter and corrupted by street noise at 10 dB SNR (Fig. 5.7(g)). Note that for both the utterances, the G.712 filtering has a slightly less adverse effect on the MPO-processing than the MIRS filtering although in both the cases most of the formant-related spectral peaks are retained by the MPO processing. The street noise has a spectral peak around 2500 Hz (Fig. 5.2). This peak is made more prominent by the MIRS filter which attenuates the lower frequencies and is evident in Fig. 5.7(f). The babble noise shown in Fig. 5.7(b) consists of a large number of speakers speaking simultaneously

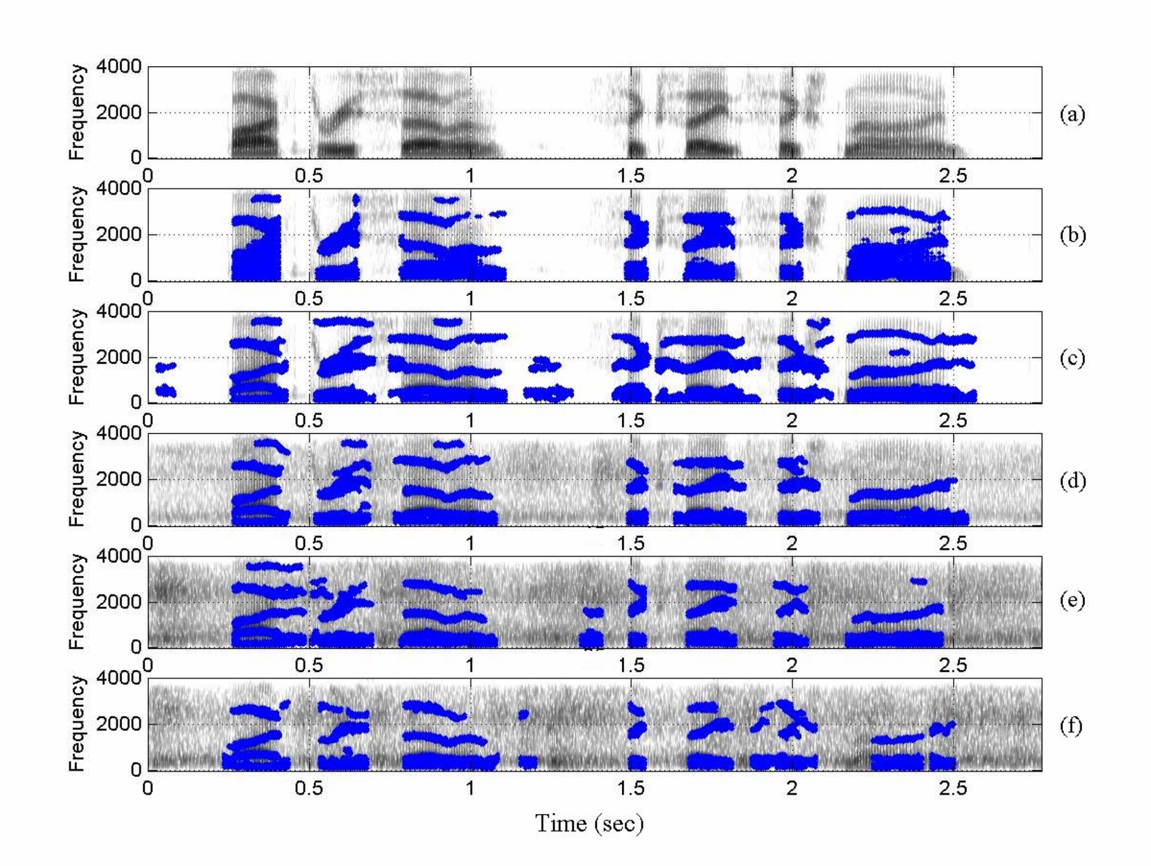

Figure 5.4: (a) Spectrogram of the utterance 'five three seven six eight six nine'; (b) the energy-based *maximal mask* (c-f) the *MPO profile* at ∞, 20, 10 and 5 dB SNR respectively when the corrupting noise is subway noise.

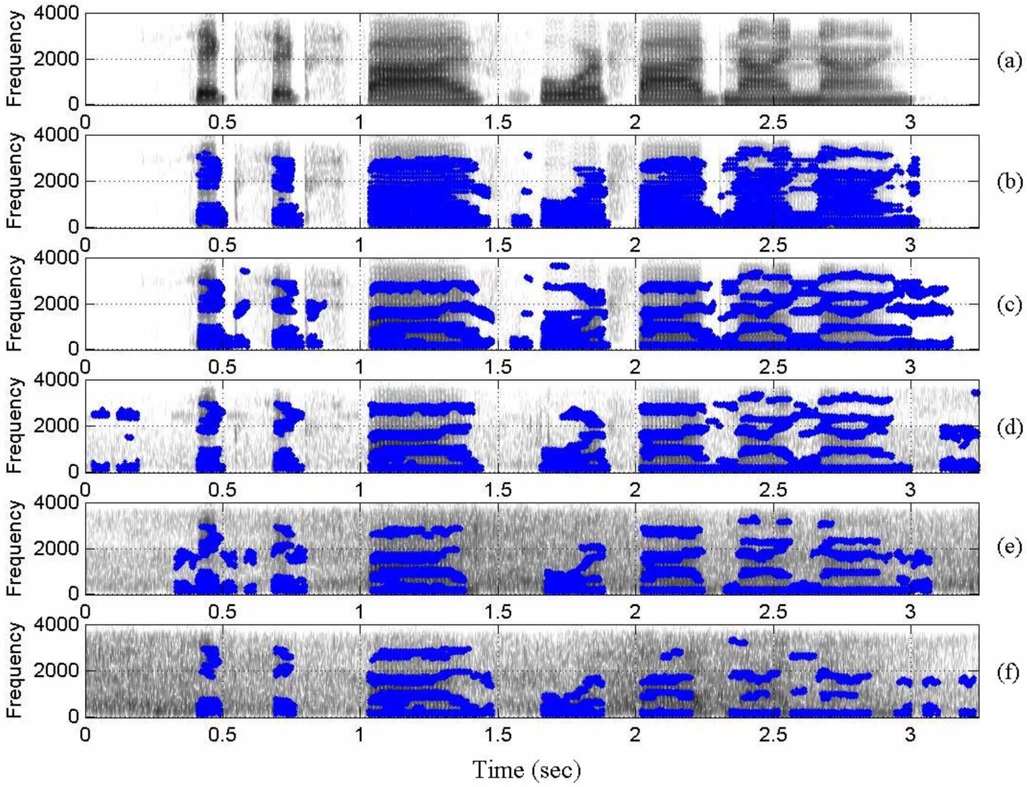

Figure 5.5: (a) Spectrogram of the utterance 'six six five four five nine nine' (b) the energy-based *maximal mask* (c-f) *MPO profile* at ∞, 20, 10 and 5 dB SNR respectively when the corrupting noise is street noise.

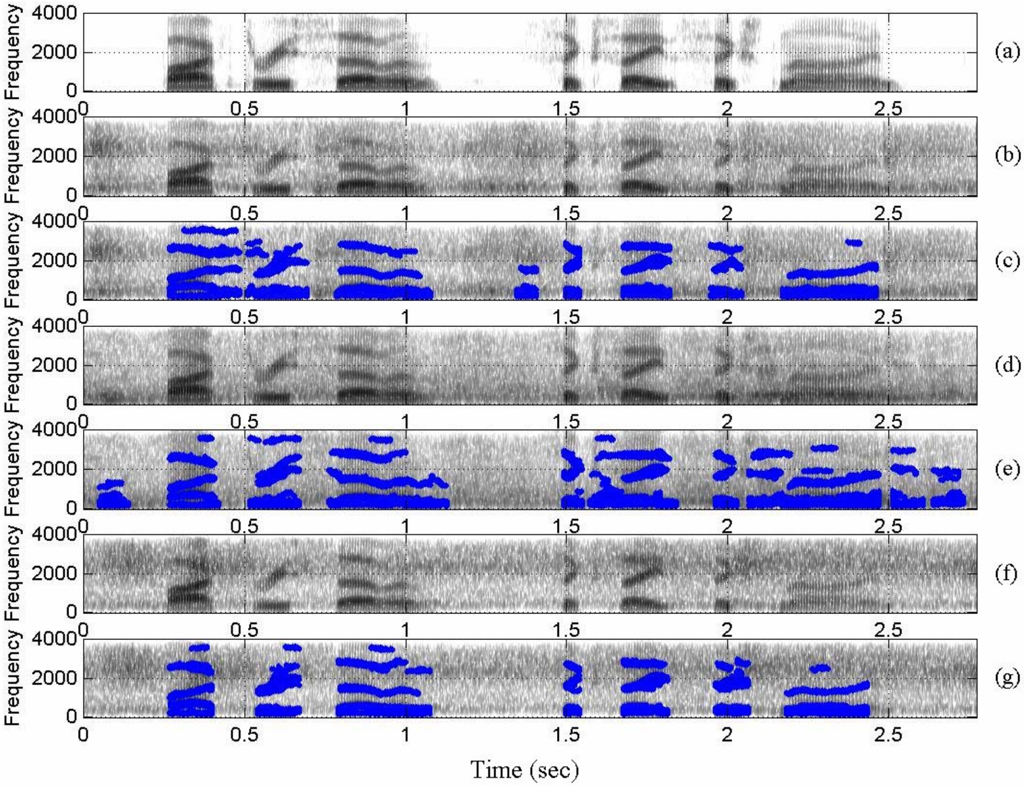

Figure 5.6: Spectrograms of the utterance 'five three seven six eight six nine' in clean (a) and when it is corrupted at 10 dB SNR by subway noise (b), restaurant noise (d) and subway noise (with MIRS filtering) (f). The corresponding *MPO profiles* are shown in (c), (e) and (f) respectively.

and thus has spectral characteristics very similar to that of speech signals. Thus, a lot more babble noise is passed as valid speech by the MPO processing (Fig. 5.7(c)) and leads to higher insertion rates (ref Table 5.2).

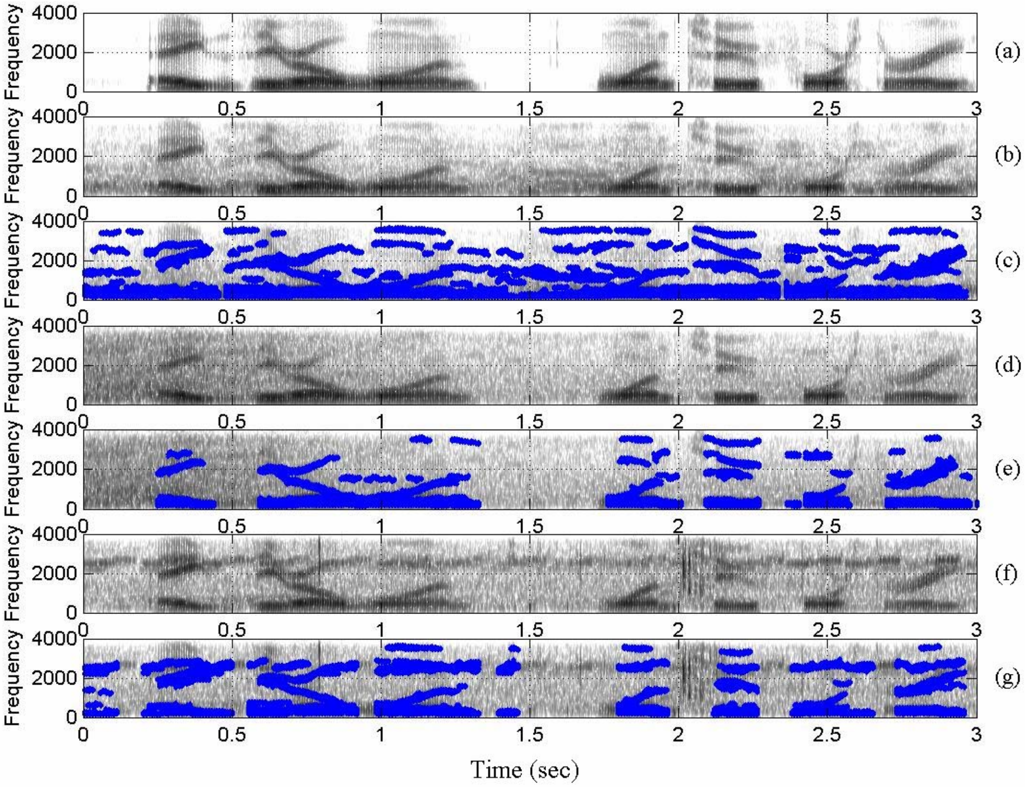

Figure 5.7: Spectrograms of the utterance ' eight zero one one two four three' in clean (a) and when it is corrupted at 10 dB SNR by babble noise (b), street noise (d) and street noise (with MIRS filtering) (f). The corresponding *MPO profiles* are shown in (c), (e) and (f) respectively.

5.3 Spectrogram displays

The binary mask based evaluations presented in the previous section can be used to evaluate the performance of only those speech enhancement techniques that split the speech signal into spectro-temporal units. In this section, the performance of the different speech enhancement techniques will be evaluated by inspecting the spectrograms. Spectrogram inspection, although not quantitatively rigorous, is a convenient tool to qualitatively analyze the nature of the speech distortion and of the residual noise.

Fig. 5.8 compares the spectrograms of the utterance 'five three seven six eight six' corrupted by subway noise at 10 dB and enhanced using the MMSE-STSA [38], GSS [33] and the proposed MPO speech enhancement technique. The spectrograms of the clean and the noisy unprocessed utterance are also shown for reference. The MMSE-STSA technique retains most of the speech signal but also passes a lot of noise (e.g. between 0 and 0.2 sec and 2-2.4 sec in Fig. 5.8(c)). The output of the GSS technique, on the other hand, contains little residual noise except for the band of energy just above 2000 Hz. However, a lot of high-frequency low-energy speech signal is suppressed. The output of the MPO enhancement technique retains the high-frequency low-energy speech signal (e.g. weak F3 information around 2500 Hz near 0.65 sec and again around 2700 Hz near 1.5 and 1.95 sec) that was suppressed by the GSS method but at the same time suppresses a lot of noise passed by the MMSE-STSA method. Thus, the MPO enhancement scheme strikes a better balance between the amount of speech signal retained and the amount of residual noise present in the enhanced speech signal. Notice that the residual noise in the MPO-enhanced output is narrowband and is relatively short in time. The residual noise is thus perceived as musical noise.

Fig. 5.9 compares the drop in performance of the three enhancement methods when the SNR is dropped to 0 dB. As expected, the amount of residual noise passed by the different methods increases. The MPO enhancement scheme is still able to retain more weak-amplitude speech information than the other two methods, but the short lax vowels (/I/ in 'six' around 1.5 sec and 2 sec) are suppressed.

Fig. 5.10 compares the performance of the different enhancement techniques when the corrupting noise is from a train station at 10 dB SNR. The utterance is 'eight four zero three zero five one'. Compared to the other two methods, the MPO enhancement method retains more speech signal while passing very little residual noise. Also notice that the MMSE-STSA method suppresses the low frequency harmonics of the speech signal. Such a behavior was observed in several of the MMSE-STSA-enhanced speech signals when the corrupting noise was from a train station. Fig. 5.11 compares the change in performance when the SNR is dropped to 0 dB. The performance of the MMSE-STSA and the GSS-based enhancement techniques deteriorates drastically while the MPO-enhancement scheme is able to retain the majority of the speech information with only a slight increase in the amount of residual noise.

Fig. 5.12 compares the performance of the different enhancement techniques when the corrupting noise is from an airport. The utterance is 'one seven five two oh four oh'. The airport noise consists of short bursts of narrowband signals.

The MPO enhancement scheme is designed to pass such narrowband signals and thus performs relatively poorly on suppressing the airport noise. The other two methods also pass a lot of noise as there is considerable overlap in the short-time spectra of the noise-only channels and speech channels. Fig. 5.13 compares the drop in performance when the SNR is dropped to 0 dB. Fig. 5.14 and 5.15 show another example of an utterance corrupted by airport noise at 10 dB and 0 dB SNR respectively. The narrowband noise is retained in the enhanced speech signal by all the three methods. This explains the high percentage of insertions in the *MPO profile* computed on speech utterances corrupted by airport noise (see Table 5.3).

