Speech enhancement methods based on perceptual wavelet filterbank

Shao, Yu

2009

Shao, Y. (2009). Speech enhancement methods based on perceptual wavelet filterbank. Doctoral thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/42244

https://doi.org/10.32657/10356/42244

SPEECH ENHANCEMENT METHODS BASED ON PERCEPTUAL WAVELET FILTERBANK

SHAO YU

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2009

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Associate Professor Chang Chip Hong, without whom motivation and encouragement, I would not have considered a graduate career in speech enhancement for robust speech recognition research area. His expertise, understanding, and patience have added considerably to my graduate experience. I appreciate his vast knowledge and skills in many areas (e.g., computer arithmetic circuits, watermarking for IP protection, algorithms and architectures for digital image and speech signal processing, computer aided design of VLSI digital circuits, interpersonal and enquiring skills), and his assistance in the preparation of technical reports, conference and journal papers and this thesis. Again, a very special thanks goes to Professor Chang for giving me the freedom to pursue this research. It was under his tutelage that I developed a focus and became interested in speech enhancement for robust speech recognition. He provided me with direction, technical support. To me, he is more of a mentor and friend than a professor. When my previous supervisor left NTU, it was through Professor Chang's persistency, understanding and kindness that I could continue my PhD research during the hardest time. I doubt that I will ever be able to convey my appreciation fully, but I owe him my eternal gratitude. What I have learned from him will benefit me well beyond my graduation in my future career and personal life. Special thanks to Mrs. Chang for her continuous concern and kindness, from which I gained lots of wonderful memories. I really appreciate the time spent with them.

I would like to thank my previous advisor, Associate Professor Tong Yit Chow for his guidance in my studies and my work. He provided many helpful suggestions on conducting research and given me the technical and academic support before his retirement. Thanks also go to my friends in Nanyang Technological University who provided me with good advices at times of critical need and for the exchanges of knowledge, skills, and venting of frustration during my graduate program. They had enriched my postgraduate study experience.

Finally, my greatest appreciation is due to my family for the support they provided me throughout my life. I am deeply indebt to my parents. Without their selfless loves and encouragement, I would not have accomplished this work. Therefore, this thesis is specially dedicated to them.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Abbreviations	Full Expressions
ARF	Autoregressive Function
ASR	Automatic Speech Recognition
ATH	Absolute Threshold of Hearing
BM	Basilar Membrane
CBW	Critical Bandwidth
CKF	Constrained Kalman Filter
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DIF	Discrimination Information Function
DWT	Discrete Wavelet Transform
EM	Expectation Maximization
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
GPTFS	Generalized Perceptual Time-Frequency Subtraction
HMM	Hidden Markov Model
HP	Highpass
IS	Itakura-Saito
JND	Just Noticeable Distortion
KF	Kalman Filter
KLT	Karhunen-Loeve Transform
LDB	Local Discriminant Bases
LP	Lowpass
LPC	Linear Predictive Coefficient
MFCC	Mel Frequency Cepstral Coefficients
ML	Maximize Likelihood
MLP	Multilayer Perceptron
MOS	Mean Opinion Score
MTCKF	Masking Threshold Constrained Kalman Filter
QMF	Quadrature Mirror Filter
PESQ	Perceptual Evaluation of Speech Quality
PSD	Power Spectral Density
PSS	Perceptual Spectral Subtraction
PWF	Perceptual Wavelet Filterbank
PWKF	Perceptual Wavelet Kalman Filter
PWPT	Perceptual Wavelet Packet Transform
Rasta-PLP	The Relative Spectral Perceptual Linear Prediction
SFM	Spectral Flatness Measure
SKF	Subband Kalman Filter
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SS	Spectral Subtraction
STFT	Short-time Fourier Transform
SVM	Support Vector Machine
SVs	Support Vectors

USE	Unvoiced Speech Enhancement
VAD	Voice Activity Detector
WKFP	Wavelet Kalman Filter with Post-filter
WPT	Wavelet Packet Transform

SUMMARY

The extraction and synthesis of quality speech is of utmost importance for a diversity of digital speech processing applications ranging from mobile communication systems, digital hearing aids, quality enhancement of old records, speech recognition systems to hands-free car kits. Most speech processing systems that work well in underrated noise condition degrade drastically in practical noisy environments. Noise induced aggravation causes inaccurate information exchange and lowers the quality of speech. An overwhelming number of methods have been developed over years to address the speech enhancement problem. This thesis investigates into versatile speech enhancement algorithms in order to increase the intelligibility of speech and reduce auditory fatigue.

A new generalized perceptual time-frequency subtraction (GPTFS) method has been proposed to meet the demand for quality noise reduction even at very low signal-to-noise ratio (SNR). A psychoacoustic model is incorporated into the generalized perceptual wavelet denoising method to reduce the residual noise and improve the intelligibility of speech. Simultaneous masking and temporal masking of the human auditory system are modelled by the perceptual wavelet packet transform. The wavelet multirate signal representation has been advantageously exploited to preserve the critical transient information. A time-frequency masking threshold is deduced to adaptively adjust the subtraction parameters of the proposed method. An unvoiced speech enhancement algorithm is also integrated into the system to improve the intelligibility of speech. Through rigorous objective and subjective evaluations, it is shown that the proposed speech enhancement system is capable of reducing noise with little speech degradation in adverse noise environments and the overall performance is superior to several competitive methods.

Another novel perceptual wavelet Kalman filtering (PWKF) approach has also been developed. It is a coherent integration of the state-space formulation of Kalman filter with the subband excision of discrete wavelet transform. Using a mixture of excitation signals consisting of periodic pulses and white noise, a subband transformed voiced-unvoiced speech model is established. It enables the long term characteristics of speech and noise to be analyzed separately in each critical band. To further improve the intelligibility of clean speech estimated by the wavelet Kalman filter, a perceptual weighting filter is designed with an adaptive time-frequency masking threshold tailored to the human auditory perception. The speech enhancement performances of a variety of noisy speeches are analyzed using both objective assessments and subjective listening tests such as spectrogram plot, segmental SNR, perceptual evaluation of speech quality (PESQ) score and mean opinion score (MOS). The results indicate consistently that the proposed perceptual wavelet Kalman filtering method produces better quality speech than several other competitive methods.

Last but not least, this thesis also explores the use of wavelet based techniques to harness the performance of automatic speech recognition (ASR) in the presence of background noise. A robust speech recognition system is realized by cascading speech enhancement preprocessing, feature extraction, and hybrid speech recognizer in time-frequency space. Speech enhancement preprocessing is applied to minimize the mismatch between the training and testing conditions of the classifier. The effect of speech enhancement on classifier performances are evaluated using the enhanced speech estimated by the proposed GPTFS and PWKF methods. The non-phonetic information is discarded while the more critical speech features are extracted and represented by the wavelet coefficients via feature extraction. Two wavelet based feature extraction methods, wavelet packet transform and local discriminant bases (WPT/LDB) and perceptual wavelet filterbank (PWF), are evaluated. The denoised wavelet features are fed to the hybrid classifier founded on Hidden Markov Model (HMM). The intrinsic limitation of HMM is overcome by augmenting it with either Multilayer Perceptron (MLP) or Support Vector Machine (SVM). The connected digit recognition experiments conducted on the proposed framework show encouraging results of various ASR configurations with the denoised wavelet features. The ingenious configuration of PWF-PWKF-SVM/HMM is shown to be particularly promising. It significantly improves recognition performance at low SNR without causing poorer performance at high SNR.

CHAPTER 1.

INTRODUCTION

1.1 Motivation

Today, computers, multimedia and networking technologies are being used by tens of millions of people worldwide to manage and operate their businesses, to deliver information, entertainment and convenience to mobile users and car drivers. Speech processing applications are in the vanguard of this revolution. Currently, users and enterprises demand high quality and high perceptibility speech processing used in real-world applications, such as hands-free car kits, the communicators in noisy cockpits, hearing aid, and the interactive voice service systems for cellular phones [AAR04], [JEA01], [JEO01], [KAH97], [LAI99], [ORT05], [PAR02], [SPR07], [WU00], [WES97]. The presence of noise of various magnitudes and forms is one of the main hurdles for accurate information exchange as it lowers the intelligibility and quality of speech. For example, the quality of communication of a hand-free car kit system is inevitably affected by a composition of several varieties of noise, such as car engine, traffic, wind, and babble noise. In hearing aid applications, background noise lowers the perception and creates listening fatigue. In automatic speech recognition (ASR), some mismatch sources, such as ambient noise and microphone distortion, which are not present in the training dataset, have also limited the effectiveness of ASR in practical applications. Hence, speech enhancement algorithms are indispensable in the plethora of digital audio communication, hearing aids, multimedia, gaming and entertainment applications.

The task of speech enhancement is to produce an estimate of clean speech to minimize the difference between the clean speech and the enhanced speech. Over the last three decades, the developments in digital signal processing have resulted in a wide variety of techniques for removal of noise in degraded speech, depending on the types of application and the characteristics of background noise [DEL99]. It is observed that majority of the audio and speech processing applications use a single microphone system. In single microphone techniques, there is no reference input available to estimate the noise and noise reduction has to be accomplished with speech dependent filtering. The process will generate some non-stationary residual noise. This residual noise is annoying and exhibits an audible tonal characteristic referred to as the musical noise [DEL99], [VAS00], [QUA02]. Furthermore, the noise has to be estimated using properties of the speech and noise signals such as stationarity and frequency content, or during silence/noise periods using a voice activity detector (VAD) [VAS00]. This makes problem of speech enhancement with a single-channel more challenging. Despite the drawbacks, speech enhancement methods based on a single microphone continue to be used in most noise scenarios and are very popular due to their simplicity and the application constraints.

Single-microphone speech enhancement algorithms fall into two main categories: nonparametric signal processing method [AYA04], [BAH01], [BER79], [BOL79], [DON95], [EPH84], [EPH95], [HAN06], [HAS04], [HER94], [HU03], [HU04], [JAB03], [KAM02], [LI01], [LIN03], [LIU06], [LU04], [MCA80], [MIT00], [SCA96], [SIM98], [VAS00], [VIR99], [WOO00] and model-based signal processing method [BRO02], [EPH92a], [EPH92b], [FED89], [GAB99], [GAB01], [GAN98], [GIB91], [GOH99], [GRA00], [GRA06], [JU07], [LEE96], [LEE97], [LEE00], [LIM78], [LOW91], [MA06], [MOU07], [NIE96], [PAL87], [POP98], [SAM98], [SRE96], [SRI06], [VAS00], [VES03], [WU98]. Methods from the first category usually process the input signal as a waveform or a digitized sequence, and do not make use of any parametric model of the signal generation process or the statistical distribution of the speech and noise [VAS00]. The main attractions of this type method are its relative simplicity and high flexibility. With no awareness of the unique characteristics of signal generation, the performances of these methods are not consistent. In spectral subtraction based algorithms [BRE79], [BOL79], [EPH84], [HAN06], [HAS04], [HU03], [HU04], [KAM02], [LI01], [LU04], [MCA80], [SCA96], [SIM98], [VIR99], [VAS00], [WOO00], noise outside the band of perceptual importance is attenuated in the frequency domain and then the speech is recovered in the time domain. The transformation of the signal from the frequency domain to the time domain usually causes additional degradation of the enhanced speech. As opposed to the nonparametric methods, the model-based methods assume a model for the signal generation process. Typically, a linear prediction model is used to model the acoustic excitation of a spoken word. Since clean speech can be represented by an autoregressive (AR) model [HAY01], [QUA02], [VAS00], the enhanced speech can be generated by estimating the parameters of AR model from the noisy signal. Model-based methods normally outperform nonparametric methods since the abstraction of the characteristic information in the form of a generalized model enables the problem to be solved by more powerful analytical and statistical signal processing techniques [VAS00]. However, the techniques chosen to solve the problem can be sensitive to the deviations from the class of signals characterized by the model. Thus, the assumptions and constraints made in any speech enhancement method are dependent on the application and the environment. Even though model-based method can remove a substantial amount of background noise, the natural sound of speech is also degraded in the synthesis process.

Although significant progresses have been made in digital speech processing algorithms in general, there are still considerable challenges in speech enhancement. Removing various types of noise is difficult due to the random nature of the noise and the inherent complexities of speech. One challenge is the difficulty to derive a good model for various adverse noises. Noise is defined as any unwanted signal that can interfere with the desired signal and it can be a byproduct of the imperfect communication channel or internally generated distortions in the process of signal processing. It is by nature a stochastic signal and is omnipresent. The success of a speech enhancement algorithm depends on its ability to characterize and model the noise process, and to use the noise characteristics advantageously to differentiate signal from noise. Unfortunately, speech and noise are not easily distinguishable. The environmental noise model has proven to be very complex. It is a difficult problem to identify a good engineering structure to accurately model both the temporal and spectral characteristics of acoustic noise.

Another challenge is the difficulty to obtain a perfect speech model from the voiced-unvoiced feature sets of the speech signal. The voiced speech consists of vowels, semi-vowels and nasals, with vowels being the largest phoneme group. Vowel has a relatively long duration of 40–250 ms. The energy of vowel ranges from 100 Hz to

4500 Hz and its power spectrum is concentrated in the range below 1000 Hz [QUA02]. Unvoiced speech consists of the stop, fricative and affricate consonants, and contributes considerably to the intelligibility of speech. The duration of unvoiced phonemes is in the range of 10-50 ms. Its power spectrum lies in much higher frequency bands than that of vowel and can extend to about 7 or 8 kHz [QUA02]. In reality, there are various types of phonemes mixed and randomly distributed in frequency with different durations. It is difficult to find an appropriate set of analysis parameters and a good representation of the speech signal due to its inherent complexities. Model selection is critical in model-based algorithms, which generally assume a white Gaussian noise excited AR model for speech [DEL99], [HAY01], [QUA02]. This model simplifies the problem formulation but it is insufficient and only valid for unvoiced speech segment.

Speech enhancement techniques usually involve a tradeoff between the amount of noise to be removed and the speech distortion introduced by the denoising process [VAS00]. A survey of existing speech enhancement methods unveils a major drawback of the single-scale representation of speech signal. This has led to inadequate suppression of the annoying residual noise and reduction of distortion simultaneously in adverse noise conditions. Transform such as wavelet or wavelet packet transform represents speech signal at multiple scales. Multiscale transform provides a flexible and efficient way to capture and represent localized information of speech signal in the time-frequency plane. The research in this thesis is motivated by the promises of multiscale representations in modelling human auditory perception. As none of the existing speech enhancement methods are known to work well in a wide variety of noisy environments at very low signal-to-noise ratio (SNR), typically below 0 dB, there exist good opportunities to research into quality noise reduction algorithms in wavelet transform domain.

1.2 Objectives

The requirements of speech enhancement algorithms vary according to the applications. Three different application goals are highlighted below.

1. For communication systems [JEA01], [ORT05], [PAR02], [WES97], the desired quality enhancement depends on the SNR of the noisy speech. For medium-to-high SNR (>5 dB) environments, the goal of speech enhancement is to produce a natural-like speech signal by reducing the noise level. For low SNR (<5 dB), the objective of speech enhancement is to decrease the noise level while retaining or improving the intelligibility.

2. For hearing aids [KAH97], [LAI99], [SPR07], the major goal is to reduce noise level while at the same time minimize perceptually induced discomfort such as speech intelligibility, abrupt amplitude variation and listener fatigue.

3. For automated speech recognition systems [JEO01], the major goal of speech enhancement is to reduce the mismatch between the training and testing procedure and increase the recognition rate in adverse noise environment.