In such cases, where the speech signal and the interfering noise type have a considerable overlap in the spectral domain, projecting the noisy speech signal in perceptually relevant higher dimensions (e.g. spectral and temporal modulation [78]) can help in achieving higher degree of separation.

5.4 Robustness to fluctuating noise

Some of the salient features of the MPO-based speech enhancement scheme are: (a) it makes minimal assumptions about the noise characteristics (the only assumption is that noise is broader than the harmonics of the speech signal), (b) it does not need to estimate the noise characteristics nor does it assume the noise satisfies any particular statistical model and (c) the noise removal performance on a given frame is independent of the performance on the adjoining frames. This scheme can thus be potentially robust when the level and the type of the background

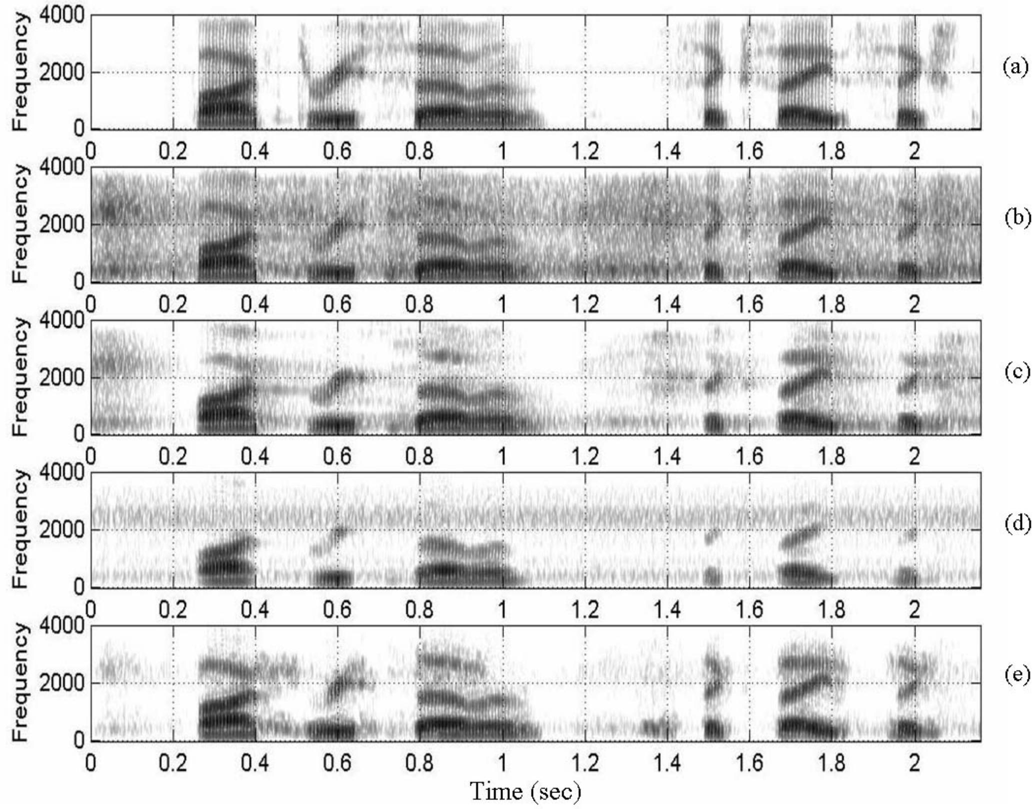

Figure 5.8: Spectrogram of (a) the clean speech signal ' five three seven six eight six' (b) the speech signal corrupted by subway noise at 10 dB SNR. (c) the speech signal enhanced using the MMSE-STSA technique (d) the speech signal enhanced using the GSS technique and (e) the speech signal enhanced using the proposed MPO technique

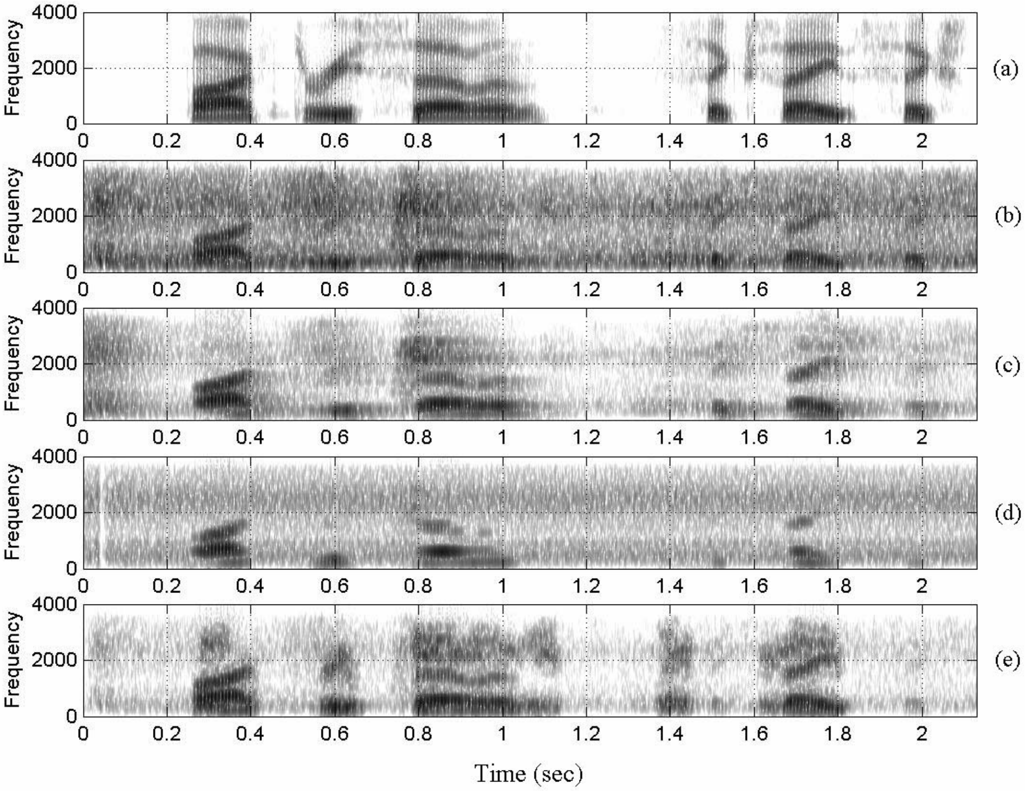

Figure 5.9: This figure compares the change in performance as the SNR drops from 10 dB (refer Fig. 5.8) to 0 dB. The speech signal and the noise type are the same as used in Fig. 5.8.

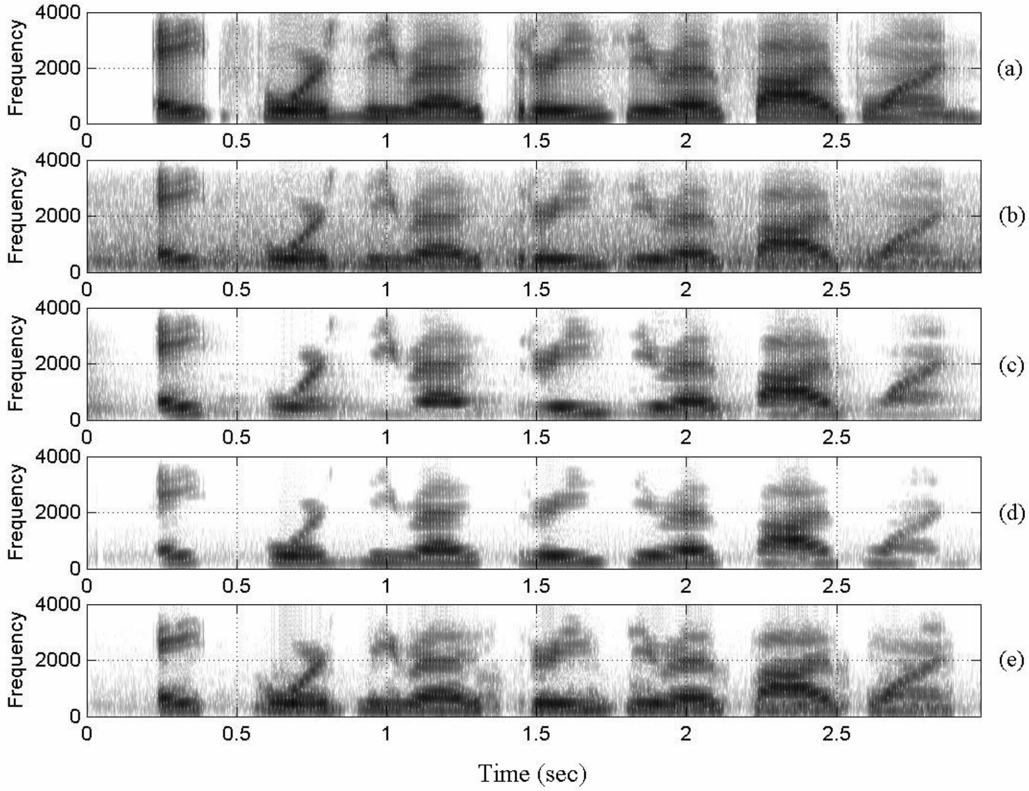

Figure 5.10: Spectrogram of (a) the clean speech signal ' eight four zero three zero five one' (b) the speech signal corrupted by train station noise at 10 dB SNR. (c) the speech signal enhanced using the MMSE-STSA technique (d) the speech signal enhanced using the GSS technique (e) the speech signal enhanced using the proposed MPO technique

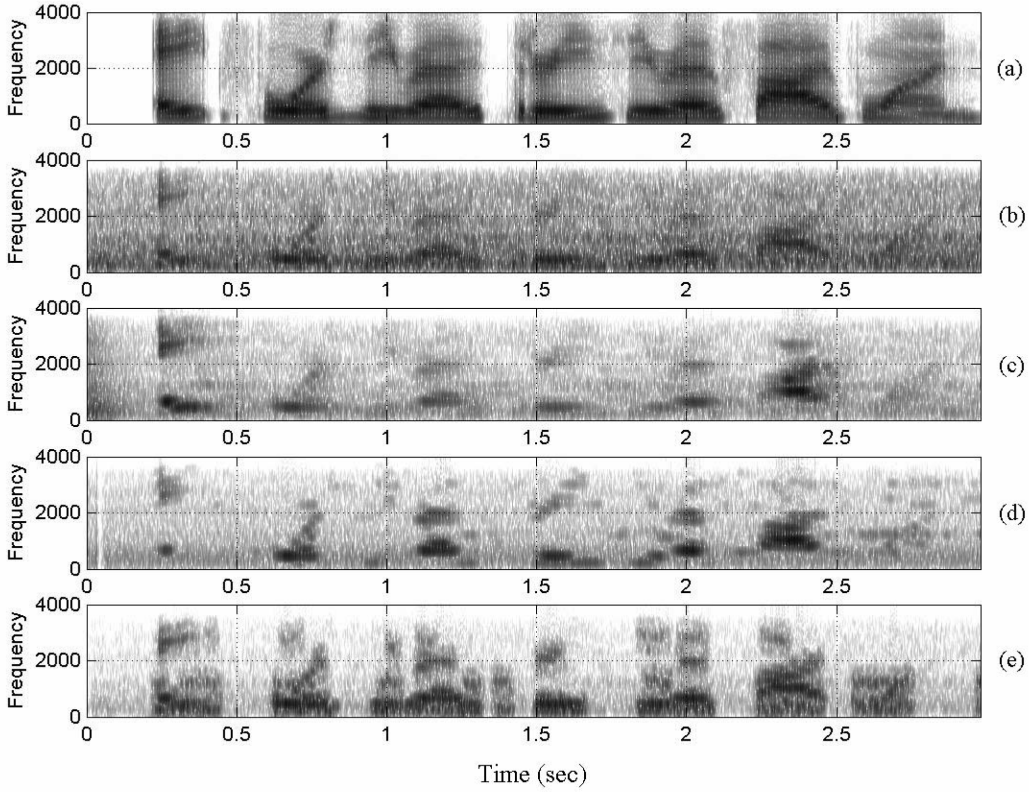

Figure 5.11: This figure compares the change in performance as the SNR drops from 10 dB (refer Fig. 5.10) to 0 dB. The speech signal and the noise type are the same as used in Fig. 5.10.

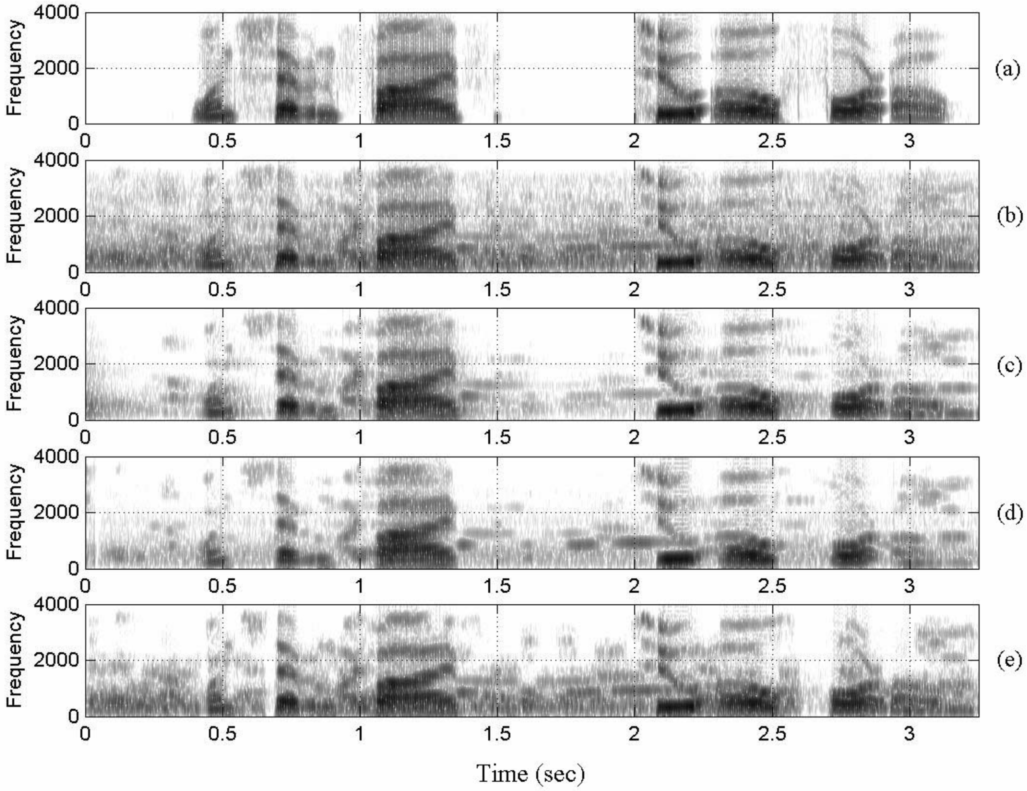

Figure 5.12: Spectrogram of (a) the clean speech signal ' one seven five two oh four oh' (b) the speech signal corrupted by airport noise at 10 dB SNR. (c) the speech signal enhanced using the MMSE-STSA technique (d) the speech signal enhanced using the GSS technique (e) the speech signal enhanced using the proposed MPO technique

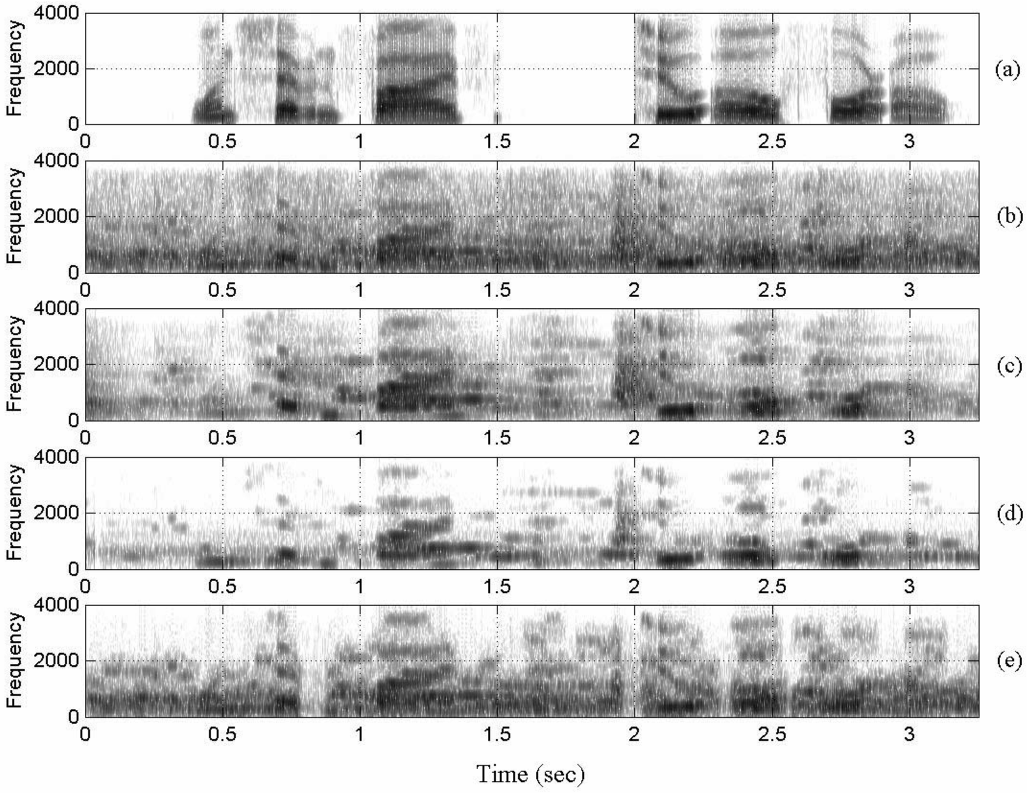

Figure 5.13: This figure compares the change in performance as the SNR drops from 10 dB (refer Fig. 5.12) to 0 dB. The speech signal and the noise type are the same as used in Fig. 5.12.

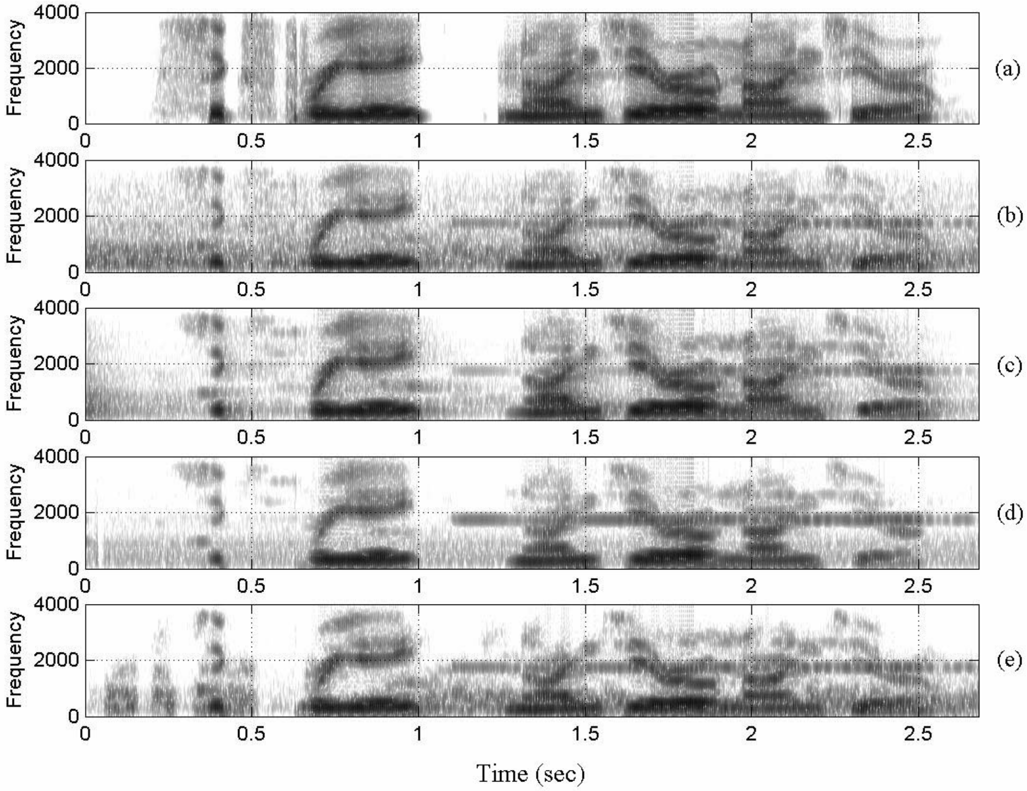

Figure 5.14: Spectrogram of (a) the clean speech signal ' six three eight nine zero nine zero' (b) the speech signal corrupted by airport noise at 10 dB SNR. (c) the speech signal enhanced using the MMSE-STSA technique (d) the speech signal enhanced using the GSS technique (e) the speech signal enhanced using the proposed MPO technique

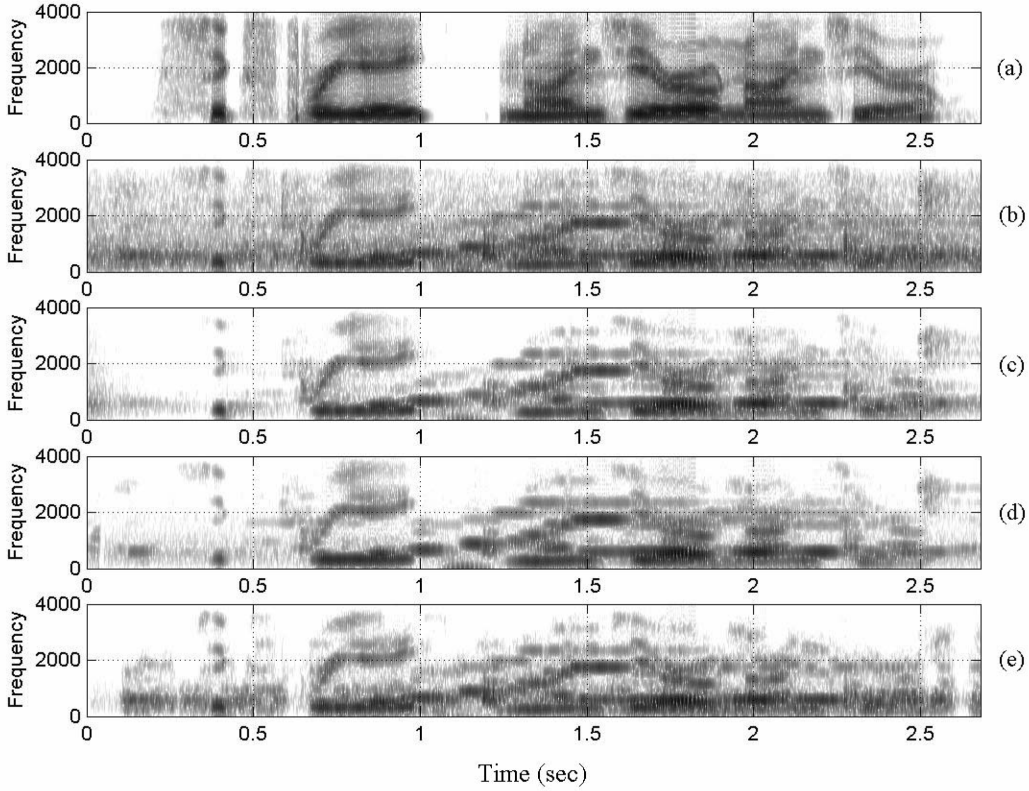

Figure 5.15: This figure compares the change in performance as the SNR drops from 10 dB (refer Fig. 5.14) to 0 dB

noise are fluctuating. To evaluate the performance on fluctuating noise, a speech utterance was formed by combining six different digits, corrupted either by subway noise, car noise or exhibition hall noise at widely varying SNRs. The digit sequence is 'nine four two eight five six' and the SNR sequence is 5, 20, 0, 15, -5, 10 dB. Fig. 5.16(a) shows the spectrogram of the clean signal. Fig. 5.16(b) shows the spectrogram of the MPO-processed clean signal. As is obvious from the figure, the MPO processing introduces little distortion when the input is clean speech. The major change is the reduction of energy in the obstruent regions. Fig. 5.16(c) shows the spectrogram of the noisy speech signal. Fig. 5.16(d-f) show the spectrograms of the speech signal enhanced using the log-MMSE-STSA method, the GSS method and the proposed MPO method respectively. The MPO method is able to retain most of the speech information while passing very little noise. For example, the transition of the weak F3 in 'four' (0.6-0.8 sec) is retained by the MPO method. The MPO method attenuates the noise in between the spectral peaks of 'five' (1.9-2.3 sec, local SNR -5 dB) while retaining the spectral peaks.

5.5 Objective evaluations

The quality of the speech signals enhanced using the MPO speech enhancement scheme was evaluated using four different objective quality measures and compared to that of speech signals using some of the other enhancement techniques proposed in the literature. One of these objective measures is based on SNR computation and has a relatively low degree of correlation with the subjective quality of the speech

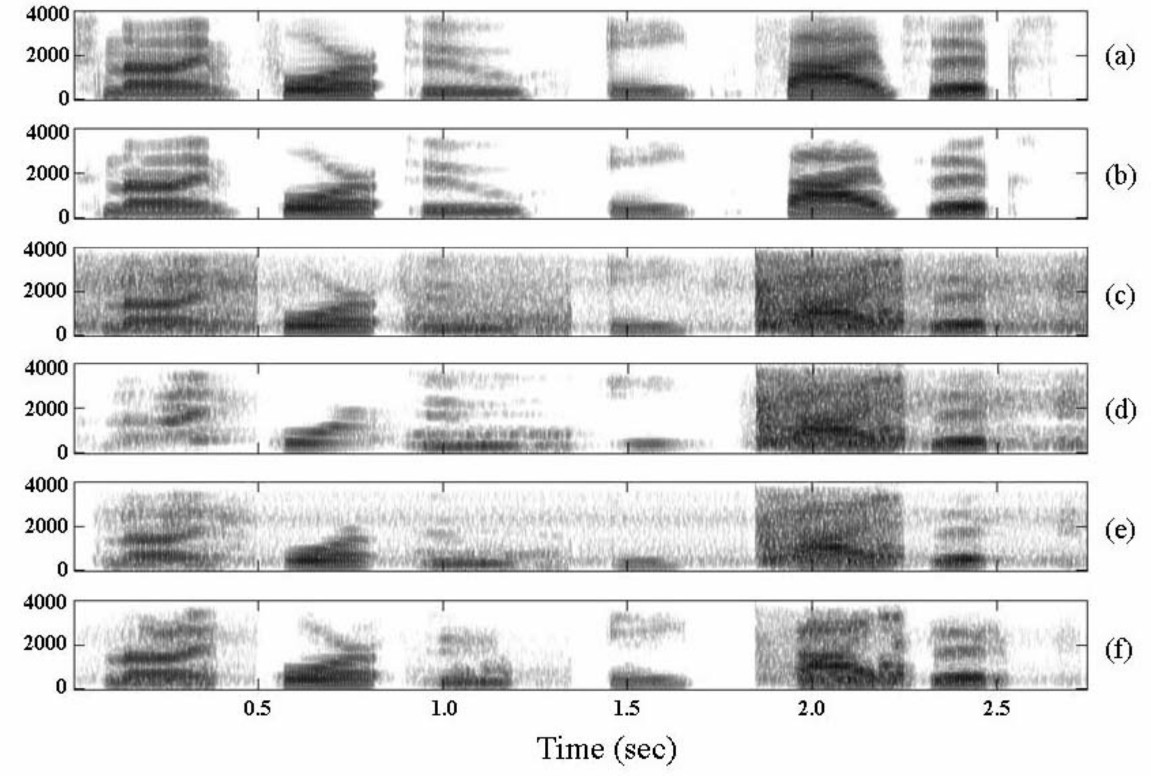

Figure 5.16: Spectrograms of (a) the clean speech signal; (b) the MPO-processed clean speech signal; (c) the utterance corrupted by fluctuating noise; (d) the speech signal enhanced using the MMSE-STSA method; (e) the speech signal enhanced using the GSS method and (f) the speech signal enhanced using the proposed MPO enhancement techinque.

signals. The other three measures are based on the computation of the Linear Predictive Coefficients (LPC) between the clean speech signal and the enhanced speech signals. These measures have a high degree of correlation with the subjective quality of the speech signals [102].

The SNR based measure is given by:

SNR improvement: The SNR improvement is expressed as the difference between the input and the output segmental SNR:

$$d_{SNR} = \frac{1}{M}\sum_{m=0}^{M-1} 10.log\frac{\frac{1}{N}\sum_{n=0}^{N-1} d^2(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}[s(n+Nm)-p(n+Nm)]^2} \tag{5.1}$$

where M is the total number of frames in the signal, N is the number of samples in a frame, $d(n)$ is the original corrupting noise, $s(n)$ is the clean speech signal and $p(n)$ is the noisy speech signal processed by an enhancement technique. Tables 5.5, 5.6 and 5.7 compare the SNR improvements obtained by the different enhancement schemes when evaluated on utterances from test subsets 'a', 'b' and 'c' respectively at different SNRs. The different techniques compared are: (a) MMSE-STSA [38] (b) logMMSE-STSA [41] (c) MMSE-STSA with non-causal SNR estimation [43] (d) GSS [33] (e) NSS [33] and (f) the proposed MPO enhancement scheme. Notice that the MPO speech enhancement scheme provides the highest SNR improvement in all of the three test scenarios at all the different SNRs except in test set 'b' and 'c' at the lowest SNR of -5 dB where it is slightly below some of the other methods. A negative SNR improvement implies that the combination of residual noise and speech distortion in the enhanced speech signal is more than the noise in the original

Table 5.5: SNR improvement (dB) obtained by the different enhancement schemes on 80 7-digit long utterances from test subset 'a'

	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	-11.348	-7.169	-3.582	-0.556	1.980	4.868
NC-MMSE	-11.423	-7.201	-3.604	-0.636	1.829	4.576
logMMSE	-11.271	-7.162	-3.570	-0.635	1.901	5.142
GSS	-11.230	-7.416	-4.023	-0.735	2.343	5.877
NSS	-15.859	-11.226	-6.789	-2.728	0.994	5.195
MPO	-3.888	-1.666	0.028	1.283	2.398	4.197

Table 5.6: SNR improvement (dB) obtained by the different enhancement schemes on 80 7-digit long utterances from test subset 'b'

	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	-11.925	-7.699	-3.992	-1.455	1.201	4.254
NC-MMSE	-11.988	-7.764	-4.066	-1.497	1.006	3.798
logMMSE	-11.868	-7.644	-3.935	-1.552	1.184	4.521
GSS	-11.752	-7.986	-4.407	-1.819	1.684	4.975
NSS	-16.398	-11.591	-6.992	-3.409	0.673	4.707
MPO	-4.333	-1.941	-0.340	0.619	1.544	2.790

Table 5.7: SNR improvement (dB) obtained by the different enhancement schemes on 40 7-digit long utterances from test subset 'c'

	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	-9.393	-5.521	-2.274	0.610	3.048	6.069
NC-MMSE	-9.550	-5.696	-2.584	0.546	2.763	6.011
logMMSE	-9.287	-5.382	-1.952	0.800	3.596	6.861
GSS	-10.415	-6.556	-2.442	0.518	4.399	7.629
NSS	-13.719	-9.090	-4.509	-0.580	3.486	7.582
MPO	-4.277	-1.618	0.595	2.397	4.465	7.383

speech signal. One of the main factors that contributes to speech distortion in an MPO-enhanced speech signal, especially at high SNRs, is the attenuation of the valley regions by the MPO processing (see Fig. 5.17(e)).

The SNR improvement measure has a poor correlation with the subjective quality of the enhanced processed signal but is a good indicator of amount of residual noise and speech distortion.