The purpose of this research is to understand the principles of speech production and perception, and apply them to the development of good engineering model. The aim is to develop versatile speech enhancement methods to combat various adverse noise conditions by appropriately modulating the acoustic parameters to enhance the quality of speech to a satisfying level for a wide spectrum of applications. The emphasis is on the reduction of the effect of residual noise and speech distortion in the denoising process and the enhancement of the denoised speech to improve its intelligibility.

Wavelet transform is used as a vehicle to devise new speech enhancement solutions. The methodologies and solutions are to be developed by exploiting the speech analysis and synthesis model of the human auditory system in the wavelet domain. This research will look into novel speech enhancement algorithms to achieve the desirable quality of speech even when an application is operating in relatively lower SNR conditions. In view of the different application requirements, the flexibility to optimally trade between noise attenuation and processing distortion will be studied and adaptive filtering will be explored to improve the intelligibility of the processed speech. For better results, we will also look into speech and noise models that could take advantage of the perceptual properties of human auditory system for noise reduction.

The following targets have been set towards fulfilling the theme of this research:

(1) To explore the use of wavelet multirate signal representation to preserve critical transient information, and model the simultaneous and temporal masking properties of human auditory perception by virtue of the frequency and temporal localization of wavelet packet transform.

(2) To develop a new subtractive type of speech enhancement method. The method shall optimize the overall reduction of background noise, residual noise and processing

distortion by tuning the time-frequency masking threshold adaptively based on the devised psychoacoustic model in the wavelet domain.

(3) To develop a new model-based speech enhancement method based on Kalman filter theory. Subband voiced-unvoiced speech model in the wavelet transform domain is to be established for the optimal estimation of clean speech observed in white and colored noise.

(4) To integrate the proposed non-parametric and model-based speech enhancement methods into a hybrid speech recognizer to reduce the mismatch between training and testing conditions in ASR application.

1.3 Major Contributions of the Thesis

This research has contributed to the speech processing applications through the development of several competitive and versatile speech enhancement methods. In particular, the following major contributions have been made in this thesis.

The first contribution is the proposal of a perceptual wavelet packet transform (PWPT) to approximate the 21 critical bands of human auditory system up to 8 kHz [SHA05b], [SHA06c], [SHA07a]. It enables the components of a complex sound to be appropriately segregated in frequency and time in order to mimic the frequency selectivity and temporal masking of human auditory system. By incorporating knowledge of human auditory system, this fixed filter bank approximation is simpler to realize than the best basis subband decomposition and yet remarkably useful for residual noise and speech distortion reduction than the use of short-time Fourier analysis. Critical high frequency components in noisy speech have been analyzed by the proposed PWPT filter bank and exploited to further improve the perceptual quality of the processed speech.

The second contribution is the proposal of a novel generalized perceptual time-frequency subtraction method (GPTFS) [SHA07a]. A parametric formulation of the spectral subtraction method for noise reduction based on the generalized perceptual wavelet transform is deduced. We have improved the Fourier transform based gain function of the generalized spectral subtraction method [VIR99] by deriving the close form expressions for the subtraction factor and noise flooring factor to optimize their tradeoffs analytically for simultaneous reduction of background noise, residual noise and speech distortion. These parameters of GPTFS can then be adaptively tuned instead of empirically determined according to the noise level and the masking thresholds derived from the human auditory model in wavelet domain.

A new system structure for speech enhancement is developed to integrate the proposed GPTFS and the unvoiced speech enhancement (USE) in wavelet domain [SHA04], [SHA07a]. The GPTFS works in conjunction with a PWPT to reduce the effect of noise contamination. This process tends to weaken the high frequency speech of low energy. An USE is proposed to tune a set of weights to discriminate the original speech in high frequency bands from the noise so that a soft thresholding can be applied to further improve the intelligibility of the processed speech.

The third contribution is the proposal of a novel perceptual wavelet Kalman filtering speech enhancement method (PWKF) [SHA06b]. Since the speech and noise signals are intrinsically non-stationary, the subtle time variation of their combined spectrum is better preserved in the wavelet domain. In PWKF, the voiced-unvoiced speech is excised into the critical band scale by the perceptual wavelet filter bank. To filter the color noise interference, the subband speech and colored noise are modeled as a white Gaussian noise excited AR process for the state-space formulation of Kalman filter. The model parameters required by the proposed algorithm are estimated from the noisy speech using an iterative expectation-maximization (EM) procedure [DUD01], [HAY01]. The main merit of this method is on the recovery of clean speech from speech corrupted by slowly varying, non-white, additive noise, when only the noisy signal is available. A new system structure for speech enhancement is then devised by amalgamating the voiced and unvoiced speech model, Kalman filter and perceptual weighting filter in the discrete wavelet transform domain. The quality of the speech is greatly enhanced by cascading PWKF with the proposed perceptual weighting filter, which is designed based on an adaptive noise masking threshold.

Last but not least, a versatile speech analysis toolbox is developed using Matlab program. This customized toolbox is originally developed to study, evaluate and compare various performance metrics of different speech enhancement methods. The modularity and flexibility of this speech processing toolbox is used to establish a platform for evaluating new configurations of automatic speech recognition system (ASR). Different configurations of ASR are realized by cascading the proposed GPTFS and PWKF speech enhancement methods with different wavelet based feature extractors and hybrid classifiers based on Multilayer Perceptron and Hidden Markov Model network (MLP/HMM) [BOU94], and support vector machines and HMM (SVM/HMM) [ZHA04]. Thanks to the successful porting of the hybrid MLP/HMM and hybrid SVM/HMM recognizers into the wavelet domain, new insights are gained from the performance evaluation of several new configurations of ASR system [SHA06a], [SHA07b]. This hybrid and hierarchical design paradigm improves the recognition performance by combining the advantages of different integral methods within a single system. The proposed speech enhancement methods help to improve the recognition rate of the ASR system in noisy environments. The connected digit recognition experiments conducted on the proposed framework show encouraging results of various ASR configurations with the denoised wavelet features.

The above contributions have led to a number of international conference and journal publications listed in the author's publications towards the end of the thesis.

1.4 Organisation of the Thesis

The thesis is organized into seven chapters. This chapter introduces the motivation and objectives of the research work. The main contributions that arise from the work presented in this thesis are also highlighted in this chapter.

Chapter 2 first anatomizes the physiology of speech production and human hearing perception to understand the underlying principles of contemporary speech processing algorithms. The autoregressive function for modelling the phoneme of speech signal is described. To provide the background of the model used in Chapter 3, the critical band and masking properties of human auditory system are introduced. A noise model based on the characteristics of noise and distortion is described. The state-of-art speech enhancement methods are reviewed with a specific emphasis on two mainstream methodologies for speech enhancement. They are the subtraction methods and the model-based methods. The limitations and challenges of the existing speech enhancement methods studied in this literature survey serve as a preamble to the more in-depth investigations of Chapters 4, and 5.

The major contributions outlined in the previous section are presented in Chapter 3 through Chapter 5. Chapter 3 presents a wavelet filterbank approach to psychoacoustic modelling of human auditory system. An overview of human hearing constraints is provided. The importance of noise reduction criteria with reference to human perception and the limitations of traditional engineering model of human auditory system are presented. A novel perceptual wavelet filter bank design is proposed to approximate the traditional critical band. The human auditory model developed from the proposed perceptual wavelet filterbank is compared with that of the short-time Fourier transform. Based on the perceptual wavelet filterbank, the simultaneous and temporal maskings are unified to estimate the auditory masking threshold. The threshold is used to adaptively tune the parameters of the speech enhancement systems.

Chapter 4 describes a new generalized perceptual time-frequency subtraction speech enhancement scheme for a single-microphone system. The psychoacoustic model, presented in Chapter 3 is incorporated into the generalized perceptual wavelet denoising method to reduce the residual noise and improve the intelligibility of speech. The time-frequency subtraction is treated as *a posteriori* SNR-dependent attenuator with an optimal time-frequency gain function. The subtraction parameters are optimally determined by a thresholding criterion. Its derivation is presented with reference to temporal and simultaneous masking properties of human perception. An unvoiced speech enhancement algorithm to augment the intelligibility of the processed speech is also presented. The performance analysis of the proposed GPTFS algorithm and its comparison with other speech enhancement algorithms are presented and discussed at the end of this chapter.

Chapter 5 presents a Kalman filtering speech enhancement approach to enhance speech corrupted by acoustic background noise. The mainstay of the proposed method is a coherent integration of the state-space formulation of Kalman filter with the subband excision of discrete wavelet transform. Using a mixture of excitation signals consisting of periodic pulses and white noise, a subband transformed voiced-unvoiced speech model is presented. This subband speech model is used to formulate the state-space equations of Kalman filter for the optimal estimation of clean speech observed in white and colored noise. To further improve the intelligibility of the clean speech estimated by the wavelet Kalman filter, a perceptual weighting filter is designed with an adaptive time-frequency masking threshold tailored to the human auditory perception. The chapter is concluded with the analysis of the speech enhancement performances of a variety of noisy speeches using both objective assessments and subjective listening tests, such as spectrogram plot, segmental SNR, perceptual evaluation of speech quality (PESQ) score and mean opinion score (MOS).

Chapter 6 commences with a brief description of a comprehensive speech analysis and enhancement toolbox. This interactive toolbox has a friendly user interface that enables various analyses of speech and noise, and the performance evaluation of different speech processing algorithms. Some features that facilitate the ease of experimentations with different speech enhancement algorithms using this toolbox are illustrated. Next, an evaluation platform that makes use of the toolbox features for robust speech recognition is presented. This is a framework of wavelet based techniques developed to harness ASR performance in the presence of background noise. The proposed robust speech recognition system is realized by cascading speech enhancement preprocessing, feature extraction, and hybrid speech recognizer in time-frequency space. Two methods for feature extraction, wavelet packet transform and local discriminant bases and perceptual wavelet filterbank are discussed. Two hybrid classifiers founded on HMM with augmenting MLP or SVM are also presented. The effect of speech enhancement on these classifier performances are evaluated using the enhanced speech estimated by the proposed GPTFS and PWKF methods. The chapter ends with the evaluation and analysis of recognition rate of noisy speech in various environments.

Finally, Chapter 7 reviews the results achieved in this thesis and highlights the features and merits of the proposed methods. The pointers to several extended topics worthy of further research are also outlined.

CHAPTER 2.

BACKGROUND AND LITERATURE REVIEW

In the past decades, suppression of additive background noise in the field of speech enhancement has been a challenging area for research. The ultimate goal of speech enhancement is to eliminate the additive noise present in a speech signal, improve the quality and intelligibility, reduce the perceptual fatigue, and synthesis the speech signal to its original form. As result of these research efforts, several different approaches have been developed with one or more of the auditory, perceptual or statistical constraints placed on the speech and noise signals. However, due to inherent complexities of speech and noise in real world situations, it is very difficult to reliably predict the characteristics of the speech and exactly estimate the characteristics of the interfering noise. Hence, the speech enhancement methods are only effective to a limited extent for reducing the amount of noise present in the signal. Due to the non-ideal nature of these methods and their models, some desire components of the speech signal can even be distorted during the process. Hence, the processing distortion becomes more noticeable as the signal-to-noise ratio decreases. The effectiveness of the speech enhancement system can therefore be measured based on how well it performs the task in the light of the trade-off between the residual noise and the processing distortions.

This chapter reviews the physiology of speech production and the hearing perception of human being. The characteristics of noise and distortion will also be investigated. Towards the end, a comprehensive literature survey of two major categories of speech enhancement methods that have been used to date is presented.

2.1 Speech Production and Perception

Speech is a dynamic and information-bearing acoustic waveform. These waves are produced by inhaling air into the lungs, and then exhaling it through the vibrating glottis cords and the vocal tract. The variation of this sound pressure is a result of some sequence of coordinated movements of a series of structures in the human vocal system. The study of the speech sound classes is called *phonemics* [QUA02]. The branch of science that studies these sound variations is known as *phonetics* [QUA02]. The process of human speech communication involves the voice-generating by the speaker and the perception of the acoustic signal by the listener. Though the process of speech perception remains largely a mystery to the scientific world, the process of speech production has been well researched and fully understood.

2.1.1 The Physiology of Speech Production

Fig 2.1 shows an anatomic view of speech production. Speech is produced by a motoric coordination of lungs, larynx (with vocal cords) and articulation tract (mouth and nose cavity) [QUA02]. The lungs act as power sources of airflow. The larynx is a complex system composed of cartilages, muscles and ligaments. It controls the vocal cords, which is made of the oral cavity from the larynx to the lips and the nasal passage. The purposes of the vocal tract are to generate sources for speech production and to spectrally color the source to make distinct speech sounds.

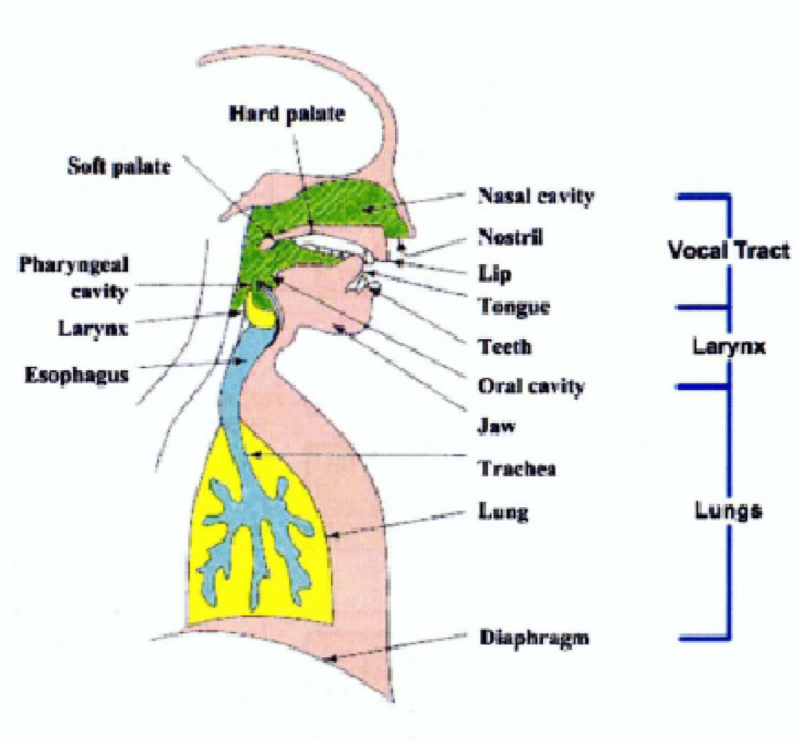


Figure 2.1 An anatomy of speech production [QUA02]

Fig 2.2(a) shows the block diagram of the physiological mechanism of speech production. The lungs and the associated muscles pump the air required to excite the vocal mechanism. The muscle force pushes air out of the lungs and through the trachea. The excitation used to generate speech can be classified into *voiced, unvoiced, mixed, plosive, whisper* and *silence* [DEL99], [QUA02]. When the vocal cords are tensed, the air flow causes them to vibrate, producing the so-called voiced speech sounds. When the vocal cords are relaxed, in order to produce a sound, the air flow must either pass through a constriction in the vocal tract, thereby becomes turbulent and produces the so-called unvoiced sounds, or it can build up pressure behind a point of the total closure within the vocal tract. When the closure is opened, the pressure is suddenly and abruptly released, causing a brief transient sound. Any combination of one or more of these actions can be blended to produce a particular type of sound.