The three LPC based distortion measures are given by [103]:

1. *Itakura-Saito (IS) distortion measure:* The IS distortion measure between a frame of a clean speech signal and the corresponding frame of the enhanced speech signal is computed by the following equation:

$$d_{IS} = \left[\frac{\sigma_c^2}{\sigma_p^2}\right]\left[\frac{L_p R_c L_p^T}{L_c R_c L_c^T}\right] + log\left[\frac{\sigma_p^2}{\sigma_c^2}\right] - 1$$

where L_c and L_p are the LPC vectors for the clean frame and the processed frame respectively, σ_c^2 and σ_p^2 are the all-pole gains for the clean frame and the processed frame respectively and R_c is the autocorrelation matrix of the clean frame.

2. *Log-Area-Ratio (LAR) measure:* The LAR measure is computed using the P^{th} order LP reflection coefficients of the clean frame and the processed frame in the following way:

$$d_{LAR} = \left[\frac{1}{P} \sum_{j=1}^{P} \left[log \frac{1 + r_c(j)}{1 - r_c(j)} - log \frac{1 + r_p(j)}{1 - r_p(j)} \right]^2 \right]^{1/2}$$

   where r_c and r_p are the reflection coefficients of the clean frame and the processed frame respectively.

3. *Log-Likelihood Ratio (LLR) measure:* The LLR measure, unlike the IS measure, does not compare the all-pole gains of the clean frame and the processed frame and thus places more emphasis on the difference in the overall spectral envelopes of the two frames. The LLR measure is computed using the following equation:

$$d_{LLR} = log \left[\frac{L_p R_c L_p^T}{L_c R_c L_c^T} \right]$$

Tables 5.8, 5.9 and 5.10 compare the increase in the IS distortion measure at different SNRs for the output of different enhancement techniques when the input consists of utterances from test set 'a', 'b' and 'c' respectively. The corresponding values for LAR and LLR distortion measures are tabulated in Tables 5.11, 5.12, 5.13 and 5.14, 5.15, 5.16. All the three distortion measures have a value of 0 when the clean speech signal and the enhanced speech signal are exactly identical. Note

Table 5.8: Increase in the IS distortion measure for 80 7-digit long utterances from test subset ’a’

	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	0.353	0.597	1.138	2.001	3.473	4.157	5.812
NC-MMSE	0.285	0.820	2.199	4.690	18.747	37.943	73.424
logMMSE	0.721	1.416	3.848	5.776	14.839	12.584	15.527
GSS	0.959	3.446	3.967	3.993	3.010	2.210	2.251
NSS	0.161	0.490	1.865	5.460	22.418	37.977	52.740
MPO	3.056	0.566	0.751	1.157	1.522	3.624	7.764

that the values in the tables indicate the increase in the distortion values. For example, in Table 5.11 the LAR distortion measure between the output of the GSS enhancement technique at 10 dB SNR and the clean speech is the sum of 2.186 (LAR measure of GSS-processed clean speech) and 3.294 (corresponding increase in the LAR measure). For all the three measures, the distortion values between the MPO-processed clean speech and the clean speech are relatively high as the MPO processing attenuates the spectral valleys in the speech signal. This leads to an increase in the dissimilarities between the LP coefficients computed on clean speech and those computed on the MPO-processed clean speech.

Fig. 5.17 plots the framewise IS distortion measure for a MPO-processed clean speech signal and also compares the spectrograms of the clean signal and the MPO-processed signal. The corresponding *MPO profile* is also shown. Most of the perceptually salient information is maintained in the MPO-processed clean speech

Table 5.9: Increase in the IS distortion measure for 80 7-digit long utterances from test subset 'b'

	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	0.353	0.306	0.747	1.355	2.671	4.167	8.415
NC-MMSE	0.285	0.349	1.886	3.758	7.306	21.959	91.236
logMMSE	0.721	0.628	1.992	4.332	8.550	12.559	23.024
GSS	0.959	2.001	2.950	3.018	3.037	2.593	3.945
NSS	0.161	0.253	1.655	3.442	12.957	27.855	46.555
MPO	3.056	0.274	0.570	0.910	1.036	3.462	8.019

Table 5.10: Increase in the IS distortion measure for 40 7-digit long utterances from test subset 'c'

	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	0.180	1.076	1.728	1.977	4.618	6.203	15.047
NC-MMSE	0.150	2.307	7.820	7.624	73.671	83.500	437.176
logMMSE	0.517	2.964	4.583	5.274	15.896	18.123	42.284
GSS	1.129	3.581	3.580	3.022	2.535	2.162	3.764
NSS	0.100	1.130	2.482	4.225	23.469	60.782	127.618
MPO	3.272	0.599	0.880	0.875	2.285	5.182	11.488

Table 5.11: Increase in the LAR distortion measure for 80 7-digit long utterances from test subset ’a’

	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	0.923	1.656	2.476	3.446	4.549	5.489	6.442
NC-MMSE	0.760	1.516	2.406	3.559	4.987	6.386	7.506
logMMSE	1.089	1.913	2.868	3.880	4.920	5.747	6.565
GSS	2.186	2.156	2.736	3.294	3.905	4.654	5.380
NSS	1.508	1.190	2.020	3.096	4.651	5.876	6.717
MPO	3.164	0.830	1.333	1.937	2.803	3.577	4.276

while the spectral valleys are attenuated leading to a higher IS distortion measure. Fig. 5.17(e-g) compares a spectral slice of the original clean speech signal, GSS-processed clean speech signal and MPO-processed clean speech signal respectively at a frame centered at 1.18 sec. MPO processing maintains all the strong harmonics but, unlike GSS processing, the weak harmonics in the valley region are greatly attenuated (around 1000 Hz and around 3000 Hz). As a result, the IS distortion measure of this frame for the MPO-processed signal is 1.17 whereas the corresponding value for GSS-processed signal is 0.05.

The MPO enhancement scheme leads to the lowest increase in the LAR measures for all the different noise types at all the different SNRs. The LLR and the IS distortion measures from the MPO-enhanced speech signals show the lowest increase in most of the cases although there are a few instances where the increase in the LLR and IS distortion values from MPO-enhanced speech signals are slightly more

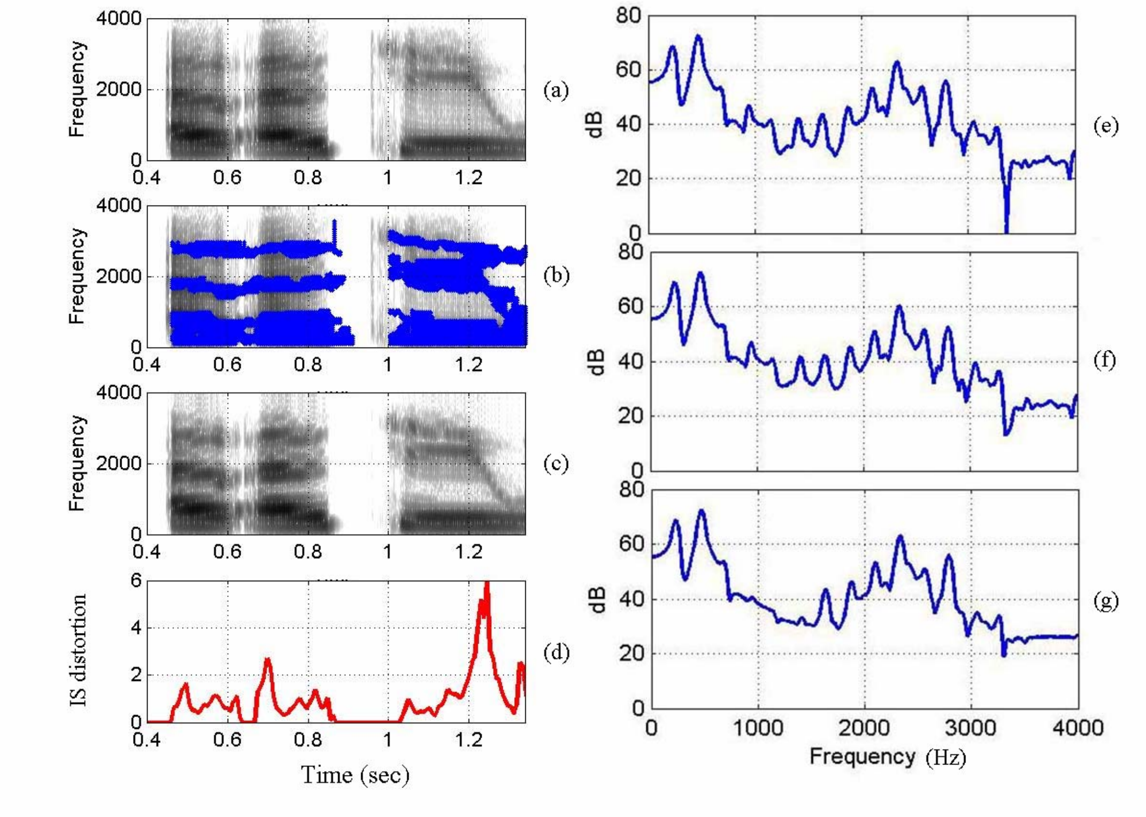

Figure 5.17: (a) Spectrogram of the utterance 'oh oh two' in clean; (b) corresponding *MPO profile*; (c) Spectrogram of the MPO-processed clean utterance; (d) framewise IS distortion measure; (e) spectral slice of a frame of the clean speech signal centered at 1.18 sec; (f) spectral slice of the corresponding frame of the GSS-processed clean speech signal (g) spectral slice of the corresponding frame of the MPO-processed clean speech signal

Table 5.12: Increase in the LAR distortion measure for 80 7-digit long utterances from test subset 'b'

	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	0.923	1.204	1.953	2.709	3.896	4.815	5.698
NC-MMSE	0.760	1.113	1.883	2.708	4.060	5.392	6.892
logMMSE	1.089	1.352	2.185	3.054	4.303	5.177	5.889
GSS	2.186	1.427	2.173	2.601	3.496	4.039	4.852
NSS	1.508	0.920	1.634	2.513	3.815	4.997	5.989
MPO	3.164	0.668	1.121	1.594	2.435	3.140	3.938

Table 5.13: Increase in the LAR distortion measure for 40 7-digit long utterances from test subset 'c'

	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	0.659	2.071	2.965	3.760	5.004	5.877	6.904
NC-MMSE	0.530	1.948	2.986	3.956	5.729	6.615	7.853
logMMSE	0.837	2.473	3.372	4.094	5.358	5.968	6.925
GSS	2.322	2.198	2.832	3.118	4.098	4.664	5.544
NSS	0.846	1.764	2.633	3.598	5.199	6.272	7.219
MPO	3.259	0.834	1.536	2.036	2.957	3.545	4.367

Table 5.14: Increase in the LLR distortion measure for 80 7-digit long utterances from test subset 'a'

	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	0.071	0.211	0.351	0.551	0.855	1.235	1.696
NC-MMSE	0.054	0.205	0.360	0.579	0.923	1.350	1.823
logMMSE	0.094	0.243	0.399	0.601	0.912	1.284	1.743
GSS	0.116	0.318	0.456	0.635	0.898	1.269	1.670
NSS	0.117	0.183	0.354	0.605	1.088	1.589	2.048
MPO	0.425	0.181	0.281	0.430	0.657	0.929	1.279

Table 5.15: Increase in the LLR distortion measure for 80 7-digit long utterances from test subset 'b'

	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	0.071	0.144	0.273	0.419	0.694	0.997	1.330
NC-MMSE	0.054	0.130	0.259	0.422	0.702	1.071	1.534
logMMSE	0.094	0.160	0.299	0.469	0.774	1.049	1.368
GSS	0.116	0.208	0.380	0.491	0.767	1.039	1.417
NSS	0.117	0.138	0.286	0.501	0.889	1.282	1.670
MPO	0.425	0.138	0.241	0.350	0.528	0.762	1.013

Table 5.16: Increase in the LLR distortion measure for 40 7-digit long utterances from test subset 'c'

	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
MMSE	0.045	0.247	0.406	0.583	0.921	1.263	1.794
NC-MMSE	0.034	0.258	0.445	0.637	1.049	1.353	1.906
logMMSE	0.072	0.308	0.463	0.613	0.942	1.252	1.815
GSS	0.138	0.335	0.506	0.621	1.040	1.320	1.894
NSS	0.043	0.234	0.391	0.612	1.077	1.533	2.103
MPO	0.461	0.183	0.312	0.408	0.669	0.932	1.347

than that for the GSS-enhanced speech signals.

The variation in performance across the different noise types is compared in Tables 5.17, 5.18 and 5.19 for MMSE-STSA, GSS and MPO based speech enhancement techniques, respectively. None of the enhancement schemes seem to favor any particular noise type over the other, although all the three methods are most affected by the subway noise and the exhibition hall noise. A similar trend was observed for distortion values using the other two objective measures.

5.6 Subjective evaluations

5.6.1 Experimental setup

The perceptual quality of the speech signals enhanced by the different techniques was evaluated by listeners using the two-alternative forced-choice preference

Table 5.17: Performance variation across the different noise types for MMSE-STSA based enhancement technique. The entries indicate the increase in the LLR distortion values.

noise	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
subway	0.066	0.267	0.437	0.661	0.969	1.365	2.036
babble	0.035	0.135	0.265	0.499	0.701	1.042	1.439
car	0.065	0.211	0.334	0.530	0.851	1.104	1.366
exhibit hall	0.115	0.232	0.366	0.514	0.897	1.427	1.945
restaurant	0.066	0.136	0.237	0.397	0.650	0.962	1.380
street	0.035	0.210	0.390	0.531	0.913	1.161	1.598
airport	0.065	0.113	0.261	0.406	0.600	0.984	1.190
train station	0.115	0.118	0.204	0.344	0.612	0.881	1.153

Table 5.18: Performance variation across the different noise types for GSS based enhancement technique. The entries indicate the increase in the LLR distortion values.

noise	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
subway	0.134	0.404	0.524	0.751	1.044	1.526	1.973
babble	0.134	0.187	0.329	0.468	0.688	1.052	1.397
car	0.107	0.297	0.423	0.609	0.864	1.099	1.481
exhibit hall	0.089	0.383	0.548	0.710	0.997	1.399	1.827
restaurant	0.134	0.187	0.352	0.450	0.674	0.970	1.322
street	0.134	0.250	0.521	0.625	0.952	1.313	1.741
airport	0.107	0.144	0.276	0.427	0.691	0.907	1.242
train station	0.089	0.249	0.371	0.464	0.753	0.967	1.365

Table 5.19: Performance variation across the different noise types for the proposed MPO based enhancement technique. The entries indicate the increase in the LLR distortion values.

noise	clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
subway	0.452	0.240	0.350	0.516	0.777	1.009	1.467
babble	0.447	0.123	0.216	0.321	0.508	0.729	0.979
car	0.371	0.147	0.265	0.405	0.627	0.868	1.102
exhibit hall	0.432	0.213	0.292	0.479	0.715	1.111	1.566
restaurant	0.452	0.138	0.230	0.354	0.492	0.708	0.963
street	0.447	0.173	0.274	0.372	0.573	0.821	1.223
airport	0.371	0.104	0.230	0.350	0.523	0.786	0.948
train station	0.432	0.138	0.230	0.324	0.523	0.733	0.917

tasks. All the six listeners, three males and three females, had American English as their first language and were screened for hearing loss. All the listeners had hearing thresholds at or below 20 dB in the frequency range 500-4000 Hz. The speech signals used for these perceptual tests consisted of 5-digit long utterances corrupted by either (a) subway noise or (b) car noise at SNRs ∞, 15, 5 or -5 dB, and enhanced using either the proposed MPO enhancement scheme or one of the following three techniques: (1) logMMSE-STSA (2) power spectral subtraction (3) Wiener filtering [104]. The corresponding unprocessed clean and noisy utterances were also used. The three techniques used here are representative techniques for speech enhancement using statistical methods and speech enhancement using signal-theoretic methods. The Wiener filtering method could not be evaluated using the objective distortion measures mentioned in section 5.5 as it introduces certain time-delay in the enhanced speech signals (which varies slightly for different utterances). The objective measures are reliable only when they are computed from the same temporal frame on both the original speech signal and the enhanced speech signal.

The listeners were divided in groups of three. Listeners in the same group were presented with the same set of 180 paired utterances. The set of utterances and the type of corrupting noise was changed with the group. The contents of the two utterances in a pair were always the same, but the processing technique was different. The following different combinations were used in random order:

1. unprocessed clean vs. unprocessed 0 dB noisy (control task)
2. unprocessed clean vs. MPO-processed clean
3. unprocessed 15dB noisy vs. MPO-enhanced 15 dB
4. unprocessed 5dB noisy vs. MPO-enhanced 5 dB
5. unprocessed 0dB noisy vs. MPO-enhanced 0 dB
6. unprocessed -5dB noisy vs. MPO-enhanced -5 dB
7. MPO-enhanced 15dB noisy vs. logMMSE-STSA-enhanced 15 dB
8. MPO-enhanced 5dB noisy vs. logMMSE-STSA-enhanced 5 dB
9. MPO-enhanced 0dB noisy vs. logMMSE-STSA-enhanced 0 dB
10. MPO-enhanced -5dB noisy vs. logMMSE-STSA-enhanced -5 dB
11. Power SS-enhanced 15dB noisy vs. MPO-enhanced 15 dB
12. Power SS-enhanced 5dB noisy vs. MPO-enhanced 5 dB
13. Power SS-enhanced 0dB noisy vs. MPO-enhanced 0 dB
14. Power SS-enhanced -5dB noisy vs. MPO-enhanced -5 dB
15. MPO-enhanced 15dB noisy vs. Wiener-filtering-enhanced 15 dB
16. MPO-enhanced 5dB noisy vs. Wiener-filtering-enhanced 5 dB
17. MPO-enhanced 0dB noisy vs. Wiener-filtering-enhanced 0 dB
18. MPO-enhanced -5dB noisy vs. Wiener-filtering-enhanced -5 dB

The first combination was used as a control task to test the attention of the listeners. It is expected that the listeners would always prefer the unprocessed-clean signal over the unprocessed-0dB-noisy signal. Each combination had 10 different utterances and each pair was presented twice. The order of the utterances in a pair were reversed the second time. For example, if the order in the first round was: 'unprocessed clean vs. MPO-processed clean' then the order in the second round was: 'MPO-processed clean vs. unprocessed clean'. All the tests were conducted in an acoustically isolated chamber and the utterances were presented binaurally through high-quality Sony MDR-7509 headphones. Listeners were asked to note their preference for the first or the second utterance based on the overall quality and ease of listening. The listeners were also asked to note the strength of their preference: (a) strong (b) moderate or (c) weak preference. All the listeners were presented an initial trial set of pairs to familiarize them with the task. The results on these trial set were not used in the final evaluations. Listeners' preferences were recorded using a Graphical User Interface (GUI) developed in Matlab. Fig. 5.18 shows the GUI.

5.6.2 Results

The outcome of each paired test is given a numerical weight of 1 if the preference was weak, 2 if the preference was moderate and 3 if the preference was strong. The score is positive if the MPO-processed output was preferred, otherwise it is negative. For example, the overall score is +3 if the MPO-enhanced signal is

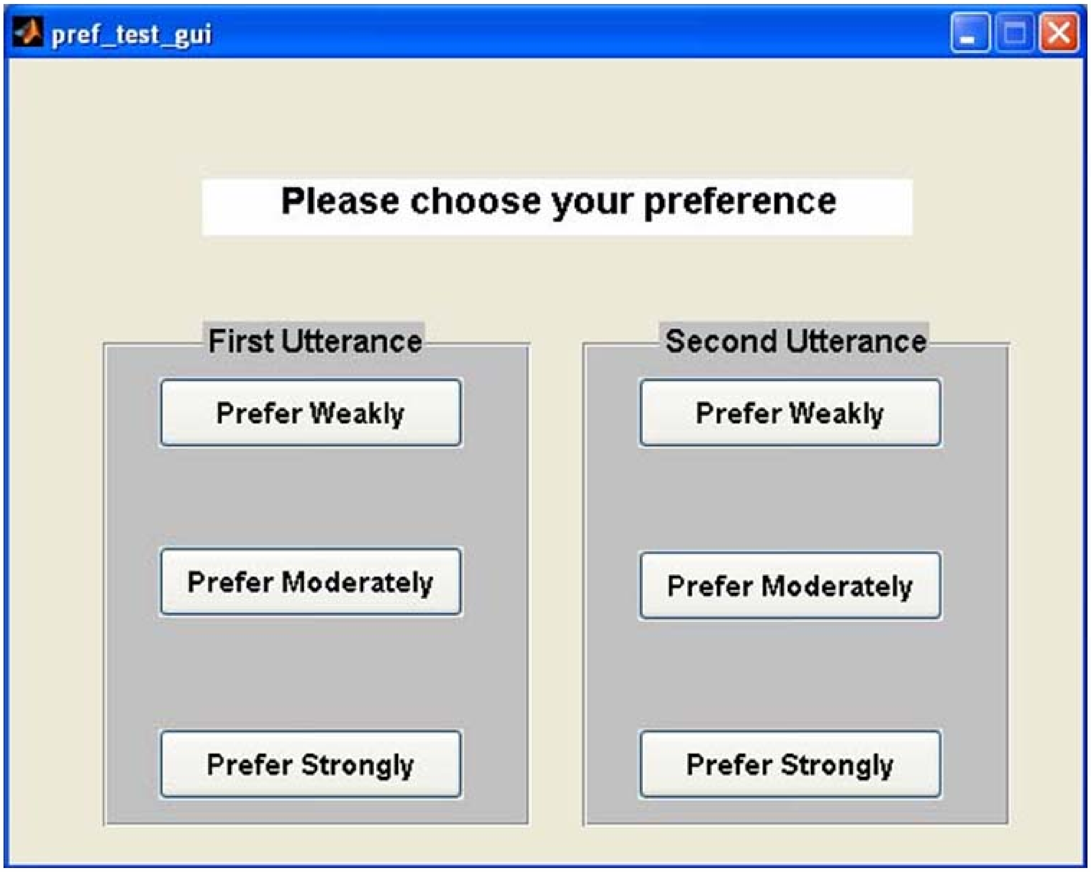

Figure 5.18: Graphical User Interface used for subjective evaluations.

strongly preferred over the logMMSE-STSA-enhanced signal and the overall score is -3 if logMMSE-STSA-enhanced signal is strongly preferred over the MPO-enhanced signal.

Charts in Figs. 5.19, 5.20, 5.21 and 5.22, 5.23, 5.24 show the preferences of each of the three listeners in the first and the second group across different SNRs. The corrupting noise is subway noise for group 1 and car noise for group 2. A positive score indicates that the output of the MPO method was preferred over that of the other method whereas a negative score indicates that the other method was preferred over the MPO method. There is considerable variance in preferences across the listeners in a group as well as across the different SNRs for a given listener. Listener 1 in group 1 weakly prefers the output of the proposed MPO enhancement scheme over the other three enhancement schemes at -5, 0, 5 dB SNR but prefers (very weakly) the logMMSE-STSA and Wiener-filtering methods at 15 dB SNR. Listener 2 in the same group weakly prefers the MPO technique over the Wiener-filtering technique at 15 dB SNR but prefers (very weakly) the Wiener filtering technique at -5 dB SNR. Listener 3 in the same group prefers (very weakly) the unprocessed signal over the MPO technique at all the SNRs but the preference among the different enhancement schemes varies as the SNR is varied.

Listeners 1 and 2 in group 2, where the corrupting noise is car noise, consistently prefer (weakly) all the other three enhancement schemes over the proposed MPO enhancement scheme at all the SNRs but have differing preferences between unprocessed noisy speech signal and the output of the MPO enhancement scheme. The preferences for listener 3 in group 2 are less consistent. To estimate the variability in preferences across the speakers, standard deviation values as well as the difference in the maximum and minimum score were computed for each combination pair across the three speakers. The difference in the maximum and the minimum score can at most be 6 $[3 - (-3) = 6]$ and indicates a case where two listeners had the two extreme preferences. For the given set of results, the difference in the maximum and the minimum score had a value of 6 about 6% of the times and a value of 5 or more about 33% of the times. The standard deviation can have a maximum value of 3.1 (when the scores are [3 3 3 -3 -3 -3]). For the given set of results, the standard deviation had a value of more than 1.6 about 60% of the times.

The results from all the listeners in a group were collapsed across the three degrees of preferences to compute the percentage of the times the MPO enhancement scheme was preferred over the other enhancement scheme. These values are tabulated in Table 5.20 and 5.21 for group 1 and group 2, respectively. In general, the output of the MPO enhancement scheme is preferred over the other techniques when the speech signals are corrupted by subway noise whereas the output of the MPO enhancement scheme is not preferred when the speech signals are corrupted by car noise.

All the six listeners had consistent preferences when the combination was: 'unprocessed clean vs. MPO-processed clean'. Each of the listeners preferred the MPO-processed clean speech signal just about as many times as (s)he preferred the unprocessed clean signal. This indicates that the MPO processing introduces minimal perceptual artifacts in clean speech.

Similar tests need to be conducted on a larger population of listeners to draw

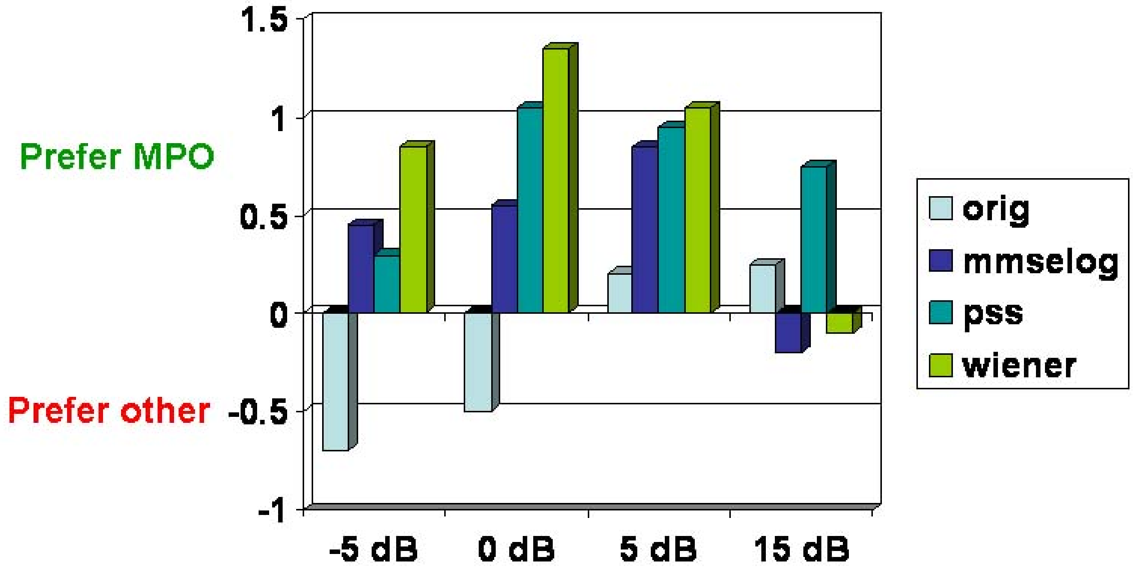

Figure 5.19: Preference chart for listener 1 in group 1. The corrupting noise is subway noise.

reliable conclusions about patterns of preferences across different noise types at different SNRs.

5.7 Robust speech recognition results

The proposed MPO speech enhancement scheme was used as a preprocessor for a robust speech recognition system. As mentioned earlier, two different databases were used for these experiments: the GRID database and the Aurora database.

5.7.1 Recognition results on the GRID database

The performance of the MPO speech enhancement technique was evaluated on the GRID database as part of the *Speech Separation Challenge* to be held as

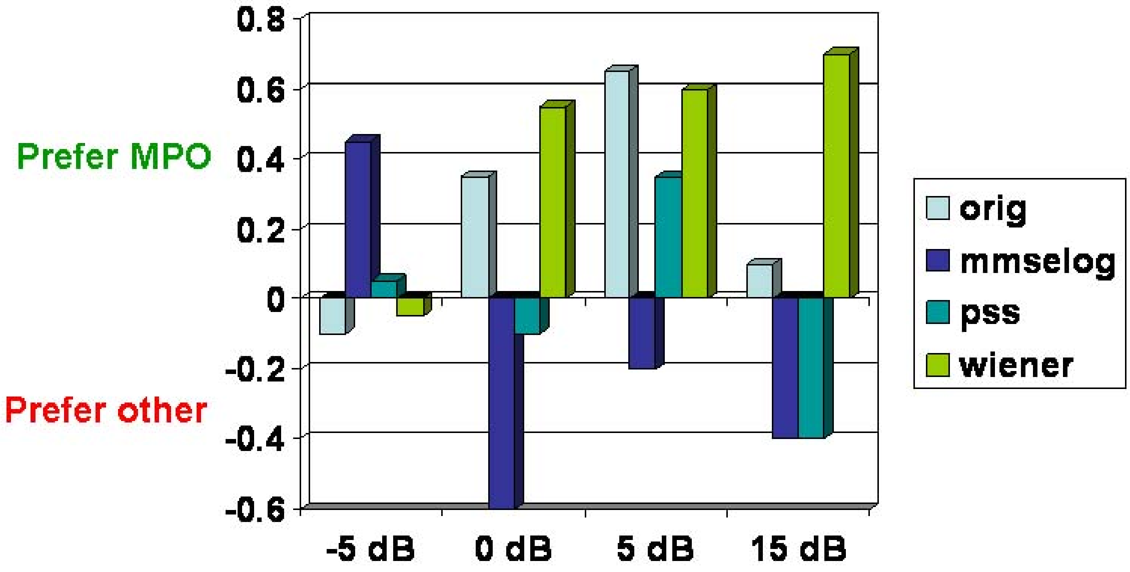

Figure 5.20: Preference chart for listener 2 in group 1. The corrupting noise is subway noise.

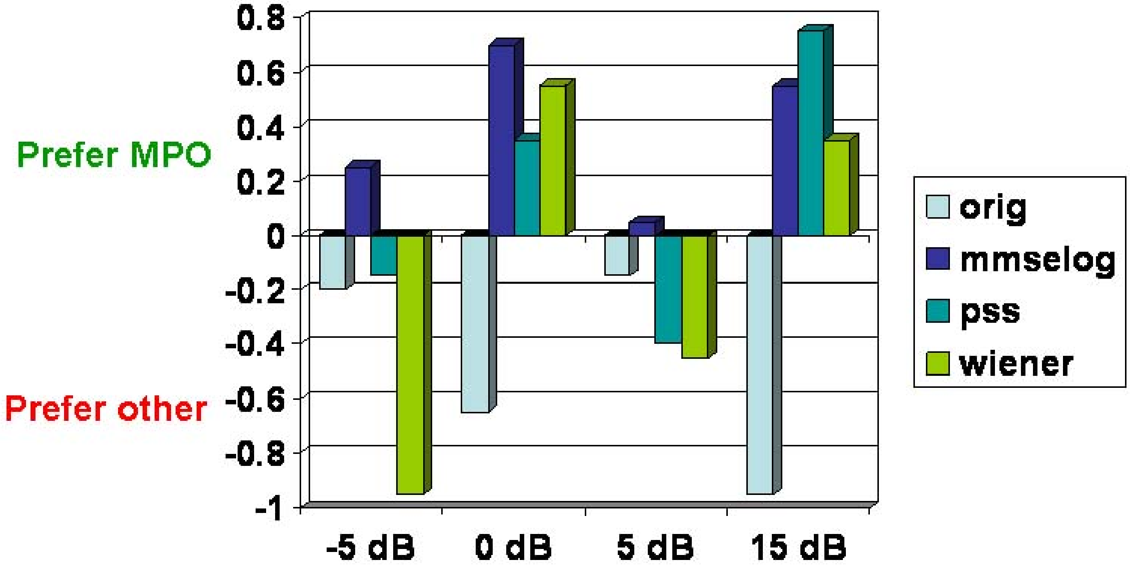

Figure 5.21: Preference chart for listener 3 in group 1. The corrupting noise is subway noise.