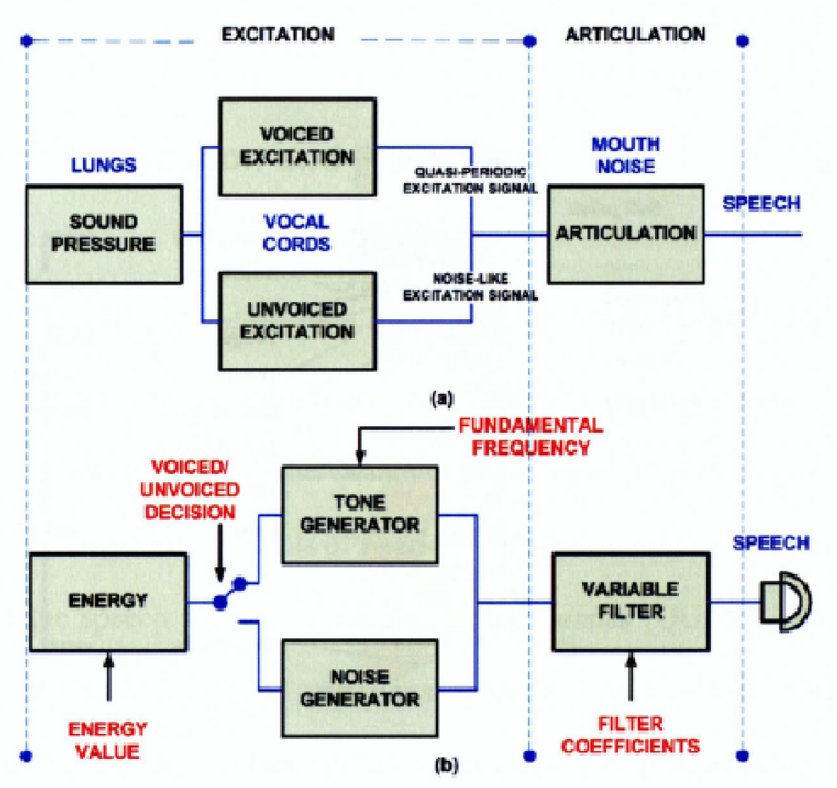


Figure 2.2 (a) the physiological mechanism of human speech production, (b) the engineering model of speech production [QUA02].

A *phoneme* describes the linguistic meaning conveyed by a particular sound of speech [DEL99], [QUA02]. The American English language consists of about 42 phonemes, which can be classified as vowels, semivowels, diphthongs and consonants (fricatives, nasals, affricatives and whisper) as shown in Fig 2.3 [DEL99], [QUA02]. Vowels are produced due to the periodic vibrations of the vocal chords in the larynx. The frequency at which the vocal chords vibrate is called the *fundamental frequency* or *pitch* of the speech [QUA02]. The fricatives are random noise-like sound caused by the turbulence of air as it passes through the narrow constrictions in the vocal tract. Nasals are caused by the acoustic coupling of the nasal cavity to the pharyngeal cavity by lowering the velum. By building up the pressure in front of the vocal tract and then abruptly releasing it produces plosives. The resonant frequencies generated by the vocal tract are called the *formant frequencies* or the *formants* [DEL99], [QUA02]. The formants depend on the length and shape of the vocal tract.

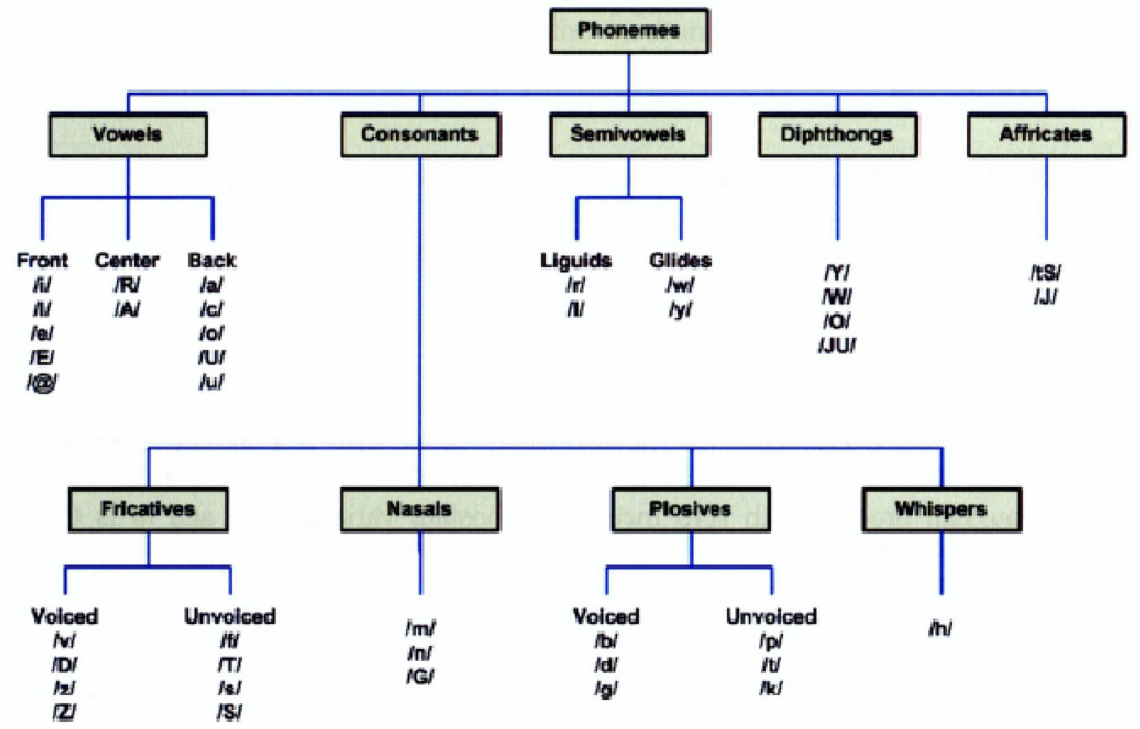


Figure 2.3 Phonemes in American English [QUA02]. Orthographic symbols are given in parentheses.

Fig 2.2(b) illustrates an engineering model of speech production. This source-filter model [QUA02] is at the heart of many speech analysis methods and it drives the research in speech perception too. The autoregressive model (AR) is used to model the speech signal. The motivation comes from visualizing the vocal tract as a lossless tube built of adjoining cylinders of different diameters. The principle put forward in the source filter model is that speech sounds are produced by the action of a filter, which is the vocal tract, on a sound source, with is either the glottis or some other constriction within the vocal tract. There are three general categories of the source for speech sounds, namely periodic, noisy and impulsive, even though combinations of these sources are often present [QUA02]. The AR model should be ideally driven by pulse train for voiced phonemes and by white noise for unvoiced ones. Thus, speech sounds are determined not only by the source, but also by different vocal tract configurations, and how these shapes combine with periodic, noisy and impulsive sources [QUA02].

To model the phoneme of the speech signal, an autoregressive function [HAY01], [VAS00], [QUA02] is defined by:

$$s[n]=\sum_{i=1}^{p}a_i s[n-i]+u[n] \tag{2.1}$$

where p is the linear predictive coefficients (LPC) order, $u[n]$ is a predictive error of a white Gaussian process with zero mean and constant variance σ_u^2, and a_i is the i^{th} autoregressive (AR) parameter.

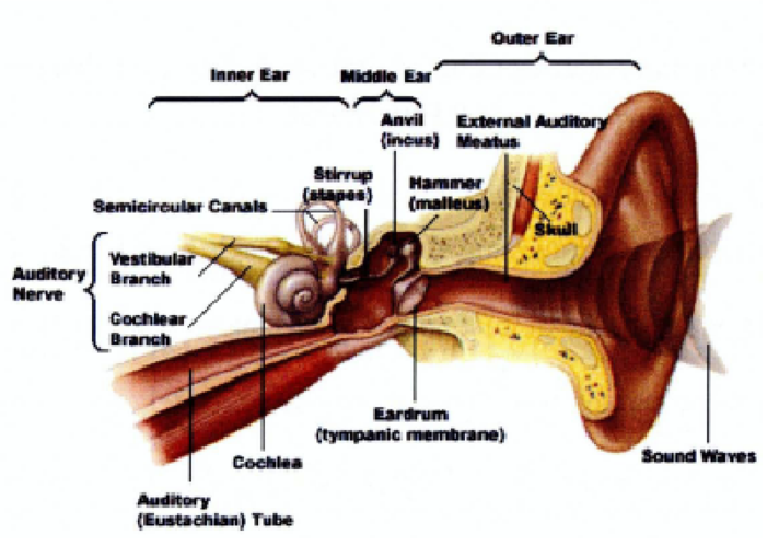


Figure 2.4 An anatomy of the human ear [YOS00].

2.1.2 The Human Hearing System

The human ear can be divided into three parts, the outer, middle and inner ear as shown in Fig 2.4. The outer ear includes the visible part of the ear (pinna) and the meatus (ear canal). The responsibility of the pinna is to collect sound waves and aid in sound localization before the sound is directed to the ear canal. The middle ear is an air-filled space separated from the outer ear by the eardrum. It contains three small bones (ossicles) that make up the ossicular chain. These bones connect the eardrum to the inner ear. The eardrum and ossicles convert sound energy received into mechanical energy transmitted to the inner ear. The inner ear consists of a liquid-filled tube called the cochlea, which has thousands of tiny nerve fibers. The mechanical action of the ossicular chain creates movement in the fluid that stimulates the nerve fibers. The nerve fibers then transmit the electrical pulses to the auditory nerve and to the brain, where these pulses are interpreted as sound. The interior of the cochlea can be subdivided by two membranes, the Reissner's membrane and the basilar membrane (BM) [ZWI90]. The organ of corti contains about 30000 sensory hair cells, which lies on the BM. These hair cells cause neural firings to propagate in the auditory nerve.

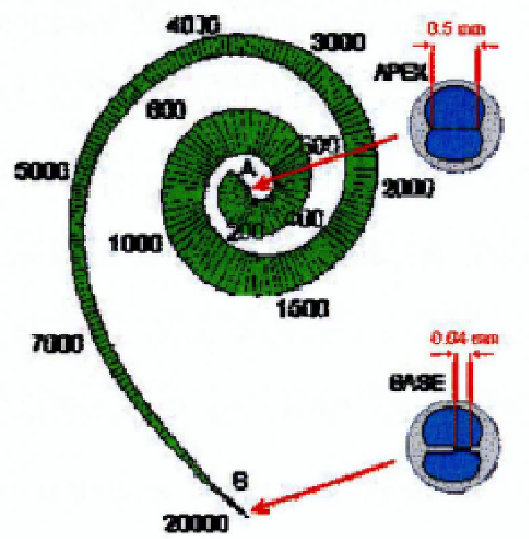


Figure 2.5 Frequency response of Basilar membrane [YOS00].

The BM within the cochlea of the inner ear is the part of the auditory system that decomposes incoming auditory signal into their frequency components. Fig 2.5 shows the characteristics of BM. It varies gradually in shape and stiffness along its length, and its frequency response also varies with its length. This induces input sound of a certain frequency to vibrate at a particular location of the membrane more than other locations. High frequency sound vibrates at the narrow, basal end of the membrane whereas low frequency sound vibrates at the wide, apical end of the membrane. The localized vibration of the BM is then transduced into neural signals by the inner hair cells of the organ of corti that sits on top of the BM. The hair cells above the most resonating segment of the membrane fire the most signals, thus the neural code leaving the ear through the auditory nerve is frequency coded. The intensity and spectrum determines if the sound can be heard, and the hearing threshold determines the perceptibility of the sound. Many perspectives of the auditory system can be replicated by the critical band analysis in the inner ear where frequency transformation occurs along the BM [ZWI90]. Experiments have shown that the multi-independent channel filterbank model can be approximately considered as the critical band behaviour of the BM. As a consequence of the highly nonlinear cochlear processes of BM vibration and neural firings, the perception of one sound is obscured by the presence of another in time and frequency. This phenomenon is called masking. There are two types of maskings effect [ZWI90]: one is frequency masking, the other is temporal masking (pre-masking and post masking). These phenomena have important implication on the modelling of human auditory perception. The BM also derives the characteristic of frequency selectivity. Selectivity is an ability of the auditory system to separate or resolve the components in a complex sound. In Chapter 3, a detailed description and modelling of critical band and masking properties will be presented.

2.2 Noise and Distortion

The focus of speech enhancement methods is noise reduction. Therefore, the modelling and removal of the effects of noise and distortion are the cornerstones of speech enhancement. Noise in speech processing refers to any unwanted sound that contaminates the original speech. Noise by nature is a stochastic signal. Distortion, on the other hand, is perceived as an undesirable change of the signal in a system. There are two types of distortion: one is the amplitude distortion and the other is the phase distortion [VAS00].

Depending on the frequency or temporal characteristics, noise can be classified into a number of categories as follows:

1. **Narrowband noise** is one that the acoustic energy is concentrated in a relatively narrow range of frequencies. The spectrum will generally show a localized "hump" or peak in amplitude. Narrowband sound may be superimposed on broadband sound. If the narrowband sound does not contain any significant discrete tones, the sound will generally lack a subjective quality of pitch or tonality. In general, the notch filters are widely adopted for removing narrowband noise [VAS00].

2. **White noise** is defined as an uncorrelated noise process with equal power at all frequencies. White noise is the generalized mean-square derivative of the Wiener

process. Let *v*(*t*) be the noise at time instance, *t*. *v*(*t*) is a continuous-time zero-mean white noise process with a variance of σ^2. The power spectrum of *v*(*t*) is defined as [VAS00]:

$$P_{vv}(f)=\int_{t=-\infty}^{\infty} E\left[v(t)v^{*}(t+\tau)\right]e^{-j2\pi ft}=\sigma^{2} \tag{2.2}$$

where $E[\cdot]$ is the statistical expectation operator. The superscript asterisk denotes the complex conjugate. $E[v(t)v^{*}(t+\tau)]$ is an autocorrelation function over the observation interval from t to $t+\tau$.

3. **Band-limited white noise** is a random signal with a flat power spectral density and a limited bandwidth. The spectrum of band-limited white noise with a bandwidth of B Hz is defined as [VAS00]:

$$P_{vv}(f)=\begin{cases}\sigma^{2}, & \text{if } |f|\le B\\ 0, & \text{otherwise}\end{cases} \tag{2.3}$$

4. **Colored noise** is noise that is not white. It does not have a flat spectrum. Theoretically, coloured noise can be generated by passing a white noise through a coloured channel. The statistical modelling of colored noise will be described in Chapter 5.

Noise and distortion are the main limiting factors in many practical applications such as the cellular mobile communication, hearing aid, speech recognition and so on. It is important to analyze the characteristics of noise and model the noise process in order to discriminate the signal from the noise. The structure for modelling both the temporal and spectral characteristics of noise can be explored to build a high-quality speech enhancement method.

The speech signal can be acquired from single or multiple channel sensors. Speech enhancement is made difficult particularly by the presence of additive noise. The enhancement effort is further complicated by the non-stationarity of the noise process. One microphone input (single channel) makes speech enhancement more challenging, as both speech and noise are present in the same channel. Separation of the two signals would require relatively good knowledge of the speech and noise models or the interfering signal is required to be present exclusively in a different frequency band than that of the speech signal. A costly solution to this problem is to use a dual microphone approach [AAR04], [HOY05], [NAN95]. Spatial analysis is beneficial in speech enhancement as it provides useful information about the signal. In such analysis, the noise source is assumed to be statistically independent and additive. This assumption is based on the fact that most environmental noise is typically additive in nature [HAY01]. The discussion in this research work will be limited to the single channel enhancement techniques, as these are the most common type of systems found in many applications. As mentioned in Chapter 1, the algorithms for speech enhancement in a single microphone system can generally be classified into two main categories: nonparametric signal processing method and model-based signal processing method [VAS00]. Two speech enhancement algorithms are investigated and proposed in this thesis. One is the subtractive-type speech enhancement algorithm, which belongs to nonparametric method [BER79], [BOL79], [HAN06], [HAS04], [KAM02], [LIN03], [SCA96], [SIM98], [VAS00], [VIR99], [WOO00]. The other is the adaptive filtering speech enhancement algorithm, which belongs to the model-based method [CHE05], [EPH92a], [GRA06], [VAS00]. In the following sections, these two types of methods will be reviewed.

2.3 Subtractive-type Speech Enhancement Methods

Here we dwell on the class of speech enhancement systems that capitalize on the short-time spectral amplitude (STSA) of speech signal for its perception. The fundamental idea is to use the STSA of the noisy speech input to recover an estimate of the STSA of the clean speech by removing those parts contaminated by the additive noise [EPH84].

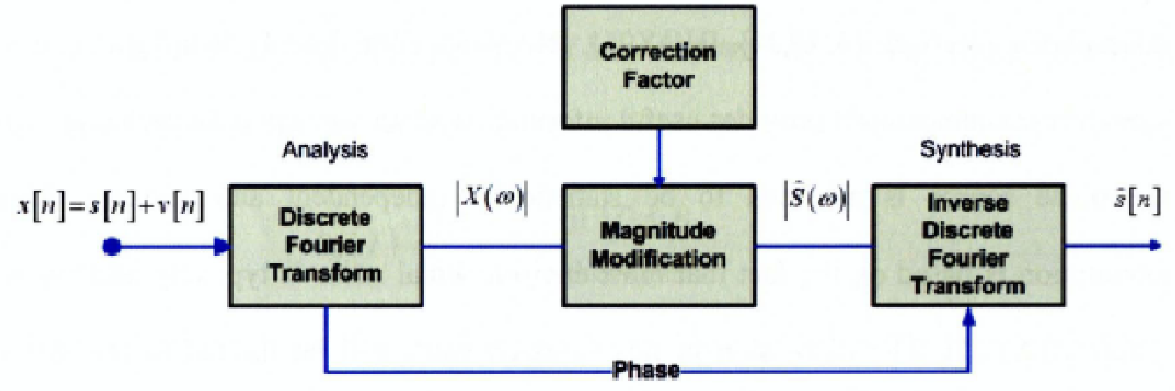


Figure 2.6 Block diagram of spectral subtraction.

Figure 2.6 illustrates the general representation of this technique. The system takes the noisy signal $x[n]$ as input, which is contaminated by ambient noise. Among many available methods for the analysis and synthesis, the Short-Term Fourier Transform (STFT) of the signal with 'OverLap and Add (OLA)' [DEL99] is the most commonly used method. The spectral amplitude, $|X(\omega)|$ of the noisy input signal, $x[n]$ is modified using a correction factor. The correction factor can be the spectral amplitude of the estimated noise signal, $v[n]$ measured and updated during periods of silence/non-activity in the speech signal. The error is corrected by subtracting the power spectrum or the magnitude spectrum of estimated noise from that of the noisy speech input. Hence, these methods are also regarded as subtraction algorithms. The assumption is that the noise is a stationary or a slowly varying process, and the speech signal is assumed to be uncorrelated with the noise. The corrected amplitude can be considered as an estimate $|\hat{S}(\omega)|$ of the original clean speech signal $s[n]$. For synthesizing the enhanced speech signal, an estimate of the instantaneous magnitude spectrum is combined with the phase of the noisy input signal under the assumption that the human ear is not able to perceive the phase distortions of the speech signal, and then it is transformed via an inverse discrete Fourier transform to the time domain [VAS00].

Spectral subtraction is a seminal noise reduction method based on the STSA estimation technique. The primary power spectral subtraction technique is popular due to its simple underlying concept and its effectiveness in enhancing speech degraded by additive noise. This method was first proposed by Boll [BOL79] in late seventies. Since then, several variations and enhancements have been made to overcome the drawbacks of this method. Given the chronicle significance of this method, the basic principle, drawbacks and improvements proposed over the years will be further elaborated.

2.3.1 Principle of the Spectral Subtraction Method

Consider $x[n]$ as the noisy input signal, which is composed of the clean speech signal, $s[n]$ and the uncorrelated additive noise signal, $v[n]$, then the noisy signal model in the time domain can be given by [BOL79]:

$$x[n]=s[n]+v[n] \tag{2.4}$$

In spectral subtraction, the incoming signal is processed frames-by-frame. The analysis of the overlapping frames of noisy signal is windowed using a Hamming window [DEL99], and then transformed via discrete Fourier transform (DFT) [DEL99]. The power spectrum of the noisy signal can be deduced as [BOL79]:

$$|X(f)|^2=|S(f)|^2+|V(f)|^2 \tag{2.5}$$

where the DFT of $X(f)$ is given by [VAS00]:

$$X(f)=\sum_{n=0}^{N-1}x[n]e^{-j\frac{2\pi fn}{N}}=|X(f)|e^{j\theta_X(f)} \tag{2.6}$$

where $\theta_X(f)$ is the phase of the noisy signal, $X(f)$.

In single microphone application, the noise spectrum $V(f)$ cannot be directly obtained. A time-averaged noise spectrum $\hat{V}(f)$ can be calculated from the periods when the signal is absent and only the noise is present. An estimate of the modified speech spectrum can be obtained by [BOL79]:

$$|\hat{S}(f)|^2=|X(f)|^2-|\hat{V}(f)|^2 \tag{2.7}$$

Due to the random variations of noise, some values of the modified spectrum calculated by Equation (2.7) may be negative. The magnitude and power spectrum are non-negative variables, and any negative estimates of these variables should be mapped into non-negative values. In this case, these values are set to zero. This process is called the half-wave rectification. With half-wave rectification, the modified spectrum can be deduced as [BOL79]:

$$\left|\hat{S}(f)\right|^2 = \begin{cases} \left|\hat{S}(f)\right|^2 & \text{if } \left|\hat{S}(f)\right|^2 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

To reconstruct the time domain signal, the modified spectrum of (2.8) is combined with the phase information from the noisy signal by using the inverse discrete Fourier transform (IDFT) in conjunction with the OLA method [BOL79].

$$\hat{s}[n] = \sum_{f=0}^{N-1} \left|\hat{S}(f)\right| e^{j\theta_X(f)} e^{j\frac{2\pi fn}{N}} \tag{2.9}$$

where $\theta_X(f)$ is the phase of the noisy signal, $X(f)$. The signal restoration of (2.9) is based on the assumption that the audible noise is mainly due to the distortion of the magnitude spectrum, and the phase distortion is largely inaudible.

The noise suppression can also be implemented as the product of the noisy signal spectrum and the frequency response of a spectral subtraction filter by rewriting the spectral subtraction method as [BOL79]:

$$\hat{S}(f) = H(f)X(f) \tag{2.10}$$

where $H(f)$ is a time-varying spectral subtraction filter represented by [BOL79]:

$$H(f) = \sqrt{1 - \frac{\left|\hat{V}(f)\right|^2}{\left|X(f)\right|^2}} \tag{2.11}$$

Here, the modified spectrum is obtained by applying a time varying weight, $H(f)$ to each frequency component. The spectral subtraction filter $H(f)$ is a zero-phase filter, with its magnitude response in the range $0 \leq H(f) \leq 1$. From (2.11), it can be deduced that the frequency dependent gain is a function of the signal-to-noise ratio (SNR) of each frequency component. The attenuation at each frequency increases and decreases with the SNR.

Through spectral subtraction processing, the enhanced signal has reduced the noise level significantly of the noise corrupted signal resulting in a better SNR and improved speech quality. Nevertheless, the subtraction process also produces a musical disturbance of unnatural quality noise called the musical noise. The intrinsic drawbacks of the method neutralize the perceivable improvement in speech quality. The success of spectral subtraction depends on the ability of the algorithm to reduce the noise variations and to remove the processing distortions.

2.3.2 Method Drawbacks of the Spectral Subtraction Method

While the spectral subtraction method is effective in reducing the additive noise present in the contaminated signal and is easily implementable, there exist some noticeable shortcomings, which are discussed below.

a) Residual Noise

The effectiveness of the noise removal process of the spectral subtraction method is dependent on the accurate spectral estimate of the noise signal. The better the noise estimate, the lesser the residual noise content in the modified spectrum [DEL99]. However, since the noise spectrum cannot be obtained directly, the average estimate of the noise spectrum can be obtained during silence period [VAS00]. Voice activity detection (VAD) is used to indicate the presence or absence of speech [VAS00]. There are some significant variations between the estimated noise spectrum and the actual noise content present in the instantaneous speech spectrum [DEL99]. Because of these random variations, the subtraction of these quantities results in the emergence of isolated residual noise levels of large variance. This residual spectral content manifests itself in the reconstructed time domain signal as varying tonal sounds [VAS00]. It is perceived as an annoying metallic sounding noise. This musical noise can be even more disturbing and annoying to the listener than the distortions due to the original noise content [KAM02].

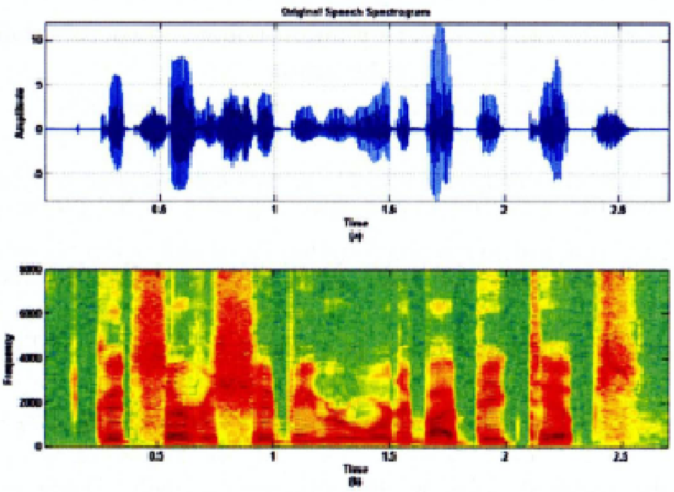


Figure 2.7 (a) Waveform and (b) spectrogram of clean speech

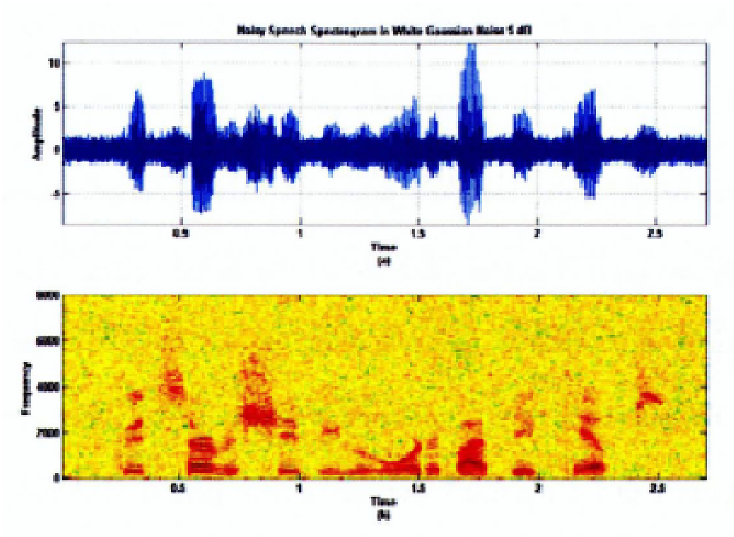


Figure 2.8 (a) Waveform and (b) spectrogram of noisy speech in white Gaussian noise of 5 dB SNR

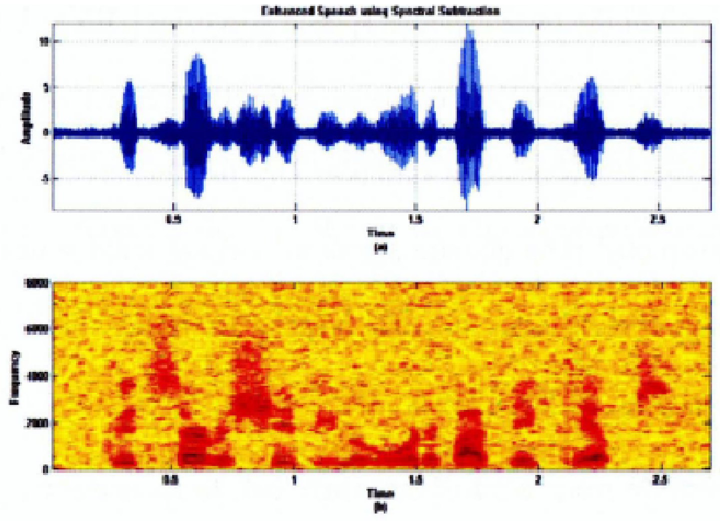


Figure 2.9 (a) Waveform and (b) spectrogram of enhanced speech obtained by spectral subtraction

Figs 2.7 to 2.9 show the waveforms and spectrograms of the clean speech, the noisy speech in white Gaussian noise of 5 dB SNR and the enhanced speech obtained by spectral subtraction. The residual noise is evident in the plot of Fig 2.9. Comparing with the noisy speech of Fig 2.8 and the enhanced results of Fig 2.9, some artifacts are observed due to the residual noise in the enhanced speech. These residual noises are characterized as short-lived narrow bands of frequencies surrounded by relatively low-level frequency components.

Several residual noise reduction algorithms have been proposed to combat this problem [BER79], [CHE91], [EPH84], [GOH98], [HU04], [MCA80], [PET81], [VAS00], [VIR99]. However, due to the limitations of the single-channel enhancement methods, it is not possible to remove this noise completely without compromising the quality of the enhanced speech. Hence there is a trade-off between the amount of noise to be reduced and speech distortion induced by its underlying process.

b) Distortions due to Non-linear Rectification

The modified speech spectrum obtained from (2.7) may contain some negative values due to the variations of the noise spectrum. In spectral subtraction, this non-linear rectification process is applied to map negative estimates of these variables into non-negative values. These values are rectified using half-wave rectification (set to zero) or full-wave rectification (set to its absolute value) [VAS00]. However, these rectifications can also distort the distribution of the enhanced signal because of the non-linear mapping of the negative and small-valued spectral estimates. This distortion is perceived as a metallic noise due to its narrowband spectrum and the tin-like sound.

c) Roughening of Speech due to the Noisy Phase

The phase of the noisy signal is directly used to combine with an estimate of the instantaneous magnitude spectrum to restore the enhanced time domain signal. This regenerating procedure is based on the fact that the presence of noise in the phase information does not contribute notably to the perception of the speech quality. This is especially true at high SNRs (>5 dB). However, at lower SNRs (<0 dB), the noisy phase can lead to a perceivable roughness in the speech signal, thus lowering the speech quality. This fact is corroborated experimentally by Schroeder [SCH75]. Estimating the phase of the clean speech is a difficult task and it will greatly increase the complexity of the method. Moreover, the audible noise is manly due to the distortion of the magnitude spectrum, and that the phase distortion is largely inaudible, especially for the high SNR cases. Hence the use of the noisy phase information is considered to be an acceptable practice in the reconstruction of the enhanced speech signal.

In summary, the spectral subtraction methods attempt to optimize the additive noise removal through subtraction of an estimate of the noise spectrum from the noisy signal spectrum. Speech communication refers to the processes associated with the production and perception of sounds used in spoken language. However, speech is heavily dependent on the frequency components and their correlation. Hence, while conventional speech enhancement algorithms can increase the quality of noisy speech by increasing the SNR, the speech intelligibility is not guaranteed. The spectral subtraction methods, as well as many other methods, suffer from this drawback.