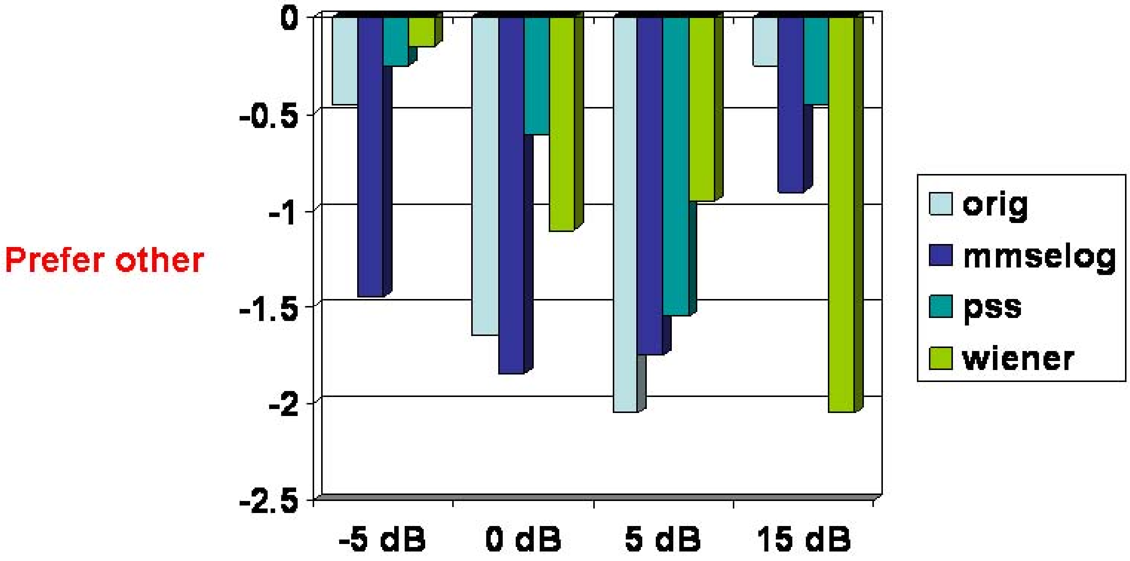

Figure 5.22: Preference chart for listener 1 in group 2. The corrupting noise is car noise.

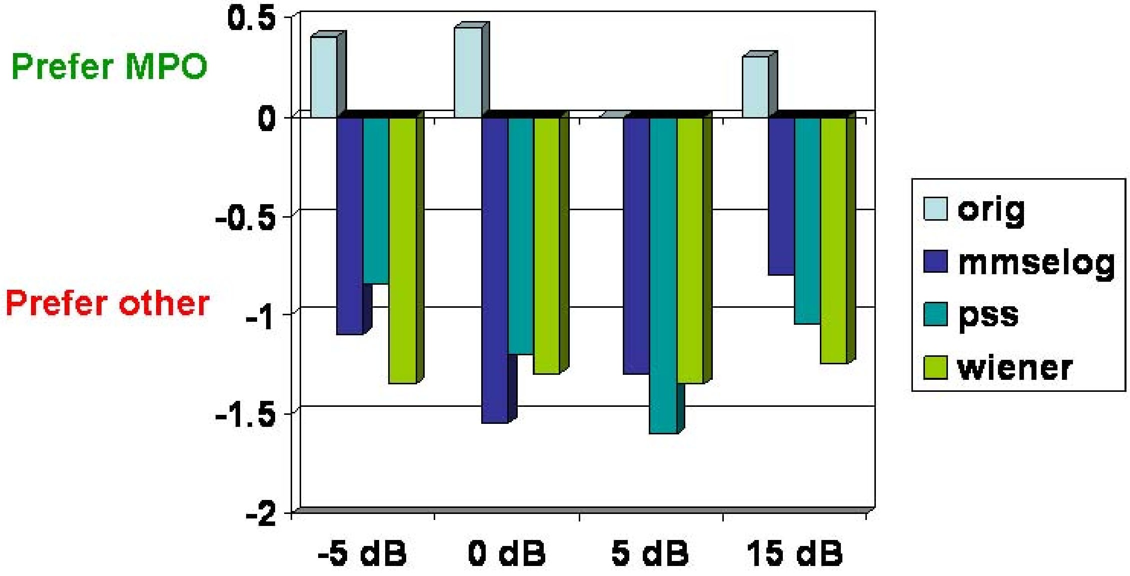

Figure 5.23: Preference chart for listener 2 in group 2. The corrupting noise is car noise.

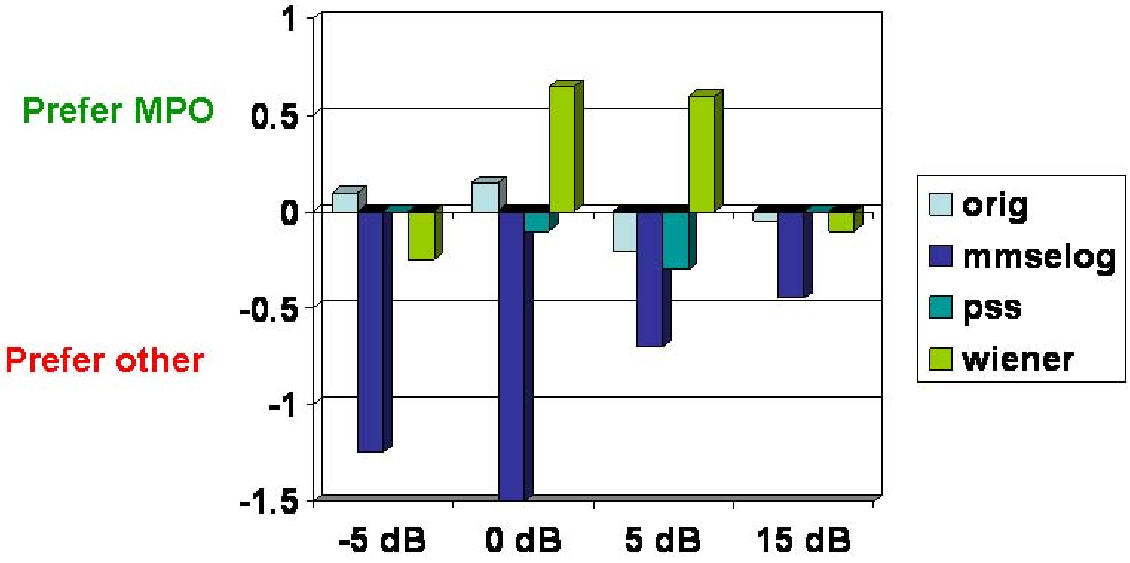

Figure 5.24: Preference chart for listener 3 in group 2. The corrupting noise is car noise.

Table 5.20: Percentage of the times the output of the proposed MPO enhancement scheme is preferred over the other enhancement methods or the unprocessed speech signal when the speech signals are corrupted by subway noise

.

	-5 dB	0 dB	5 dB	15 dB
unprocessed	40.00	36.66	55.00	46.66
logMMSE-STSA	63.33	58.33	63.33	50.00
Power SS	48.33	65.00	58.33	55.00
Wiener	71.66	78.33	71.66	55.00

Table 5.21: Percentage of the times the output of the proposed MPO enhancement scheme is preferred over the other enhancement methods or the unprocessed speech signal when the speech signals are corrupted by car noise

	-5 dB	0 dB	5 dB	15 dB
unprocessed	51.66	48.33	31.66	48.33
logMMSE-STSA	11.66	10.00	18.33	25.00
Power SS	40.00	31.66	20.00	31.66
Wiener	33.33	31.66	33.33	18.33

a special session at the 2006 International Conference on Spoken Language Processing [105]. The speech recognizer was trained on clean speech and was tested on speech corrupted by speech shaped noise or competing speech at various SNRs. The thresholds for the MPO processing were developed using the two extremes of strictly narrowband signals and white noise (refer Section 3.3). The thresholds were not retrained nor was the MPO enhancement scheme tailored in any way to suit the GRID database. The present version of the MPO enhancement scheme consists of MPO structures placed at regular frequency spacing from 100 Hz to just below 4000 Hz. The GRID database is sampled at 25 kHz and thus has relevant information till about 12.5 kHz. The *MPO profile* is computed till about 4 kHz and can be adjusted to the higher frequencies in one of the three ways: (a) downsample the database to 8 kHz and apply the *MPO profile* as-is, (b) apply the *MPO profile* from 0–4 kHz and pass the high frequency information as-is (i.e. set the *MPO profile* to one for all spectro-temporal channels with CF $>$ 4 kHz) or (c) apply the *MPO profile* from 0–4 kHz and suppress the high frequency information (i.e. set the *MPO profile* to zero for all spectro-temporal channels with CF $>$ 4 kHz). These three different methods are referred to as MPO$_{4k}$, MPO$_{hon}$ and MPO$_{hoff}$ respectively and results are presented for each of these methods as well as for the 'no-processing' case where the noisy test data is used without any processing. To minimize mismatch in the training and the testing conditions, the training utterances and the testing utterances use the same technique to extend the *MPO profile* to higher frequencies.

The experiments were conducted using the baseline recognizer provided with the database. The recognizer is based on the widely used Hidden Markov Model Toolkit (HTK) [108]. The speech signal is parameterized into 12 Mel-cepstral coefficients and the energy along with the first and the second-order derivatives to form a 39-dimensional parameter vector. Each word in the dictionary is modeled as a whole-word HMM with a left-to-right state-transition topology with no skips allowed over the states. The output of each state is modeled as a mixture of 32 Gaussians with diagonal covariance matrices. The number of states for each word is based on the phoneme-length of the word and varies from 4, for short words like 'at', 'one', to 10 states for long words like 'seven'. The grammar is modeled so that only the valid structured sentences (refer Section 5.1) are permitted.

Table 5.22 shows the recognition accuracy when the speech signals in the test subset are corrupted by speech shaped noise at various SNRs and enhanced using one of the three different extensions of the MPO speech enhancement scheme mentioned above. In all the cases, the recognizer is trained using only the MPO-processed clean speech. The row corresponding to 'no-processing' shows the baseline results obtained

Table 5.22: Recognition accuracy for speech-shaped-noise condition

Type	clean	6 dB	0 dB	-6 dB	-12 dB
No processing	98.56	56.67	18.94	11.78	11.67
MPO$_{hon}$	97.89	73.67	40.67	19.11	13.28
MPO$_{hoff}$	96.44	71.06	41.94	18.72	14.50
MPO$_{4k}$	96.00	73.83	50.06	26.00	14.33

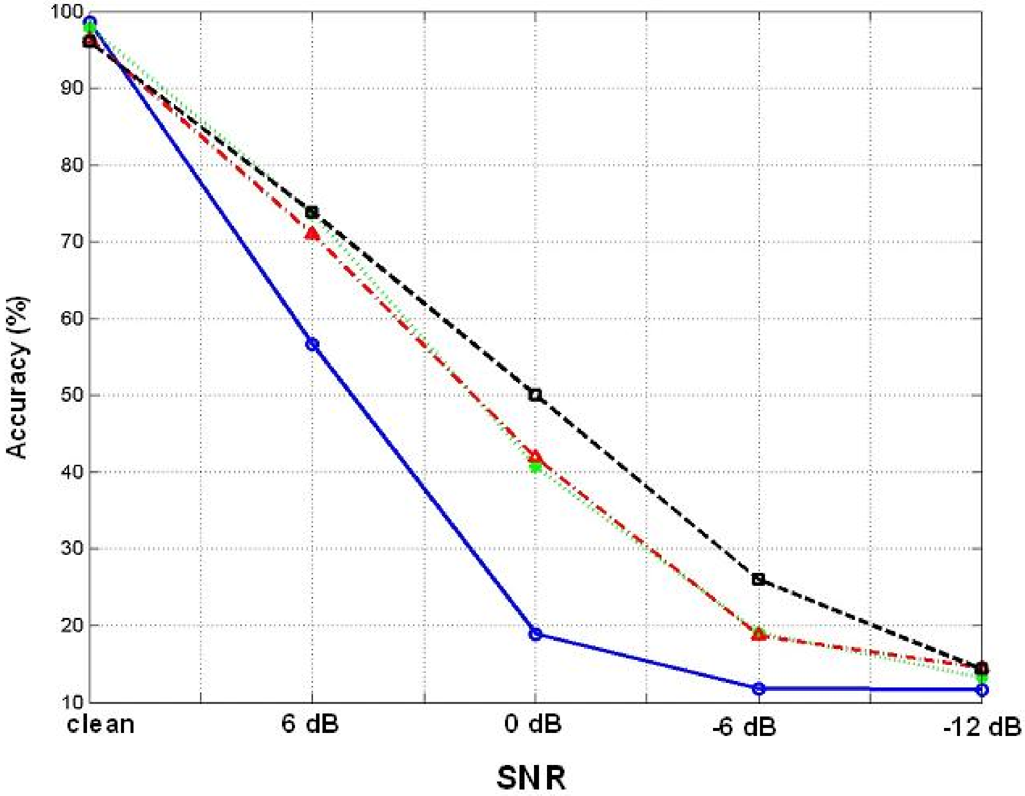

Figure 5.25: Recognition accuracy when the speech signals are corrupted by the speech-shaped-noise. blue solid curve with o : no processing, green dotted curve with * : MPO$_{hon}$, red dash-dotted curve with △ : MPO$_{hoff}$, black dashed curve with □ : MPO$_{4k}$

Table 5.23: Recognition accuracy for two-talker condition

Type	6 dB	3 dB	0 dB	-3 dB	-6 dB	-12 dB
no processing	63.58	45.75	31.92	19.42	11.75	6.75
MPO$_{hon}$	56.17	41.42	29.42	18.58	12.83	8.25
MPO$_{hoff}$	53.08	42.33	33.17	24.58	18.67	13.58
MPO$_{4k}$	53.75	44.42	34.25	26.00	18.58	12.75

using the noisy test utterances. The results are also plotted in Fig. 5.25. It is evident from the figure that all of the three ways in which the *MPO profile* is applied to the test set result in an improvement in the accuracy. The results obtained in the clean condition with either of the three methods are very similar to the ones obtained in 'no-processing' condition implying that the MPO-processing retains most of the speech information when the input is clean speech. The slight drop in accuracy (from 98.56% to about 97%) could be because the MPO-processing removes most of the obstruent information. The MPO$_{4k}$ processing leads to an increase in the accuracy of about 31% at 0 dB SNR.

Table 5.23 shows the recognition accuracy on the test set when the speech signals are corrupted by other competing utterances at various TMRs. The row corresponding to 'no-processing' shows the baseline results obtained using the noisy test utterances. These results are also plotted in Fig. 5.26. The figure shows that MPO-processing leads to a slight drop in the accuracy at positive SNRs and a slight increase in the accuracy at negative SNRs. These results are not surprising as the corrupting noise in this case is a competing speech signal which is also narrowband.