2.3.3 Derivatives of Spectral Subtraction Method

Several variants of the spectral subtraction method originally proposed by Boll have been developed to address the musical noise problem of the basic technique [BER79], [CHE91], [EPH84], [GOH98], [HU04], [MCA80], [PET81], [VAS00], [VIR99]. Spectral subtraction based methods have also been extended to perform noise suppression in the autocorrelation, cepstral, logarithmic and sub-space domains [AHM89], [BRE07], [EPH95], [HER94], [HU02], [HU03], [JAB03], [MIT00], [REZ01], [ZHO05]. Some pre- and post-processing techniques applied before and after spectral subtraction have also helped to minimize the musical noise and speech distortion, [HOY05], [MA06], [NAN95]. This section scrutinizes some of these important enhancement techniques that have been proposed over the years.

a) Magnitude Averaging

Boll [BOL79] first introduced the magnitude averaging in his primary paper. The aim is to modify the spectral subtraction to smooth out the spectral content that contributes to residual noise. The principle of magnitude averaging of the input spectrum is to reduce spectral errors by averaging them across neighbouring frames. This has the effect of lowering the noise variance while reinforcing the speech spectral content and thus preventing destructive subtraction [DEL99]. The concomitant problem of magnitude averaging is the short-term stationarity of speech. Therefore, averaging is only allowed over a limited number of neighbouring frames. If this constraint is ignored, certain slurring of speech can be detected due to the temporal smearing of short transitory sounds [VAS00]. A generalized representation of the averaging operation can be expressed as [BOL79]:

$$\widehat{X}_i(f) = \frac{1}{2M+1}\sum_{j=-M}^{M} W_j X_{i-j}(f) \tag{2.12}$$

where i is the frame index. M is the frame number for the averaging operation. The weights W_j can be used to weigh the frames. When $W_j = 1\ \forall\, j$, the equation reduces to the basic magnitude averaging operation. In the case where the frames are weighted by different values of W_j, the operation is referred to as the weighted magnitude averaging. Inspired by Boll's method, Goh *et al.* [GOH98] identified them by using multi-blade median filtering over several frames of speech.

b) Generalized Spectral Subtraction

A generalized form of the basic spectral subtraction of (2.7) is proposed by Berouti *et al.* [BER79] as follows:

$$\left|\widehat{S}(f)\right|^{\gamma} = \left|X(f)\right|^{\gamma} - \left|\widehat{V}(f)\right|^{\gamma} \tag{2.13}$$

where the exponent γ can be chosen to optimize certain performance metric. When $\gamma = 2$, the subtraction is carried out on the Short-term power density spectra and is referred to as the power spectral subtraction. When $\gamma = 1$, the equation reduces to the basic spectral subtraction method proposed by Boll, the subtraction is performed on the magnitude spectra.

c) Spectral Subtraction using Over-subtraction and Spectral-flooring

Berouti *et al.* [BER79] proposed an *oversubtraction* of noise spectrum to reduce the overall residual noise level and an inclusion of a fraction of noise as a spectral flooring to mask the musical tone artifacts. A couple of parameters have been introduced into the power spectral subtraction equation as follows:

$$\left|\hat{S}(f)\right|^2 = \left|X(f)\right|^2 - \alpha\left|\hat{V}(f)\right|^2 \tag{2.14}$$

$$\left|\tilde{S}(f)\right|^2 = \begin{cases} \left|\hat{S}(f)\right|^2 & \text{if } \left|\hat{S}(f)\right|^2 > \beta\left|\hat{V}(f)\right|^2 \\ \beta\left|\hat{V}(f)\right|^2 & \text{otherwise} \end{cases} \tag{2.15}$$

where α is the over-subtraction factor, and β is the spectral floor. α is a function of the noisy signal-to-noise ratio and is calculated as [BER79]:

$$\alpha = \alpha_0 - \frac{3}{20}SNR, \quad -5dB \le SNR \le 20dB \tag{2.16}$$

where α_0 is the desired value of α at 0 dB SNR. The over-subtraction factor, which can be seen as a time-varying factor, attenuates the power spectrum of the noise more than necessary. This leads to a reduction of residual noise peaks but also to an increased audible distortion. It provides a degree of control over the noise removal process between periods of noise update. The spectral floor factor determines the minimum value taken by the spectral components of the enhanced spectrum, $\beta\left|\hat{V}(f)\right|^2$. This operation fills out the valleys between spectral peaks. The neighbouring residual noise components are masked by the addition of background noise into the spectrum. This leads to a reduction of residual noise but an increased level of background noise remains in the enhanced speech. The values of these parameters were usually empirically adjusted to improve the performance based on some subjective listening tests [LIM79], [VIR99]. To a large extent, the proposed technique has proved to be successful in suppressing the residual noise, but over-subtraction of the noise estimate also causes severe speech distortions.

d) Spectral subtraction with an MMSE STSA estimator

Ephraim and Malah [EPH84] derived an approach, which optimally estimate the short-time spectral amplitude and complex exponential of the phase of the speech signal under the minimum mean-square error criterion (MMSE). This MMSE STSA estimator took into account the uncertainty of the signal present in the noisy spectral components. In this method, a gain function was calculated based on the *a priori* and *a posterior* SNRs, and the variance of each noise spectral component. The method can be described by the following equations.

$$\hat{S}(f)=H(f)X(f) \tag{2.17}$$

$$H(f)=\frac{\sqrt{\pi}}{2}\sqrt{\frac{1}{\lambda_v}\cdot\frac{SNR_{priori}}{1+SNR_{priori}}}\cdot F\left(SNR_{post}\cdot\frac{SNR_{priori}}{1+SNR_{priori}}\right) \tag{2.18}$$

where λ_v is the variance of the spectral component of the noise. SNR_{priori} is the *a priori* SNR, which is calculated as:

$$SNR_{priori,i}(f)=0.98\cdot\left(\frac{\left|\hat{S}_{i-1}(f)\right|^2}{\left|\hat{V}_i(f)\right|^2}\right)+(1-0.98)\cdot P\left(SNR_{post,i}-1\right) \tag{2.19}$$

Here, i is the frame index with $P(x) = x$ if $x \geq 0$ and $P(x) = 0$ otherwise. SNR_{post} is *a posterior* SNR and the function, F is given by:

$$F(x) = e^{-\frac{x}{2}}\left[(1+x)I_0\left(\frac{x}{2}\right)+I_1\left(\frac{x}{2}\right)\right] \tag{2.20}$$

where $I_0(y)$ and $I_1(y)$ are zero and first order modified Bessel functions, respectively.

The MMSE STSA estimator depends on the parameters of the statistical model. The *priori* SNR was found to be a key parameter of the estimator. Unlike magnitude averaging where the averaging is performed irrespective of whether the frame contains speech or noise, the MMSE STSA estimator performs non-linear smoothing. It results in significantly less mean square error and bias when the frame contains noise predominantly. The residual noise due to this technique is observed to be colorless. The averaging effectively reduces the distortions in the speech part.

e) Spectral subtraction based on perceptual properties

Conventional single channel subtraction algorithms are characterized by a tradeoff between the amount of noise reduction, the speech distortion, and the level of musical residual noise, which can be modified by varying the subtraction parameters. While these methods improve the speech quality of the noisy speech by increasing the SNR, there is no significant improvement in speech intelligibility due to the quasi-stationary nature and other subtle properties of speech. They are also usually limited to the use of fixed optimized parameters, which are difficult to choose for all speech and noise conditions. To solve this problem, attempts have been made to incorporate the knowledge of human perceptual properties into the denoising process. Methods based on the perceptual loudness [PET81] and lateral inhibition [CHE91] showed some success in preserving the speech content. Virag [VIR99] proposed a technique based on the masking properties of the human auditory system. A masking threshold is calculated by modeling the frequency selectivity of human ear and the masking property, which is the property that weak sounds are masked by simultaneously occurring stronger sounds. It paves an avenue for the automatic adaptation in time and frequency by parametric enhancement system, where the tradeoff is determined by a criterion that is correlated with human perception.

Using the spectral subtraction technique given by (2.10), the gain function is calculated as:

$$G(f)=\begin{cases}\left(1-\alpha\left[\dfrac{|\hat{V}(f)|}{|X(f)|}\right]^{\gamma_1}\right)^{\gamma_2} & \text{if } \left[\dfrac{|\hat{V}(f)|}{|X(f)|}\right]^{\gamma_1}<\dfrac{1}{\alpha+\beta}\\ \left(\beta\left[\dfrac{|\hat{V}(f)|}{|X(f)|}\right]^{\gamma_1}\right)^{\gamma_2} & \text{otherwise}\end{cases} \tag{2.21}$$

where the over subtraction factor α, and the spectral floor β, is a function of the masking threshold, $T(f)$. The range of over-subtraction factor α is $\alpha > 1$, and the range of spectral flooring β is $0 < \beta < 1$. The exponents γ_1 and γ_2 determine the sharpness of the transition of $G(f)$. The masking threshold, $T(f)$ is calculated by applying a spreading function across the critical bands of the speech spectrum. It will be discussed in detail in Chapter 3.

Kim *et al.* [KIM00] proposed a speech state-dependent spectral subtraction method. The method employed a modified subtraction rule by incorporating the acoustic characteristics of phonemes with residual noise reduction using the masking threshold. While these methods have led to a significant reduction of the unnatural residual noise and improved speech quality over those that are based on the pure mathematical models of speech and noise signals, their implementation complexity has also increased substantially.

f) Frequency-dependent Spectral Subtraction Methods

Most implementations and variations of the basic subtraction technique advocate the subtraction of noise spectrum estimate over the entire speech spectrum. However, real world noise is mostly colored and does not affect the speech signal uniformly over the entire spectrum. In recent years, researchers have turned to the frequency adaptive subtraction factor based on the segmental noisy SNR [LOC92], [HE99], [WU01], [KAM02]. The frequency-dependent spectral subtraction approach considers the fact that colored noise affects the speech spectrum differently at various frequencies.

Lockwood and Boudy [LOC92] proposed a non-linear spectral subtraction (NSS) method that utilizes the estimates of local SNRs. The over-subtraction factor is frequency dependent within every frame of speech input. In the case that the speech signal may be lost out to noise, over-subtraction followed by non-linear processing of the negative estimates results in a higher overall attenuation of the noise. Hence, the subtraction is non-linear over the range of frequencies in the spectrum. The enhanced speech spectrum can be expressed as [LOC92]:

$$\left|\hat{S}_i(f)\right| = \left|X_i(f)\right| - \frac{\alpha_i(f)}{1+\varphi\cdot\rho_i(f)} \tag{2.22}$$

where i is the frame index. $X_i(f)$ is the smoothed noisy speech spectrum of the i-th frame, and $\alpha_i(f)$ is the frequency-dependent overestimation factor calculated over M frames as:

$$\alpha_i(f) = \max_{i-M\leq j\leq i}\left(\left|\hat{V}_j(f)\right|\right) \tag{2.23}$$

It can be approximated to

$$\alpha_i(f) = 1.5\cdot\left|\hat{V}_i(f)\right| \tag{2.24}$$

The scaling factor, φ of (2.22) is dependent on the variation of the frequency-dependent SNR, $\rho_i(f)$, which is given by [LOC92]:

$$\rho_i(f) = \frac{\left|X_i(f)\right|}{\left|\hat{V}_i(f)\right|} \tag{2.25}$$

The subtracting term in (2.22) is bounded to reduce large variation in the modified spectrum. The bounds are given by [LOC92]:

$$\left|\hat{V}_i(f)\right| \leq \frac{\alpha_i(f)}{1+\varphi\cdot\rho_i(f)} \leq 3\left|\hat{V}_i(f)\right| \tag{2.26}$$

To avoid negative values in the enhanced spectrum, spectral flooring is applied as follows:

$$\left|\hat{S}_i(f)\right| = \begin{cases} \left|\hat{S}_i(f)\right| & \text{if } \left|\hat{S}_i(f)\right| \geq \beta\left|\hat{V}_i(f)\right| \\ \beta\left|X_i(f)\right| & \text{otherwise} \end{cases} \tag{2.27}$$

where the typical value of β is 0.1, usually the lower bound $\beta \geq 0.01$.

The algorithm demonstrates that overall, frequency-dependent processing can be used to suppress musical noise to achieve better speech quality. The performance improvement resulted from calculating the optimal over-subtraction value for each frequency in the frame depending on the SNR. However, the drawback of this algorithm is that large variations may exist between neighbouring frequency components due to errors in the noise estimate [VAS00].

Other approaches based on frequency-dependent subtraction have also been proposed. He and Zweig [HE99] proposed a two-band spectral subtraction method for the lower frequency band and weighted magnitude averaging for the higher frequency band, which is considered to be stochastic in nature. The cut-off frequency between the two bands is the highest frequency below which the separation between adjacent peaks is approximately equal to the fundamental frequency. This cut-off frequency is determined adaptively for each frame. Another method proposed by Wu and Chen [WU01] uses the spectral subtraction method of Berouti et al. [BER79] on each critical band over the speech spectrum. Kamath and Loizou [KAM02] also proposed a multi-band spectral subtraction approach. Their method accounted for the fact that colored noise affects the speech spectrum differently at various frequencies.

g) Wavelet based Spectral Subtraction Methods

In order to simplify the mapping of audio signals into a scale that could well preserve the time-frequency related information, wavelet packet transform has been incorporated recently in some proposed speech and audio coding systems. Sinha and Tewfik [SIN93] first introduced adaptive wavelets to transparent audio compression. Srinivasan and Jamieson [SRI98] proposed a high-quality audio compression using the adaptive wavelet packet decomposition and psychoacoustic modeling.

Based on the principle of subtractive type of speech enhancement algorithm, Li et al. [LI01] proposed the perceptual time-frequency subtraction algorithm for noise reduction in hearing aids. Motivated by the generalized spectral subtraction method of [VIR99], Lu and Wang [LU04] applied the critical band wavelet packet transform to speech enhancement. The phenomenal success met by these methods has shown that the subtraction proceeded in the wavelet domain is a valid line of research. However, the quality enhancement of the processed speech of the above methods diminishes as more processing distortion is introduced and when the high frequency speech signals are grievously reduced.

Several other speech enhancement methods based on wavelets have been developed. Bahoura and Rouat [BAH01] proposed modified wavelet speech enhancement method using the teager energy operator. This technique does not require an explicit estimation of the noise level or *a priori* knowledge of the SNR, which is usually needed in many popular speech enhancement methods. Veselinovic and Graupe [VES03] introduced a wavelet transform approach to the blind adaptive filtering of speech from unknown noises. It can reduce the noise effect well, but the complexity is high. Ayat *et al.* [AYA04] proposed a wavelet based speech enhancement method using the step-garrote thresholding algorithm.

2.4 Model-based Speech Enhancement Algorithm

The category of model-based speech enhancement algorithms is also called the statistical model based method [EPH92a], which utilizes a parametric model of the speech signal generation process, such as autoregressive-moving average (ARMA) and autoregressive (AR) processes [HAY01]. This process involves the estimation of the speech model parameters from the noisy observations, and then applies an optimal filtering, such as Wiener or Kalman filter, to obtain an enhanced speech signal. The model-based speech enhancement method is illustrated in Fig. 2.10.

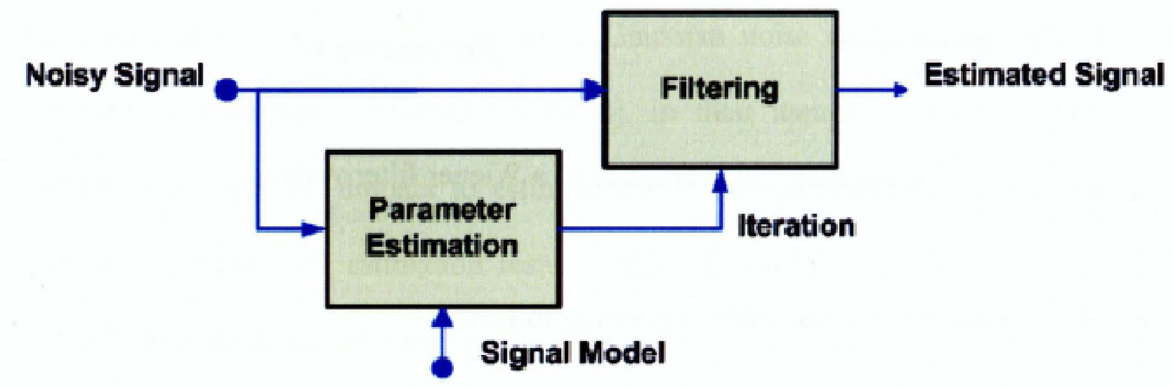


Figure 2.10 Block diagram of model-based speech enhancement

2.4.1 Wiener Filtering Speech Enhancement Method

Early studies of model-based methods have focused mainly on Wiener filtering, which is a popular adaptive technique that has been used in many enhancement methods [HAY01], [VAS00]. Fig. 2.11 illustrates a Wiener filter, which takes noisy speech, $x[n]$ as input, and produces an output signal $\hat{s}[n]$. The filter is represented by the coefficient vector, $\mathbf{w}$, which is calculated to minimize the Mean Square Error (MSE) between the desired signal, $s[n]$ and the estimated signal, $\hat{s}[n]$.

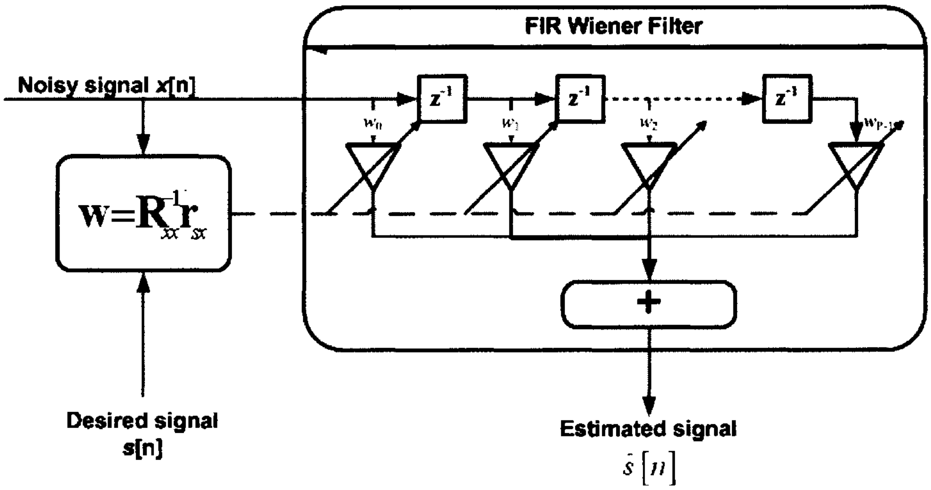


Figure 2.11 Block diagram of a Wiener filter structure

The filter input-output relationship is given as follows:

$$\hat{s}[n] = \sum_{k=0}^{P-1} w_k x[n-k] = \mathbf{w}^T \mathbf{x} \tag{2.28}$$

where $\mathbf{w}$ is the wiener filter coefficient vector. The Wiener solution is obtained by calculating the gradient of the minimum mean square error function with respect to the filter coefficient vector.

$$\mathbf{w} = \mathbf{R}_{xx}^{-1}\mathbf{r}_{xs} \tag{2.29}$$

From (2.29), the calculation of the Wiener filter coefficients requires the autocorrelation matrix of the noisy input signal, $\boldsymbol{R}_{xx}$, and the cross-correlation vector of the input and the desired signals, $\boldsymbol{r}_{xs}$.

The Wiener filter can be expressed in the frequency domain by:

$$H(f) = \frac{E\left[\left|X(f)\right|^2\right] - E\left[\left|\hat{V}(f)\right|^2\right]}{E\left[\left|X(f)\right|^2\right]} \tag{2.30}$$

From (2.30), it is obvious that *a priori* knowledge of the speech and noise power spectra is necessary. The speech power spectrum is estimated using the estimated speech model parameters.

An enhancement algorithm for the white Gaussian noise contaminated speech was proposed by Lim and Oppenheim [LIM78]. In their approach, Wiener filtering is applied in the spectral domain to estimate the speech parameters. The procedure is iterated to enhance the estimation results. This method is extended by Feder *et al.* [FED89] to a dual-microphone system. The parameter estimation procedure becomes more complicated and it is formulated as a maximum likelihood (ML) problem. Due to the complexity and the difficulty of the ML solution, expectation maximization (EM) algorithm [DEM77] is used to approximate the solution. Lowerre [LOW91] further extended the EM algorithm in frequency domain to time domain in noise cancellation. However, these algorithms need block processing due to the requirement of Wiener theory.

2.4.2 Kalman Filtering Speech Enhancement Method

An alternative to Wiener filter is Kalman filter. Kalman filter is a recursive least square error method for the estimation of a signal. Kalman filter theory is based on a state-space approach in which a state equation models the dynamics of the signal process and an observation equation models the noisy observation signal. The advantage of employing Kalman filter in speech enhancement algorithms is its sequential processing and adaptation to the nonstationary characteristics of the speech signal.

Table 2.1 Summary of Kalman Variables and Parameters

Variable	Definition	Dimension
$\boldsymbol{s}[n]$	State at time n	M-by-1
$\boldsymbol{x}[n]$	Observation at time n	N-by-1
$\mathbf{F}[n+1,n]$	Transition matrix from time n to time $n+1$	M-by-M
$\mathbf{C}[n]$	Measurement matrix at time n	N-by-M
$\mathbf{Q}_\mathrm{e}[n]$	Correlation matrix of process noise $\boldsymbol{e}[n]$	M-by-M
$\mathbf{Q}_\mathrm{V}[n]$	Correlation matrix of measurement noise $\boldsymbol{v}[n]$	N-by-N
$\hat{\mathrm{s}}[n\mid x_{n-1}]$	Predicted estimate of the state at time n given the observations $x[1]$, $x[2]$, ..., $x[n$-1]	M-by-1
$\hat{\mathrm{s}}[n\mid x_{n}]$	Filtered estimate of the state at time n given the observations $x[1]$, $x[2]$, ..., $x[n]$	M-by-1
$\mathbf{G}[n]$	Kalman gain at time n	M-by-N
$\boldsymbol{a}[n]$	Innovations vector at time n	N-by-1
$\mathbf{R}[n]$	Correlation matrix of the innovations vector $\boldsymbol{a}[n]$	N-by-N
$\mathbf{K}[n-1, n]$	Correlation matrix of the error in $\hat{\mathrm{s}}[n\mid x_{n-1}]$	M-by-M
$\mathbf{K}[n]$	Correlation matrix of the error in $\hat{\mathrm{s}}[n\mid x_{n}]$	M-by-M

In general, for a clean speech signal, $s[n]$ and a noisy input signal, $x[n]$, the state equation model and the observation model are defined as:

$$\mathbf{s}[n+1]=\mathbf{F}[n+1,n]\mathbf{s}[n]+\mathbf{e}[n] \tag{2.31}$$

$$\mathbf{x}[n]=\mathbf{C}[n]\mathbf{s}[n]+\mathbf{v}[n] \tag{2.32}$$

where n is the discrete-time sampling index. The definition of the Kalman variables and parameters are given in Table 2.1 and its computations are summarized in Table 2.2 [HAY01].

Table 2.2 Kalman Filtering Process

Input vector process: Observations = $\{y[1], y[2], \ldots, y[n]\}$ Known parameters: Transition matrix = $\mathbf{F}[n+1, n]$ Measurement matrix = $\mathbf{C}[n]$ Correlation matrix of process noise = $\mathbf{Q}_e[n]$ Correlation matrix of measurement noise = $\mathbf{Q}_v[n]$ Computation: $\mathbf{G}[n]=\mathbf{F}[n+1,n]\mathbf{K}[n,n-1]\mathbf{C}^H[n]\left[\mathbf{C}[n]\mathbf{K}[n,n-1]\mathbf{C}^H[n]+\mathbf{Q}_v[n]\right]^{-1}$ $\mathbf{a}[n]=\mathbf{x}[n]-\mathbf{C}[n]\hat{\mathbf{s}}[n \mid x_{n-1}]$ $\hat{\mathbf{s}}[n+1 \mid x_n]=\mathbf{F}[n+1,n]\hat{\mathbf{s}}[n \mid x_{n-1}]+\mathbf{G}[n]\mathbf{a}[n]$ $\mathbf{K}[n]=\mathbf{K}[n,n-1]-\mathbf{F}[n,n+1]\mathbf{G}[n]\mathbf{C}[n]\mathbf{K}[n,n-1]$ $\mathbf{K}[n+1,n]=\mathbf{F}[n+1,n]\mathbf{K}[n]\mathbf{F}^H[n+1,n]+\mathbf{Q}_e[n]$ Initial conditions: $\hat{\mathbf{s}}[1 \mid x_0]=E[\mathbf{s}[1]]$ $\mathbf{K}[1,0]=E\left[(\mathbf{s}[1]-E[\mathbf{s}[1]])(\mathbf{s}[1]-E[\mathbf{s}[1]])^H\right]$
Note: the superscript H indicates Hermitian matrix.

The application of Kalman filter to speech enhancement was first proposed by Paliwal and Basu [PAL87], with the speech parameters obtained from the clean speech signal and the noise characteristics obtained from the non-speech frames. Their method demonstrated the viability of Kalman filter to outperform Wiener filtering for the enhancement of speech signal corrupted by white Gaussian noise [LIM78]. Since then, there has been a flurry of research interest and development activities on the application of Kalman filter theory to speech enhancement problems [GIB91], [GRA06], [LEE00], [MA06], [POP98], [PAL87], [WU98]. Motivated by [LIM78], Gibson *et al.* [GIB91] proposed the first Kalman filter that used EM-based iterative estimation for the reduction of colored noise interference. Kalman filtering of colored noise for speech enhancement was also proposed by Popescu *et al.* [POP98]. The speech and colored noise are modelled as AR processes excited by white Gaussian noise. This Kalman filtering speech enhancement for colored noise performs even better than Wiener filter and the traditional Kalman filter for white Gaussian noise. However, the standard algorithm for Kalman filtering involves multiplications of very large matrices and has high computational cost. The computational complexity is reduced by dividing the signal into subbands using a quadrature mirror filterbank and then applying the low-order Kalman filter in each subband [WU98]. Multiple Kalman filters are also used with EM algorithm in the time domain to enhance speech in nonstationary noise [LEE00]. The most recent emphasis of Kalman filter based speech enhancement methods is on the exploitation of human perception. Grancharov *et al.* [GRA06] showed that the causal Kalman algorithm is in conflict with the basic properties of human perception and proposed a weighted Kalman filter solution to improve the perceptual quality, while preserving its efficient structure. A Kalman filtering speech enhancement method constrained by a human auditory system masking threshold was also proposed in [MA06].

2.5 Robust Speech Recognition

To evaluate the speech enhancement methods proposed in this thesis, a framework of techniques based on denoised features are presented. These techniques are applied to improve automatic speech recognition (ASR) performance in the presence of background noise. Some of the state-of-the-art speech recognizers and feature extraction methods are reviewed here.

The task of the recognizer is to use the feature vectors provided by the feature extractor to assign the object to a proper category. Many researchers are concerned with the design of good recognizer to build robust speech recognition system. Statistical approach based on Bayesian rule [DUD01] is widely used in modern speech recognition systems. Typically, the acoustic modeling components of a speech recognizer are based on the hidden Markov models (HMM) [DUD01], [MOO97]. A HMM uses the probability distribution associated with each state to model the temporal variability that occurs in speech across speakers or phonetic context via an underlying Markov process. This distribution is typically a Gaussian mixture model (GMM). A GMM provides a sufficiently general parametric model as well as an efficient and robust mathematical framework for speech estimation and analysis. Widespread use of HMMs for modeling speech can be attributed to the availability of efficient parameter estimation procedures [DUD01] to maximize the likelihood (ML) of data given the model. The expectation-maximization (EM) [DUD01] algorithm provides an iterative framework for ML estimation with good convergence properties. There are, however, problems with ML formulation for speech recognition. In some cases, the ML training of a Gaussian model will never achieve good classification. By learning the decision regions discriminatively can improve the classification performance. The discriminative approaches are a key ingredient for creating robust and more accurate models. Many promising methods [BOU94], [DUD01] have been introduced for using discriminative techniques to improve the estimation of HMM parameters. One such technique is the artificial neural network (ANN) [DUD01], which represents an interesting and important class of discriminative techniques that have been successfully applied to speech recognition. Although ANN attempts to overcome many of the problems previously described, there are apparent shortcomings. Some of the most notable deficiencies include the difficulty to design optimal model topologies, slow convergence during training, and tendency to overfit data. The other is the support vector machine (SVM) [CRI00], [DUD01], [VAP98]. SVM is a new universal learning machine, which uses a device called kernel mapping to map the data in the input space to a high-dimensional feature space wherein the problem becomes linearly separable. The decision function of an SVM is related not only to the number of support vectors (SVs) and their weights but also to the selected kernel called the support vector kernel [CRI00]. Many different kinds of kernels can be used, such as the linear, Gaussian and polynomial kernels. Currently, SVM has been widely used in speech recognition applications [LIN05], [RAM06], [ZHA04].

The traditional goal of the feature extractor is to characterize an object to be recognized by measurements whose values are very similar for objects in the same category, and very different for objects in different categories. This brings the idea of looking for distinguishing features that are invariant to irrelevant transformations of the input. In the beginning, these attributes primarily include the vocal tract spectrum envelop, such as linear predictive coefficients (LPC) [DEL99], and to a lesser extent, instantaneous pitch and glottal flow excitation, as well as temporal properties such as source event onset times and modulations in formant trajectories [DEL99], [QUA02]. To reduce the sensitivity of ASR to speech variability and adverse noise effects, robust features are also developed by considering the modus of phonetic perception. The excitation level that the follicles in the cochlea experienced is differentiated by the amplitude of speech.

The number of pressure changes occur in a given period of time is perceived as the pitch. The less intuitive perception is the time of occurrence of each amplitude and frequency pair. Davies and Mermelstein [DAV80] introduced the mel-cepstrum, which exploits auditory principles, as well as the decorrelating property of the cepstrum. The mel-frequency cepstral coefficients (MFCC) obtained good performance when they are used with HMM recognizer [KAR01], [VER99]. The relative spectral perceptual linear prediction (Rasta-PLP) feature is a term used to describe a range of speech recognition front end algorithms developed by Hermansky *et al.* [HER94]. It has been successfully applied to large vocabulary hybrid HMM-GMM and HMM-MLP methods [HER94], [PUJ05]. A good domain for speech feature extraction and representation is therefore the wavelet transform [STR96] as it has been known to be effective for the time–frequency analysis of non-stationary and quasi-stationary signals. Unlike other features such as LPC, MFCC, and RASTA-PLP, wavelet coefficients [FAR04] provide flexible and efficient manipulation of speech signal localized in the time-frequency plane.

2.6 Chapter Summary

In this chapter, background knowledge related to the speech production and perception, as well as a comprehensive literature survey on speech enhancement and speech recognition algorithms have been provided.

This chapter has described qualitatively the main functions of the speech production and perception mechanisms and the associated anatomy. Articulatory and acoustic descriptors of speech sounds were given. Based on these features, a brief introduction of phonemics was presented. Some implications of sound production and perception mechanisms for signal processing algorithms were discussed.