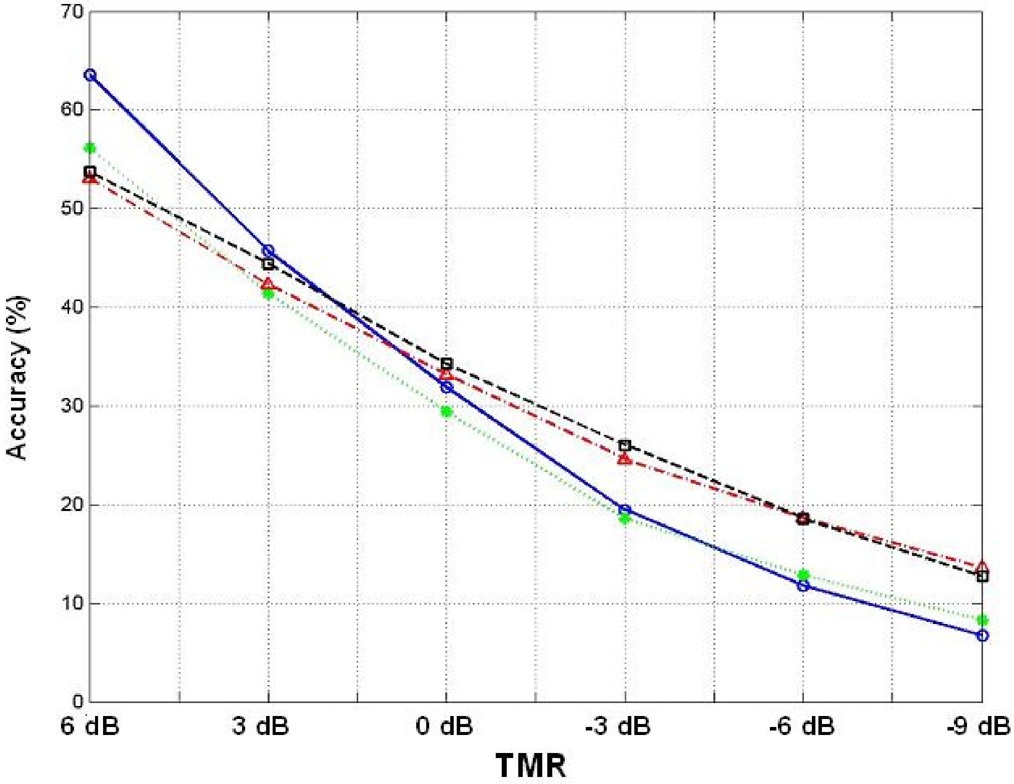

Figure 5.26: Recognition accuracy for speech corrupted by simultaneous speech from one more speaker. blue solid curve with o : no processing, green dotted curve with * : MPO$_{hon}$, red dash-dotted curve with △ : MPO$_{hoff}$, black dashed curve with □ : MPO$_{4k}$

Table 5.24: Categorized recognition results for two-talker condition

SNR	same talker	same gender	diff gender	average
6dB	52.94	55.59	53.00	53.75
3dB	44.34	45.81	43.25	44.42
0dB	30.54	35.75	37.00	34.25
-3dB	24.43	27.37	26.50	26.00
-6dB	16.29	20.11	19.75	18.58
-9dB	11.31	14.53	12.75	12.75

In this case, the MPO-processing will retain both the target speech signal as well as the masking signal. The results for the two-talker case can be categorized further based on whether the talker and the masker are the same, have the same gender or have different genders. These results are tabulated for the MPO$_{4k}$ case in table 5.24. MPO-processing does not favor any one category over the others as the interfering noise in all the categories is still narrowband.

It might be possible to use the MPO-processing in conjunction with the spectro-temporal profile of proportion of periodicity and aperiodicity at each time-frequency unit as well as the pitch estimates generated by the APP detector [95] to improve the overall performance when the corrupting noise has spectral characteristics very similar to that of speech signals.

5.7.2 Recognition results on the Aurora database

The robust speech recognition experiments on the Aurora database were conducted using the baseline recognizer provided with the database. The recognizer is based on the HTK speech recognition software. The speech signal is parameterized into 12 Mel-cepstral coefficients and the energy along with the first and the second-order derivatives to form a 39-dimensional parameter vector. Each of the 11 digits is modeled as a whole-word 16-state HMM with a left-to-right sate-transition topology with no skips allowed over the states. The output of each state is modeled as a mixture of 3 Gaussians with diagonal covariance matrices. The grammar is modeled so that a sequence of any number of digits is permitted.

Tables 5.25, 5.26 and 5.27 compare the recognition accuracies when the recognizer is trained on clean speech and tested on the different noise types in test subsets 'a', 'b' and 'c', respectively, at different SNRs when the unprocessed noisy speech signals are used for testing and when the noisy speech signals are replaced by the corresponding MPO-enhanced speech signals. To minimize the mismatch in the training and testing conditions, the recognizer was trained using the original clean utterances when unprocessed noisy speech signals were used for evaluations and the recognizer was trained on MPO-processed clean speech signals when the MPO-enhanced noisy speech signals were used for evaluations.

Replacing the noisy speech signals by the corresponding MPO-enhanced speech signals results in an increase in the accuracy for most of the noise types at low SNRs, but the performance drops in high SNR situations. One of the main reasons for this drop is the inability of the MPO enhancement scheme to retain the obstruent information. As a result, the trained models have inadequate information about the obstruents which are more prominent at higher SNRs. A significant drop in performance is noticed at all the SNRs when babble-corrupted speech signals are replaced by the corresponding MPO-enhanced speech signals, mainly because a significant amount of noise is passed as valid speech signals, leading to numerous insertion errors and hence a negative accuracy.

The obstruent information in the speech signal can be retained by applying the MPO processing only in the non-obstruent regions and passing the obstruent regions without any modifications (i.e. setting the *MPO profile* uniformly to 1 in all the spectro-temporal channels in obstruent regions). The APP detector, which does a reliable job of separating obstruent regions from sonorant regions, can be used to pull out the obstruent regions. Such a strategy may lead to improved performance especially at higher SNRs.

Table 5.25: Results for test subset 'a' when only clean data was used for training. orig: The recognizer is trained using original clean speech utterances and evaluated on unprocessed noisy speech utterances. MPO: The recognizer is trained using MPO-processed clean speech utterances and evaluated on MPO-processed noisy speech utterances.

SNR	N1		N2		N3		N4	
	orig	MPO	orig	MPO	orig	MPO	orig	MPO
∞	98.83	96.93	98.97	96.67	98.81	96.60	99.14	96.95
20 dB	96.96	88.92	89.96	62.52	96.84	92.96	96.20	88.74
15 dB	92.91	84.74	73.43	52.18	89.53	88.91	91.85	83.28
10 dB	78.72	75.74	49.06	39.24	66.24	80.55	75.10	73.43
5 dB	53.39	59.72	27.03	22.19	33.49	62.42	43.51	53.04
0 dB	27.30	35.06	11.73	4.02	13.27	34.60	15.98	29.07
-5 dB	12.62	17.10	4.96	-5.23	8.35	15.63	7.65	12.16

Table 5.26: Results for test subset 'b' when only clean data was used for training. orig: The recognizer is trained using original clean speech utterances and evaluated on unprocessed noisy speech utterances. MPO: The recognizer is trained using MPO-processed clean speech utterances and evaluated on MPO-processed noisy speech utterances.

SNR	N1		N2		N3		N4	
	orig	MPO	orig	MPO	orig	MPO	orig	MPO
∞	98.83	96.93	98.97	96.67	98.81	96.60	99.14	96.95
20	89.19	62.11	95.77	85.34	90.07	65.97	94.38	84.29
15	74.39	55.20	88.27	79.38	76.89	57.02	83.62	78.37
10	52.72	43.02	66.75	69.71	53.15	42.32	59.61	68.68
5	29.57	27.36	38.15	52.33	30.69	26.54	29.74	51.00
0	11.70	11.67	18.68	30.11	15.84	12.79	12.25	27.74
-5	5.00	-1.35	10.07	14.21	8.11	1.10	8.49	12.77

Table 5.27: Results for test subset 'c' when only clean data was used for training. orig: The recognizer is trained using original clean speech utterances and evaluated on unprocessed noisy speech utterances. MPO: The recognizer is trained using MPO-processed clean speech utterances and evaluated on MPO-processed noisy speech utterances.

SNR	N1		N2	
	orig	MPO	orig	MPO
∞	99.02	97.11	98.97	96.92
20	94.47	78.32	95.19	82.41
15	87.63	72.98	89.69	77.06
10	75.19	59.04	75.27	62.24
5	52.84	38.62	48.85	45.71
0	26.01	19.68	21.64	27.36
-5	12.10	11.05	10.70	15.21

Chapter 6

Conclusions and future directions

Several different approaches have been proposed in the literature to bridge the gap between the performance of automatic speech recognition system and human speech perception, especially when the ambient noise levels are not negligible. The performance of human speech perception is robust till very low SNRs whereas the performance of the automatic speech recognizers drops drastically, even at moderate to high SNRs. In the present work, a speech enhancement technique called the Modified Phase Opponency model was developed from a model of the auditory system. The proposed MPO speech enhancement technique does not need to estimate the characteristics of the corrupting noise, nor does it make any limiting assumptions about the noise. The MPO speech enhancement scheme is based on the fact that speech signals, for most part, are composed of narrowband signals (i.e. harmonics) with varying amplitudes and that the harmonics that are higher in amplitude are perceptually more significant. The MPO speech enhancement scheme detects presence of narrowband signals embedded in wideband noise by using a combination of a bandpass filter and an allpass filter tuned to different center frequencies over the frequency range of interest.

It was shown that, compared to some of the other enhancement techniques, the MPO enhancement scheme strikes a better balance between the amount of noise removed and the amount of perceptual distortion introduced in the enhanced speech signals, even when the speech signal is corrupted by noise with time-varying levels and spectral characteristics. The performance of the proposed speech enhancement scheme was evaluated and compared with that of some of the other schemes proposed in the literature using several different LPC-based objective quality assessment measures which estimate the spectral distortion in the clean speech signal and the enhanced speech signal. For most of the cases, the MPO enhancement techniques leads to the lowest increase in the distortion values as the SNR is reduced.

A small set of perceptual hearing tests were conducted on human subjects with normal hearing to evaluate the subjective quality of the MPO-enhanced speech signals. These tests indicate that there is little perceptual difference in the MPO-processed clean speech signals and the corresponding original clean signals as all the listeners preferred the MPO-processed clean speech signals over the original clean speech signals just about as many times as they preferred the original clean speech signals over the MPO-processed speech signals. In general, the MPO-enhanced output was preferred over the output of the other enhancement methods when the speech signals were corrupted by subway noise, but the other enhancement schemes were preferred when the speech signals were corrupted by car noise. The results indicate considerable variance in the preferences across listeners as well as across different SNRs for a given listener, and further perceptual tests on a larger population of listeners are needed to draw reliable conclusions.

The MPO enhancement scheme was also used as a preprocessor block for robust speech recognition systems. Replacing the noisy speech signals with the corresponding MPO-enhanced speech signals leads to an improvement in the recognition accuracies at low SNRs, but at high SNRs it leads to a drop in the performance. The amount of improvement varies with the type of the corrupting noise. When the corrupting noise has speech-like characteristics (e.g. babble noise), the MPO enhancement scheme does not lead to any improvement at any SNR. The drop in performance at high SNRs can be attributed to the fact that the MPO processing does not retain obstruent speech information.

The present work has shown that the MPO enhancement scheme is a promising candidate to enhance speech signals corrupted by additive noise. There are a lot of different ways to extend the MPO enhancement scheme to improve the overall quality of the enhanced speech signals and to improve the performance of robust speech recognition systems.

6.1 Future work

Some of the paths that can be pursued to improve the performance of the MPO speech enhancement technique are:

1. *Noise specific adaptation of the MPO processing scheme*: The MPO enhancement scheme does not need to estimate the noise characteristics nor does it make any assumptions about the noise type. However, if such information were to be available, the MPO processing can be tailored to improve the overall enhancement output. For example, the enhancement objective can be different at low SNRs compared to the objective at high SNRs. At low SNRs,

where the noise level is high, the objective can be to reduce the noise level to a greater extent while sacrificing some of the speech signal. On the other hand, at high SNRs, the objective can be to preserve the speech signal to a greater extent while letting some of the (low amplitude) noise pass. The MPO analysis generates an estimate of the *speech-present* and *speech-absent* spectro-temporal regions. The signal in the *speech-absent* frequency channels is the instantaneous estimate of the corrupting noise and the signal in the *speech-present* frequency channels is the instantaneous estimate of the speech signal for a given temporal region. The relative amplitudes of these estimates can be used to estimate the local SNR and refine the MPO processing in that region in the second pass.

Such an estimate of the SNR can also be used to change the weighting scheme employed to attenuate the *speech-absent* regions (Section 4.3). In the present version, the valleys in the high SNR regions are attenuated by the same factor as the valleys in the low SNR regions. A more appropriate strategy is to use weights closer to 1 when the estimated SNR is high and lower weights when the estimated SNR is low.

2. *Restoring the fricatives*: The present version of the MPO speech enhancement scheme can detect the frequency onset of frication as well as the formant movement through the fricatives particularly well in clean or at 20 dB SNR when the frication is strong. This information can be used to locate the fricatives and pass the high frequency information in the corresponding regions without

any modifications. Detailed analysis of a sizeable set of utterances corrupted by various noise types at high SNRs is needed to develop a strategy that can retain the fricatives without increasing the amount of residual noise. The output of the APP detector can also be used to separate the likely fricatives from the sonorants so that no extra high frequency noise is retained in the sonorants.

3. *Frequency spacing of the MPO filters:* In the present version of the MPO enhancement scheme, the MPO filters are spaced every 50 Hz from 100 Hz to just below the Nyquist's frequency. Preliminary analysis shows that increasing the spacing from 50 Hz to 100 Hz reduces the computational cost tremendously with minimal loss of robustness. More detailed analysis is needed to quantize the effect of the filter spacings on the performance of the MPO enhancement scheme. A perceptually more relevant ERB spacing scheme can also be explored as an alternative to the current linear spacing.

4. *From MFCCs to more robust features*: All the robust speech recognition experiments conducted in this work used the standard MFCC-based front-end. We previously showed [96] that the speech-production-knowledge-based Acoustic Parameters (APs) are more robust to linear-filtering distortions as compared to the MFCCs. The performance of the APP detector drops slightly when the speech signals are spectrally impoverished [107]. Thus, using the AP-based front-end in conjunction with the MPO-enhanced speech signals may lead to a further increase in the performance of robust speech recognition.

5. *MPO-based front end*: The *MPO profile* generated by the MPO enhancement scheme can also be used to compute a set of robust features for speech recognition in noise. The *MPO profile* can be thought of as a binary matrix with each column representing the MPO-features at the corresponding temporal frame. The regions where the *MPO profile* is 1 can be replaced by the output of the corresponding MPO filter (Fig. 3.2) to compute the MPO-features. Thus, the MPO-features will have a highly negative value in the spectral regions with speech information and a value of zero in the *speech-absent* regions. Some of the frequency normalization strategies can be employed to reduce the inter-speaker variability.

BIBLIOGRAPHY

[1] L. Carney, M. G. Heinz, M. E. Evilsizer, R. H. Gilkey, H. S. Colburn . “Auditory phase opponency: A temporal model for masked detection at low frequencies”, Acta Acustica vol. 88, pp. 334–347, 2002.

[2] L.C.W. Pols, ”How humans perform on a connected-digits data base”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing-82, vol. 2, pp. 867–870, 1982.

[3] A. Varga , H. J. M. Steeneken, “Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems”, Speech Communication, vol. 12, pp. 247–251, 1993.

[4] N. Deshmukh, R.J. Duncan, A. Ganapathiraju, J. Picone, ”Human Performance on the 1995 CSR Hub-3 corpus”, DARPA SRW, pp. 129–134, 1996.

[5] P.C. Woodland, M.J.F. Gales, D. Pye, V. Valtchev, ”The HTK large vocabulary recognition system for the 1995 ARPA H3 Task”, DARPA SRW, pp. 99–104, 1996.

[6] R. Lippmann, ”Speech recognition by machines and humans”, Speech Communication, vol. 22, 1–15, 1997.

[7] L.R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, vol. 77(2), pp. 257–286, 1989.

[8] M.J.F. Gales, ”Predictive Model-Based Compensation Schemes for Robust Speech Recognition”, Speech Communication vol. 25, 1998.

[9] M.J.F. Gales and S.J. Young, ”Robust Continuous Speech Recognition using Parallel Model Combination”, IEEE Transactions on Speech and Audio Processing, vol. 4, pp. 352–359, 1996

[10] M.J.F. Gales and S.J. Young, ”A fast and flexible implementation of parallel model combination”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing-95, pp. 133–136, 1995.

[11] P. J. Moreno, B. Raj, and R. M. Stern, ”A Vector Taylor Series Approach For Environment-Independent Speech Recognition,” Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing-96, 1996.

[12] N.S. Kim, ”Statistical linear approximation for environment compensation”, IEEE Signal Processing Letters, vol. 5(1), pp. 8–10, 1998.

[13] N.S. Kim, “Non-stationary environment compensation based on sequential estimation. IEEE Signal Processing Letters,” vol. 5(3), pp. 57–59, 1998.

[14] M. Afifiy and O. Siohan, ”Sequential estimation with optimal forgetting for robust speech recognition”, IEEE Transactions on Speech and Audio Processing, vol. 12(1), pp. 19–26, 2004.