Besides, the approaches to reduce additive noise in a speech signal over a single channel have been studied. We investigated techniques of filtering one spectral slice at a particular time instant, with the view of STFT. These techniques include spectral subtraction, Wiener filtering, Kalman filtering, and optimal spectral magnitude estimation. The fundamentals of speech enhancement have been introduced. The spectral subtraction and its variants have been reviewed. The adaptive filtering speech enhancement approaches have also been presented. To obtain an enhanced speech signal, the adaptive process involves the estimation of the speech model parameters from the noisy observations and the application of an optimal filtering, such as Wiener or Kalman filter. Wiener filter, Kalman filter and their related methods were discussed. We concluded from the literature review that the Kalman filter method is more suitable for the reduction of white and colored noise. We have also observed that in spite of the significant improvements in reducing signal distortion and residual artifacts, improved intelligibility of the enhanced waveform for the human listener remains elusive. The algorithms to solve this problem have been studied and this topic will be further discussed in Chapters 4 and 5, where two new speech enhancement algorithms are proposed.

The literature survey also briefly covers a few existing algorithms on recognizers for automatic speech recognition. The recognizers based on hidden Markov model, artificial neural network, and support vector machines, as well as the different types of training methods have been presented. Some popular features used for ASR have also been introduced. To build a robust ASR system, the hybrid recognizer and several feature extraction approaches will be evaluated and discussed in Chapter 6.

CHAPTER 3.

WAVELET FILTERBANK FOR PSYCHOACOUSTIC MODELING OF HUMAN AUDITORY SYSTEM

3.1 Introduction

A great deal of research has been done to build a good psychoacoustic model [GOL94], [LOI98], [ZHE99], [ZWI90], which is the backbone of a quality speech enhancement system. In the past few decades, good progress has been made in understanding how the human auditory system processes sound in general and how it suppresses noise in particular [LI01], [VIR99]. It is found that the human auditory system resembles a powerful speech signal processor endowed with an immense capability to resolve both the spectral and temporal features of sound simultaneously. These unveiled characteristics of human auditory system, such as critical bands, and masking properties, *etc.*, have been widely applied in speech quality evaluation [KAR06], speech coding [BEN93], [JOH88], [SCH79], [SRI98], and speech recognition [CHO07], [WAN04].

In this chapter, the preprocessing of sound in the peripheral auditory system is addressed. The peripheral auditory system is composed of the outer, middle, and inner ear [YOS00]. This peripheral preprocessing structure is characterized by nonlinearities [ZWI90]. In other words, the information received by the auditory system can be described as nonlinear responses to frequency, whose selectivity is carried out by the inner ear. The sound is transferred to and concentrated at different places along the basilar membrane (BM) corresponding to its different frequencies [MOL73]. The BM is the part of the peripheral auditory system that decomposes incoming auditory signal into their frequency components in critical band scale. The concept of critical bands was proposed by Fletcher [FLE40], who laid the foundation that the peripheral auditory system behaves as a bank of overlapping bandpass filters having relatively constant Q values through psychoacoustic tests [PET83]. To model the behavior of the BM in the inner ear, these auditory filters act as frequency weighting functions across the critical bands, as explained in Chapter 2, Section 2.1.2. The short-time Fourier transform (STFT) [DEL99] is widely used to analysis the critical band filter responses and to realize the frequency selective properties of the peripheral auditory system. A STFT-based algorithm uses overlapping windows of a fixed duration to perform the spectral estimation by fast Fourier transform (FFT) algorithm [DEL99]. The time and frequency resolutions are fixed by the window length and the same FFT are applied across the entire time–frequency range. Therefore, it is insensitive to the varying features of transient signal [MAL99], [MER99]. Masking occurs because the auditory system is incapable of distinguishing two signals close in the time or frequency domain. Masking is manifested by a shift of auditory threshold in signal detectability and it plays an important role in determining the characteristics of critical bands [ZWI90]. Typically, only the simultaneous masking is considered in the STFT-based algorithms [JOH88], [SCH79], [VIR99], [HU04]. In a nutshell, the traditional auditory model exploits critical bands to perform the subband processing of signals with fixed time resolution and the temporal masking is often omitted.

Current research suggests that the wavelet approach to the modeling of the BM is preferable [BEN93], [KAR07], [SRI98], [YAO01], [YAO02]. In this chapter, we propose a new multiresolution human auditory system, which is obtained by modeling the auditory phenomena of absolute hearing threshold, perceptual wavelet scale, simultaneous masking, and temporal masking. The proposed auditory system adopts the base structure of the traditional auditory model but replace the time-invariant bandpass filter with wavelet packet transform (WPT). In order to mimic the time-frequency analysis of critical bands, the WP tree is appropriately derived using a criterion based on the perceptual frequency scale. The bandwidths of the proposed wavelet filter bank can be designed to have similar resolutions as the critical bands of human cochlea. The WPT coefficients of the wavelet filter bank are squared to obtain the energy components, which are decomposed into bands that closely match the critical band (CB) structure. The time-resolution of the critical band energy components are determined by the respective decomposition depth of the WP tree. It is capable of producing a good frequency resolution at low-frequency and good time resolution at high-frequency [MAL99], [MER99], [STR96]. These energy components are subsequently processed by spectral spreading to account for the simultaneous masking phenomenon due to the finite frequency resolution of the ear. To account for the temporal masking phenomenon, the model of temporal smearing is applied to individual spread critical band energy components. In a word, the proposed auditory system based on WPT processes the input acoustic signal by modeling the time-frequency representations of speech signals in the human auditory system to achieve effective noise suppression.

The remainder of this chapter is organized as follows. The basis of conventional psychoacoustic model of human auditory system is given in Section 3.2. The pertinent terms of hearing area, critical bands, absolute hearing threshold and masking properties are briefly explained. Section 3.3 presents the proposed perceptual wavelet filterbank design in critical band scale. Its performance is evaluated and compared with the traditional critical band. A unification of simultaneous and temporal masking in wavelet domain is elaborated in Section 3.4. Section 3.5 summarizes the chapter. A part of the work presented in Sections 3.2 and 3.3 has been presented at the *2005 and 2006 IEEE International Symposiums on Circuits and Systems* [SHA05a], [SHA06c]. A large portion of the work presented in this chapter has been published in the *IEEE Transactions on System, Man and Cybernetics - Part B* [SHA07a].

3.2 An Overview of Psychoacoustic Modeling

To engineer a viable psychoacoustic model, we need to understand how the human auditory system adapts to noise [GOL94], [LOI98], [ZHE99], [ZWI90]. Hearing and sound perception [ZWI90] are two phenomena that depict how the sound signal enters the listener's ear and be converted into a linguistic message. In physics terms, the pitch indicates the frequency of sound and the tone represents the quality or character of sound [ZWI90]. Our auditory system can sense and judge the pitch and tone, which are essential to the appreciation of music and contribute directly to our understanding of speech.

The essence of complex biological processing sequence in the human auditory system can be modeled and engineered by a number of digital signal processing stages [CHA95]. (1) A set of general bandpass filters are used to provide the filtering effect of basilar membrane in the cochlea. These bandpass filters can be realized by filterbank in the critical band scale. (2) An envelope detector is employed to estimate the intensity of the signal from each of the bandpass filters as detected by the hair cells. (3) The signal is logarithmically compressed with nonlinear intensity gain. Dynamic compression is used to model the transformation of mechanical vibrational energy into neural signals by the hair cells; (4) Loudness adaptation is provided by *high pass filters*. It is at this stage that the instantaneous intensity is derived from the intensity of the envelope of the acoustic pressure field. (5) The firing of the inner hair cells and the associated neural network is functionally mimicked by the *hyperbolic tangents*; (6) *Multiplicative intrinsic noise sources*, which are additive in the logarithmic domain, operate in conjunction with the exponentiators and detectors to efficiently synergize the function of loudness detection in the brain. This model exploits the property of critical bands, but it does not reveal the masking properties of human auditory system. To augment this psychoacoustic model, the human auditory response to sound is modeled by making use of the wavelet filterbank to approximate the critical bands and improvise the masking properties. The proposed model takes into account the effects of hearing area, critical band, hearing threshold and masking properties, which will be exposited in the following subsections.

3.2.1 Hearing Area

Fig. 3.1 illustrates the hearing area and the threshold in quiet. The hearing area is a plane in which audible sounds can be displayed. In its normal form, the hearing area is plotted with frequency on a logarithmic scale as the abscissa, and sound pressure level in dB and Watt per square meter (W/m^2) on a linear scale as the ordinate. Along the abscissa, our hearing organ produces sensations for pure tones within three decades in frequency range from 20 Hz to 20 kHz.

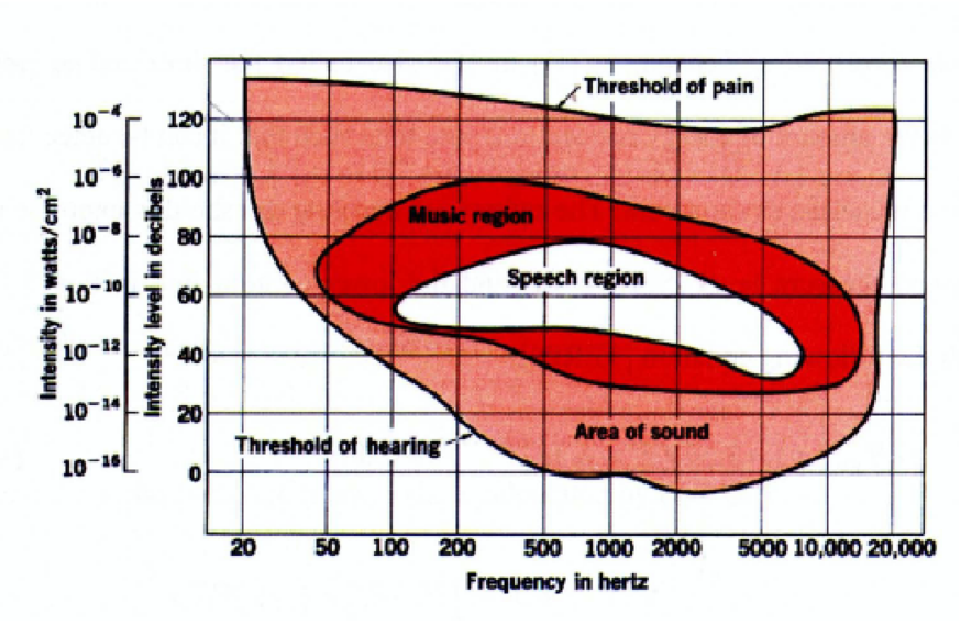


Figure 3.1 Hearing area: the area between the threshold in quiet and the threshold of pain. The areas encompassed by music and speech are also indicated. (Figure excerpted from [ZWI90])

The actual hearing area represents the range that is bounded within the thresholds in quiet (the limit towards low levels) and the threshold of pain (the limit towards high levels) [ZWI90]. These thresholds are demarcated in Fig. 3.1 by solid contour lines. If a speech is resolved into spectral components, the region it normally occupies is also illustrated in the hearing area. In Fig. 3.1, the range of speech starts near 100 Hz and ends near 7 kHz [ZWI90] as indicated by the white color region in Fig. 3.1. The components of music have a larger distribution in the hearing area as indicated by the red color region in Fig. 3.1. It starts at frequency as low as around 40 Hz, and extends to about 10 kHz [ZWI90]. The high level border of the pink color region marks is threshold of pain. What has not shown in this figure is the limit of damage risk, which is very important in our everyday life. This border is lower than the threshold of pain. It is reached at comparatively higher sound pressure levels (120 dB) at very low frequencies, but it decreases to nearly 90 dB in the range between 1 and 5 kHz [ZWI90].

The *absolute threshold of hearing* (ATH) in Hz, also called the threshold in quiet, characterizes the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment. The referenced absolute threshold is expressed in terms of sound pressure level (SPL) in dB and the lower contour line in Fig. 3.1 is plotted with the following equation [TER79]

$$ATH_{SPL}\left(f\right)=3.64f^{-0.8}-6.5e^{-0.6(f-3.3)^2}+0.001f^4 \tag{3.1}$$

3.2.2 Critical Band

The concept of critical bands was proposed by Fletcher [FLE40]. It models the masking of a narrow band (sinusoidal) signal by a wideband noise source. As mentioned in Chapter 2, the critical band analysis can be used to explain many characteristics of the auditory system due to the frequency transformation taken place along the BM in the inner ear [ZWI90]. A critical band can be deemed as a frequency selective 'channel' for psychoacoustic processing; only noise that falls within the critical bandwidth can contribute to the masking of a narrow band signal. The mammalian auditory system consists of a whole series of critical bands, each filter out a specific portion of the audio spectrum. The critical band has a perceptual as well as a physical correlation with the auditory system. From empirical study, there are 25 critical bands over the frequency of human hearing. The range is from 20 Hz to 20 kHz. Table 3.1 shows the lower (f_l) and upper (f_u) frequencies, center frequency (f_c) and bandwidth (Δf) in hertz (Hz) for each critical band. The critical band scale is approximately linearly related to the frequency scale at low frequencies but logarithmically related to the frequency scale at high frequencies. Thus, the referenced critical band rate, Z_B, in Bark is approximated by [ZWI90]:

$$Z_B = 13\tan^{-1}\left(7.6\times10^{-4} f\right) + 3.5\tan^{-1}\left(1.33\times10^{-4} f\right)^2 \tag{3.2}$$

where the frequency, f is measured in Hz.

The critical bandwidth (CBW) in Hz is calculated by [ZWI90]:

$$CBW\left(f\right) = 25 + 75\left(1 + 1.4\times10^{-6} f^2\right)^{0.69} \tag{3.3}$$

The bandwidth remains constant at approximately 100 Hz for center frequency below 500 Hz. For critical bands with center frequency above 500 Hz, the bandwidths increase with the center frequencies.

Table 3.1 Critical-band rate ZB (data excerpted from [ZWI90])

Z_B	$[f_l\ f_u]$	f_c	Δf_g
BarkScale	***[lower upper] Hz***	***Center Hz***	***Bandwidth Hz***
1	[0 100]	50	100
2	[100 200]	150	100
3	[200 300]	250	100
4	[300 400]	350	100
5	[400 510]	450	110
6	[510 630]	570	120
7	[630 770]	700	140
8	[770 920]	840	150
9	[920 1080]	1000	160
10	[1080 1270]	1170	190
11	[1270 1480]	1370	210
12	[1480 1720]	1600	240
13	[1720 2000]	1850	280
14	[2000 2320]	2150	320
15	[2320 2700]	2500	380
16	[2700 3150]	2900	450
17	[3150 3700]	3400	550
18	[3700 4400]	4000	700
19	[4400 5300]	4800	900
20	[5300 6400]	5800	1100
21	[6400 7700]	7000	1300
22	[7700 9500]	8500	1800
23	[9500 12000]	10500	2500
24	[12000 15500]	13500	3500
25	[15500 20000]	19500	4500

3.2.3 Masking

As a consequence of the highly nonlinear cochlear processes of BM vibration and neural firings, the perception of one sound is obscured by the presence of another in time and in frequency [ZWI90]. This phenomenon is called masking, which can be replicated using the notion of critical band. The masking property is exploited through adaptive thresholding to mask signals at nearby frequencies, making them inaudible to the listener. For a masked signal to be heard, its power will need to be increased to a level greater than a *masking threshold* determined by the sound pressure level, the frequency of the masker tone and its strength, and the *tonalities* of the masker and the maskee. There are two dominant types of masking. The *simultaneous masking* [ZWI90] is the masking between two concurrent sounds and it is best modeled in the frequency domain. The *temporal masking* [ZWI90] is the masking of other sounds immediately preceding or following a sudden stimulus sound, and it is usually modeled in the time domain. Two main time domain masking phenomena have been observed in human audition: *pre-masking*, which is also called backward masking, and *post-masking*, which is also called forward masking. Post-masking has a more important effect than pre-masking since it has a longer duration. Pre-masking appears approximately 20 ms before the masker, whereas post-masking lasts for about 100 to 200 ms [ZWI90]. These two kinds of masking are widely accepted as separate mechanisms for the purpose of modeling. However, the behaviour of temporal masking depends on the frequency location of the masker with respect to the maskee.