[15] A. Shankar, C-H. Lee, ”Stochastic matching for robust speech recognition”, IEEE signal processing letters, vol. 1(8), pp. 124–125, 1994

[16] A. Shankar, C-H. Lee, ”A maximum-likelihood approach to stochastsic matching for robust speech recognition”, IEEE Transactions on Speech and Audio Processing, vol. 4(3), pp. 190–202, 1996

[17] J. C. Junqua, ”The Lombard reflex and its role on human listeners and automatic speech recognizer”, Journal of Acoustical Society of America, vol. 93, pp. 510–524. 1993.

[18] J. C. Junqua, ”The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex”, Speech Communication, vol. 20, pp. 13–22, 1996.

[19] J. Benesty, S. Makino, J. Chen. (Eds.), “Speech Enhancement”, Springer-Verlag, Netherlands, 2005.

[20] P Price, W M Fisher, J Bernstein, and D S Pallett. The DARPA 1000-word Resource Management database for continuous speech recognition”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing-88, pp. 651–654, 1988.

[21] L. Deng, A. Acero, M. Plumpe, X. Huang, ”Large-vocabulary speech recongition under adverse acoustic environments”, Proc. Int. Conf. Spoken Language Processing, 2000

[22] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, ”High-performance robust speech recognition using stereo training data”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, 2001.

[23] L. Deng, J. Droppo, A. Acero, ”Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition”, IEEE Transactions on Speech and Audio Processing, vol. 11(6), pp. 568–580, 2003.

[24] L. Deng, J. Droppo, A. Acero, ”Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise”, IEEE Transactions on Speech and Audio Processing, vol. 12(2), pp. 133–143, 2004.

[25] L. Deng, J. Droppo, A. Acero, ”Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features”, IEEE Transactions on Speech and Audio Processing, vol. 12(3), pp. 218–233, 2004.

[26] Q. Zhu and A. Alwan, ”Non-linear feature extraction for robust recognition in stationary and non-stationary noise,” Computer, Speech, and Language, vol. 17(4), pp. 381–402, 2003.

[27] L. Rabiner, B. Juang,”Fundamentals of speech recognition”, Printice Hall, 1993

[28] J. S. Lim, A. V. Oppenheim, ”All-pole modeling of degraded speech”, IEEE Transactions on Acoustics Speech, Signal Processing, vol. ASSP-26, pp. 197–210, 1978

[29] Y. Ephraim, D. Malah, and B.-H. Juang, ”On the application of hidden Markov models for enhancing noisy speech,” IEEE Transactions on Acoustics Speech, Signal Processing, vol. ASSP-37, pp. 1846–1856, 1989.

[30] M. R. Sambur, ”Adaptive noise canceling for speech signals”, IEEE Transactions on Acoustics Speech, Signal Processing, vol. ASSP-26, pp. 419–423, 1978

[31] S. F. Boll, ”Suppression of acoustic noise in speech using spectral subtraction”, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-27(2), pp. 113–120, 1979.

[32] M. Berouti, R. Schwartz, J. Makhoul, ”Enhancement of speech corrupted by additive noise”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing-79, pp. 208–211, 1979

[33] D. V. Compernolle, ”DSP techniques for speech enhancement”, ETRW-92, pp. 1–10, 1992

[34] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system”, IEEE Transactions on Speech and Audio Processing, vol. 7(2), pp. 126–137, 1999.

[35] H. Gustafsson, S. Erik Nordholm and I. Claesson, “Spectral subtraction using reduced delay convolution and adaptive averaging”, IEEE Transactions on Speech and Audio Processing, vol. 9(8), pp. 799–807, 2001

[36] J. Beh, H. Ko, ”A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing-03, pp. 648–651, 2003

[37] R. J. McAulay, M. Malpass, ”Speech enhancement using a soft-decision noise suppression filter”, IEEE Transactions on Acoustics Speech Signal Processing, vol. ASSP-28(2), pp. 137–145, 1980

[38] Y. Ephraim and D. Malah, ”Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator”, IEEE Transactions on Acoustics Speech and Signal Processing, ASSP-32(6), pp. 1109–1121, 1984.

[39] A. Papoulis, “Probability, Random Variables, and Stochastic Processes”, 3rd edition, McGraw-Hill, 1991.

[40] O. Cappe, ”Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor”, IEEE Transactions on Speech and Audio, vol. 2(3) pp. 345–349, 1994

[41] Y. Ephraim and D. Malah, ”Speech enhancement using a minimum mean-square log-spectral amplitude estimator”, IEEE Transactions on Acoustics Speech and Signal Processing, vol. ASSP-33(2), pp. 443–445, 1985.

[42] P. C. Loizou, “Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum”, IEEE Transactions on Speech and Audio Processing, vol. 13(5), pp. 857–869, 2005

[43] I. Cohen, “Speech enhancement using a noncausal a-priori SNR estimator,” IEEE Signal Processing Letters, vol. 11(9), 2004.

[44] Y. Hu, P. C. Loizou, ”Speech enhancement based on wavelet thresholding the multitaper spectrum”, IEEE Transactions on Speech and Audio Processing, vol. 12(1), pp. 59–67, 2004.

[45] Y. Ephraim, H. L. Van Trees, ”A signal subspace approach for speech enhancement”, IEEE Transactions on Speech and Audio Processing, vol. 3, pp. 251–266, 1995

[46] F. Jabloun, B, Champagne, ”Incorporating the human hearing properties in the signal subspace approach for speech enhancement”, IEEE Transactions on Speech and Audio Processing, vol. 11(6), pp. 700–708, 2003

[47] R. Martin, ”Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 253–256, 2002

[48] R. Martin, ”Statistical methods for the enhancement of noisy speech”, IWAENC03, Sept., 2003

[49] B. C. J. Moore, “Introduction to the pschyology of hearing”, Academic Press, London, 1997

[50] S. Greenberg, ”Acoustic transduction in the auditory periphery” Journal of Phonetics, vol. 16, pp. 3–17, 1988.

[51] S. Greenberg, ”The ear as a speech analyzer”, Journal of Phonetics, vol. 16, pp. 139–149, 1988.

[52] L. Deng and C. D. Geisler, ”A composite auditory model for processing speech sounds”, Journal of Acoustical Society of America, vol. 82(6), pp. 2001–2012 1987

[53] L. Deng and C. D. Geisler and S. Greenberg, ”A composite mode of the auditory for processing speech sounds”, Journal of Phonetics, vol. 16, pp. 93–108, 1988

[54] M. B. Sachs, C. C. Blackburn and E. D. Young, ” Rate-place and temporal-place representations of vowels in the auditory nerve and anteroventral cochlear nucleus”, Journal of Phonetics, vol. 16, pp. 37–53, 1988

[55] S. Shamma, ”The acoustic feature sos peech sounds in a model of auditory processing: vowels and voiceless fricatives” Journal of Phonetics, vol. 16, pp. 77–91, 1988.

[56] X. Yang, K. Wang, S. Shamma, ” Auditory representations of acoustic signals”, IEEE Transactions on Information Theory 38(2) pp. 824-839 1992

[57] K. Wang and S. Shamma, ” Self normalization and noise-robustness in early auditory representations”, IEEE Transactions on Speech and Audio 2(3) pp. 421–435 1994

[58] Y. M. Cheng, D. I’Shanughnessy, ”Speech Enhancement based conceptually on auditory evidence”, IEEE Transactions on Signal Processing, 39(9), pp. 1943-1954, 1991

[59] J. H. Hansen and S. Nandkumar, "Robust estimation of speech in noisy backgrounds based on aspects of the auditory process", J. Acoust. Soc. Am. 97 (6) pp. 3833-3849, 1995

[60] S. Seneff, "A Joint Synchrony/Mean-rate Model of Auditory Speech Processing," Journal of Phonetics, vol. 16, pp. 55–76, 1988.

[61] S. Seneff, "Pitch and spectral estimation of speech based on auditory synchrony model", Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 36.2.1-36.2.4, 1984

[62] S. Seneff, "A computational model for the peripheral auditory system" application to speech recognition research", Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 37.8.1-37.8.4, 1986

[63] A. M. A. Ali, J. V. derSpiegel, P. Mueller, "Robust Classification of Stop Consonants Using Auditory-Based Speech Processing", Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 81-84, 2001

[64] H. You, Q. Zhu, A. Alwan, "Entropy-based Variable Frame Rate Analysis of Speech Signals and Its Application to ASR", Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 549-553, 2004.

[65] B. K. W. Mak, Y-C Tam and P. Q. Li, " Discriminative auditory-based features for robust speech recognition", IEEE Transactions on Speech and Audio 12(1) pp. 27-36 2004

[66] K. Kasper, H. Reininger and D. Wolf, "Exploiting the potential of auditory preporcessing for robust speech recognition by locally recurrent neural networks", Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, vol. 2, pp. 1223-1227, 1997

[67] M. Kleinschmidt, J. Tchorz, B. Kollmeier, "Combining Speech Enhancement and Auditory Feature Extraction for Robust Speech Recognition", Speech Communication, 34(1-2). pp. 75-91. 2001

[68] A. Ivanov, A. Petrovsky, "Analysis of the IHC adaptation for the anthropomorphic speech processing system," Eurasip Journal on applied signal processing, 9 pp. 1323-1333, 2005.

[69] K. J. Palomaki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation", Speech Communication, vol. 43, pp. 361-378. 2004

[70] J. Barker, L. Josifovski, M. P. Cooke, P. D. Green, ”Soft decisions in missing data techniques for robust automatic speech recognition”, Proc. Int. Conf. Spoken Language Processing, pp. 373-376, 2000

[71] A. S. Bregman, “Auditory scene analysis”, MIT Press, Cambridge, 1990

[72] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, ”Robust automatic speech recognition with missing and uncertain acoustic data”, Speech COmmunication, vol. 34 pp. 267-285, 2001

[73] J. Ming, P. Jancovic and F. J. Smith, ”Robust speech recognition using probablistic union models,” IEEE Transactions on Speech Audio Processing, 10(6), pp. 403-414, 2002.

[74] Y. Ohshima and R. M. Stern Jr., ”Environmental robustness in automatic speech recognition using physiologically-motivated signal processing”, Proc. Int. Conf. Spoken Language Processing, 1994

[75] O. Ghitza, ”Temporal non-place information in the auditory nerve firing patterns as a front-end for speech recognition in a noisy environment,” Journal of Phonetics, vol. 16 , pp. 109-124. 1988

[76] O. Ghitza, ”Auditory models and human performance in tasks related to speech coding and speech recognition,” IEEE Transactions on Speech and Audio 2(1) pp. 115-132 1994.

[77] K. Wang and S. Shamma, ”Spectral shape analysis in the central auditory system”, IEEE Transactions on Speech and Audio 3(5) pp. 382-395, 1995

[78] N. Mesgarani, S. A. Shamma, ”Speech enhancement based on filtering the spectrotemporal modulations”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, 2005

[79] B. S. Atal, ”Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification”, J. Acoust. Soc. Am. 55 (6), pp. 1304-1312. 1974

[80] B. H. Juang, L. R. Rabiner, “Signal restoration in spectral mapping”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 2368-2371, 1987

[81] Q. Zhu and A. Alwan, ”The Effect of Additive Noise on Speech Amplitude Spectra: a Quantitative Approach,” the IEEE Signal Processing Letters, vol. 9, pp. 275-277, 2002

[82] B. Strope and A. Alwan, ”A model of dynamic auditory perception and its application to robust word recognition”, IEEE Transactions on Speech and Audio Processing, 5(2), pp. 451-464, 1997

[83] S. Ikbal, H. Misra, H. Bourlard, ”Phase autocorrelation (PAC) derived robust speech features”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, 2003

[84] C. R. Jankowski, H. D. H. Vo, R. P. Lippmann, ”A Comparison of Signal Processing Front Ends for Automatic Word Recognition”, IEEE Transactions on Speech and Audio Processing , 3(4), pp.286-293, 1995

[85] H. Hermansky, ”Perceptual linear predictive (PLP) analysis of speech”, Journal of Acoustical Society of America, vol. 87 (4), pp. 1738-1752. 1990

[86] H. Hermansky, ”RASTA processing of speech”, IEEE Transactions on Speech and Audio 2(4) pp. 578-589 1994.

[87] H. Hermansky, ”Auditory modeling in automatic recognition of speech”, ECSAP-97, pp. 17–21, 1997

[88] T. V. Sreenivas, K. Singh, R. J. Niederjohn, ” Spectral resoultion and noise robustness in auditory modeling”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 817-820, 1990

[89] B. Gajic, K. K. Paliwal, ”Robust speech recognition using features based on zero crossing with peak amplitudes”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 64-67, 2003

[90] D-S Kim, S-Y Lee, R M. Kil, ”Auditory processing of speech signals for robus speech recognition in real-world noisy environments”, IEEE Transactions on Speech and Audio Processing, 7(1), pp. 55-69, 1999

[91] N. Bitar, ”Acoustic Analysis and Modeling of Speech Based on Phonetic Features”, Ph.D. thesis, Boston University, 1997

[92] O. Deshmukh, C. Espy-Wilson, ”Speech Enhancement Using Auditory Phase Opponency Model”, Proc. Eurospeech, pp. 2117–2120, 2005.

[93] O. Deshmukh, M. Anzalone, C. Espy-Wilson, L. Carney, “A noise reduction strategy for speech based on phase-opponency detectors”, 149th Meeting of the ASA, 2005.

[94] T. Pruthi, C. Y. Espy-Wilson, “Acoustic parameters for automatic detection of nasal manner”, Speech Communication, vol. 43, pp. 225-239, 2004.

[95] O. Deshmukh, C. Y. Espy-Wilson, Ariel Salomon, J.singh, “Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech”, IEEE Transactions on Speech and Audio Processing, 13(5), pp. 776-786, 2005.

[96] O. Deshmukh, C. Espy-Wilson and A. Juneja, ”Acoustic-phonetic speech parameters for speaker independent speech recognition”, Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 593-596, 2002.

[97] Stevens K., “Acoustic Phonetics”, M.I.T. Press, Cambridge, 1999

[98] D. L. Wang, “On ideal binary mask as the computational goal of auditory scene analysis.”, in Divenyi P. (ed.), Speech Separation by Humans and Machines, Kluwer Academic, Norwell, 181–197, 2005

[99] H.G. Hirsch, D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions”, Proceedings of ISCA Tutorial and Research Workshop ASR2000, Paris, France, 2000.

[100] R.G. Leonard and G. Doddington, “Tidigits speech corpus,” Texas Instruments, Inc. 1993.

[101] M. Cooke, J. Barker, S. Cunningham and X. Shao, “ An audio-visual corpus for speech perception and automatic speech recognition”, submitted to Journal of Acoustical Society of America

[102] J. H. Hansen, L. M. Arslan, “Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus”, IEEE Transactions on Speech and Audio Processing, vol. 3(3), pp. 169–184, 1995.

[103] J. Hansen , and B. Pellom, “An effective quality evaluation protocol for speech enhancements algorithms,” Proceedings of Inter. Conf. on Spoken Language Processing, pp. 2819–2822, 1998.

[104] P. Scalart and J. Vieira-Filho, “Speech enhancement based on apriori signal to noise estimation,” Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, pp. 629-632, May 1996.

[105] O. Deshmukh, C. Espy-Wilson, “Modified phase opponency based solution to the speech separation challenge”, to appear, Proc. Interspeech Int. Conf. Spoken Language Processing, 2006.

[106] J. Barker, and M.P. Cooke, “Modelling speaker intelligibility in noise,” submitted to Speech Communication.

[107] A. Salomon, C. Espy-Wilson, O. Deshmukh, ”Detection of Speech Landmarks: Use of Temporal Information”, Journal of Acoustical Society of America, vol. 115, pp. 1296-1305, March 2004.

[108] S. Young, ”The HTK book”, http://htk.eng.cam.ac.uk, 2002.

[109] R. Martin, “Statistical methods for the enhancement of noisy speech”, Inter. Workshop on Acosut. Echo and Noise Control, Kyoto, Japan, Sept. 2003

[110] G. Fant and A. Risberg “Auditory matching of vowels with two formant synthetic sounds”, STL-QPRS 4, pp. 7–11, Royal Institute of Technology, Stockholm.

[111] L.A. Chistovich, “Central auditory processing of peripheral vowel spectra”, Journal of Acoustical Society of America, 77, pp. 789–805, 1985.