3.3 Perceptual Wavelet Filterbank Design in Critical Band Scale

Experiments have shown that the critical band behaviour of the BM can be approximated by a multi-independent channel articulation band model [GOL94]. Such filter bank modeling is useful for mimicking the sensorineural responsiveness to signal-to-noise ratio and analyzing the cochlear selectivity to the perceived quality of the acoustic environment.

Since wavelet analysis can be considered as constant-Q or octave analysis [MER99] similar to the critical band structure of the human auditory system, we propose a new implementation of the human auditory model by replacing the time-invariant bandpass filters with WPT. The time-frequency analysis is done using WPT. The energy of the input signal is decomposed into bands that closely match the critical band structure. Unlike the best WP basis method, which is obtained by minimizing the classical entropy-based criterion, the WP tree is appropriately derived using a criterion based on the perceptual frequency scale for the input acoustic signal. The quadrature mirror filters (QMF) in the WP tree are designed to obtain the WPT coefficients, which are grouped according to the critical band structure. To propose a WP-based auditory model in the context of critical band decomposition and auditory masking, the idea of *multiresolution analysis* and the efficient realization of the discrete wavelet transform based on multirate filter banks will be addressed. Before presenting the proposed wavelet packet decomposition, a brief description of the discrete wavelet transform (DWT) is given.

Wavelets are families of basis functions. A DWT is a set of linearly independent functions that can be used to produce all admissible functions $x[n]$ in $l^2(\mathbb{Z})$. That is, $x[n]$ is a finite energy function [MAL99] and $\sum_{n=-\infty}^{+\infty}\left|x[n]\right|^2 < +\infty$. A signal, $x[n]$ can be mathematically synthesized from a combination of DWT basis functions as follows [MAL99]:

$$x[n]=\sum_{k=0}^{K}c_0[k]\phi_{0,k}[n]+\sum_{k=0}^{K}\sum_{j=0}^{J-1}d_j[k]\psi_{j,k}[n] \tag{3.4}$$

where the index *j* represents the scale of decomposition and *k* the coefficient number. *J* is the maximum number of decomposition levels and *K* the total number of coefficients. $\phi[n]$ and $\psi[n]$ are the scaling and wavelet functions, respectively. $c_j[k]$ is a scaling coefficient and in DWT, $j = 0$. $d_j[k]$ is a wavelet coefficient and $j = 0, 1, \ldots, J-1$. The DWT can be implemented by the pyramid algorithm [MER99]. It should be noted that $\phi_{j,k}[n]$ and $\psi_{j,k}[n]$ are special functions of the wavelet basis, which means that all functions of $\phi_{j,k}[n]$ and $\psi_{j,k}[n]$ are constructed from the single basis function. Typically, $\phi_{j,k}[n]$ and $\psi_{j,k}[n]$ are obtained by compressing the mother wavelet *j* times and shifting it *k* times.

$$\phi_{j,k}[n] = 2^{-\frac{j}{2}}\phi\left[2^{-j}n-k\right] \tag{3.5}$$

$$\psi_{j,k}[n] = 2^{-\frac{j}{2}}\psi\left[2^{-j}n-k\right] \tag{3.6}$$

The functions $\phi_{j,k}[n]$ and $\psi_{j,k}[n]$ form an orthonormal basis. The dilation equations for the scaling and the wavelet functions are defined as:

$$\phi[n] = \sqrt{2}\sum_{k=0}^{K} h[k]\phi[2n-k] \tag{3.7}$$

$$\psi[n] = \sqrt{2}\sum_{k=0}^{K} g[k]\psi[2n-k] \tag{3.8}$$

where $h[k]$ is the transfer function of a lowpass filter which leads to the scaling function, and $g[k]$ is the transfer function of a highpass filter which leads to the wavelet function. The lowpass and highpass filters form a quadrature mirror filter (QMF) pair. The wavelet packet transform (WPT) is an extension of DWT, where both the scaling and wavelet coefficients are decomposed. The WPT can be implemented using an extension of the pyramid algorithm where both the scaling and wavelet coefficients are decomposed in a tree-structured, QMF filterbank. The synthesis of signal $x[n]$ using WPT can therefore be mathematically expressed as follows [MAL99].

$$x[n] = \sum_{k=0}^{K}\sum_{j=0}^{J-1} c_j[k]\phi_{j,k}[n] + \sum_{k=0}^{K}\sum_{j=0}^{J-1} d_j[k]\psi_{j,k}[n] \qquad (3.9)$$

The simplified analysis equation for WPT is defined as:

$$\left\{w_{j,k}(x)\right\} = WPT\left(x[n]\right) \qquad (3.10)$$

where $w_{j,k}(x)$ is a wavelet transform coefficient, (j,k) in the subscript of w corresponds to its scale and translation indices.

In the proposed wavelet packet decomposition, two-channel wavelet filter banks are used to split both the lowpass and highpass bands as opposed to the decomposition of only the low frequency bands in the usual wavelet decomposition [STR96]. A cascaded 2-channel wavelet analysis and synthesis filter bank structure is depicted in Fig. 3.2.

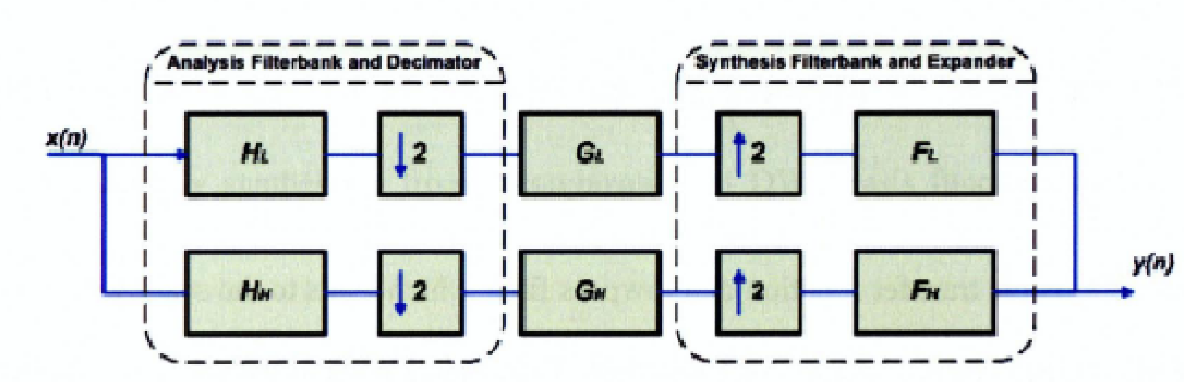

Analysis Filterbank and Decimator
Synthesis Filterbank and Expander
x(n)
H_L
↓2
G_L
↑2
F_L
H_H
↓2
G_H
↑2
F_H
y(n)

Figure 3.2 Two-channel filterbank: the input is separated into frequency bands for analysis, and reassembled to synthesize the output

The output y(n) corresponds to an input x(n) through the two-channel filterbank, in z-transform, is given by:

$$\begin{aligned} Y(z) &= \frac{1}{2}X(z)\left[F_L(z)G_L(z^2)H_L(z)+F_H(z)G_H(z^2)H_H(z)\right] \\ &+\frac{1}{2}X(-z)\left[F_L(z)G_L(z^2)H_L(-z)+F_H(z)G_H(z^2)H_H(-z)\right] \end{aligned} \tag{3.11}$$

where $G_L(z)$ and $G_H(z)$ are the transfer functions of the subband processing filters. $H_L(z)$ and $H_H(z)$ are the lowpass and highpass transfer functions, respectively before the decimation-by-two operation in each stage of the analysis filter bank. $F_L(z)$ and $F_H(z)$ are the lowpass and highpass transfer functions, respectively after the upsampling-by-two operation in each stage of the synthesis filter bank.

(3.11) can be reformulated as follows:

$$\begin{aligned} Y(z) &= X(z)T_0(z)+X(-z)T_1(z) \\ T_0(z) &= \frac{1}{2}\left[F_L(z)G_L(z^2)H_L(z)+F_H(z)G_H(z^2)H_H(z)\right] \\ T_1(z) &= \frac{1}{2}\left[F_L(z)G_L(z^2)H_L(-z)+F_H(z)G_H(z^2)H_H(-z)\right] \end{aligned} \tag{3.12}$$

where $T_0(z)$ is the distortion transfer function, and $T_1(z)$ is the aliasing transfer function. The criterion for the perfect reconstruction (PR) of conventional QMF filterbank is given as follows [STR96]:

$$\begin{aligned} &F_L(z)H_L(z)+F_H(z)H_H(z)=Cz^{-l} && \text{No distortion} \\ &F_L(z)H_L(-z)+F_H(z)H_H(-z)=0 && \text{Alias cancellation} \end{aligned} \tag{3.13}$$

where l is an integer and C is a constant. (3.13a) implies that perfect reconstruction can be accomplished with a l-step delay in the z-domain. (3.13b) indicates the condition for aliasing-free reconstruction. However, for arbitrary transfer functions, $G_H(z)$ and $G_L(z)$, the aliasing-free equations need to be modified to

$$F_L(z)H_L(-z)=0 \quad and \quad F_H(z)H_H(-z)=0 \tag{3.14}$$

It should be noted that the criteria imposed by (3.13a) and (3.14) are so strict that they are very difficult to satisfy by standard filter design methodologies. Many methods have been proposed to ease this problem [TAN97], [VAI87], [VAI93].

By omitting the effect of transfer functions $G_H(z)$ and $G_L(z)$, the analysis and synthesis filter banks are derived from (3.13b) [MAO00] as follows:

$$\begin{aligned} h_H[n]&=(-1)^n f_L[n] \leftrightarrow H_H(z)=F_L(-z) \\ f_H[n]&=-(-1)^n h_L[n] \leftrightarrow F_H(z)=-H_L(-z) \end{aligned} \tag{3.15}$$

The above relationship between the lowpass and highpass filters reduces the number of filters to be implemented for each stage of the two-channel filter bank by half. Once the lowpass (LP) filters, $H_L(z)$ and $F_L(z)$ are designed, the highpass (HP) filters, $H_H(z)$ and $F_H(z)$ can be directly derived from (3.15). Since all the filters are linear phase finite impulse response filters, they can be realized with low complexity symmetrical filter structure after the quantization of real-valued coefficients.

From this two-channel filter bank, multirate digital filters consisting of an analysis section and a synthesis section [STR96] can be built. The analysis filter bank is a bank of M digital filters with a common input. The transfer functions of the analysis filters are denoted by $H_1(z)$, $H_2(z)$, …, $H_M(z)$. The input signal, $x[n]$ is thereby partitioned into a new set of signals denoted by $\{x_k[n]\}_{k=1}^{M}$, which are called the subband signals. The subband signals are down sampled by a bank of decimators by virtue of their narrower bandwidth than the full-band signal, $x[n]$. The analysis section of the multirate digital filter processes each decimated signal in such a way that the special properties (energy levels or perceptual significance) of the k-th decimated subband signal are extracted. The resulting signals are then applied to the synthesis section of the multirate digital filter for further processing. The synthesis section consists of two functional blocks of its own. The bank of expanders is used to upsample the respective inputs. Each expander performs an interpolation function but a filter is needed to convert the zero-valued samples of the expander into interpolated samples to complete the interpolation. The synthesis filter bank consists of a parallel connection of a set of M digital filters with a common output. The transfer functions of the synthesis filters are denoted by $F_1(z)$, $F_2(z)$, …, $F_M(z)$, and the resulting output of the synthesis section is denoted by $y(n)$. The output signal, $y[n]$ differs from the input signal, $x[n]$ due to: (1) the external processing performed on the decimated signals in the analysis section and (2) the aliasing errors. In the context of our present discussion, aliasing refers to the phenomenon of a high-frequency component taking on the identity of a lower frequency component in the spectrum of its decimated version. This phenomenon, arising because of the non-ideal nature of the analysis filters, also includes the aliasing of a low-frequency band into a high-frequency band. As the decimation creates multiple copies of lower frequency components, it is possible for a low-frequency-band signal to show up elsewhere.

Let $T(z)$ denote the overall transfer function of the multirate digital filter. Suppose that, in the absence of any external processing performed on the output signals of the analysis section, the transfer functions $H_1(z)$, $H_2(z)$, …, $H_M(z)$ of the analysis filters and the corresponding transfer functions $F_1(z)$, $F_2(z)$, …, $F_M(z)$ of the synthesis filters are chosen in such a way that $T(z)$ is forced to be a pure delay, that is [STR96],

$$T(z) = cz^{-\Delta} \tag{3.16}$$

where c is a scaling factor and Δ is the processing delay introduced by the cascaded analysis and synthesis filters. When this condition is satisfied, the alias-free multirate digital filter is said to have the *perfect reconstruction property* [STR96].

Table 3.2 Critical-band rate Z_B for Different Sampling Frequencies

Sampling Rate f_s	No. of decomposition depth for WPT J	Critical band rate Z_B
16 kHz	6	21
32 kHz	7	24
44.1 kHz	8	25

One of the most instrumental signal processing tasks in the perceptual auditory modeling [KAR07] is the realization of critical bands. For some of the widely used sampling frequencies, i.e., 16kHz, 32kHz, and 44.1 kHz, the optimal numbers of decomposition depth of WPT, obtained based on the perceptual criterion, are shown in Table 3.2.

In our proposed system, we deduced a perceptual wavelet packet transform to decompose the speech signal from 20 Hz to 8 kHz into 21 frequency subbands that approximate the critical bands. The decomposition is implemented with an efficient 6 level wavelet packet tree structure. The output of this stage is a set of wavelet coefficients. This binary tree decomposition is attractive due to the following reasons. First, the smoothness property of wavelet is determined by the number of vanishing moments. If a wavelet with a large number of vanishing moments is used, the stringent bandwidth and stopband attenuation of each subband, as specified by the human auditory model, can be more closely approximated by the wavelet decomposition. Secondly, the psychoacoustic study of human ears in the previous section and in Chapter 2 suggests that a frequency to bark transformation needs to be performed to accurately model the frequency dependent sensitivity of human ears. Such a transformation is accomplished in audio processing systems by dividing the frequency range into critical bands.

Table 3.3 Critical-band rate Z_B, and Perceptual Wavelet Filter-Banks, Z_W (the critical-band data are excerpted from [ZWI90])

Z_B	$[f_l\ f_u]$	f_c	Δf_g	Z_W	$[wf_l\ wf_u]$	wf_c	Δwf_g
Bark Scale	***[lower upper] Hz***	***Center Hz***	***Bandwidth Hz***	***Wavelet Scale***	***[lower upper] Hz***	***Center Hz***	***Bandwidth Hz***
1	[0 100]	50	100	**1**	[0 125]	62.5	125
2	[100 200]	150	100	**2**	[125 250]	187.5	125
3	[200 300]	250	100	**3**	[250 375]	312.5	125
4	[300 400]	350	100	**4**	[375 500]	437.5	125
5	[400 510]	450	110	**5**	[500 625]	562.5	125
6	[510 630]	570	120	**6**	[625 750]	687.5	125
7	[630 770]	700	140	7	[750 875]	812.5	125
8	[770 920]	840	150	**8**	[875 1000]	937.5	125
9	[920 1080]	1000	160	**9**	[1000 1250]	1125	250
10	[1080 1270]	1170	190	**10**	[1250 1500]	1375	250
11	[1270 1480]	1370	210	**11**	[1500 1750]	1625	250
12	[1480 1720]	1600	240	**12**	[1750 2000]	1875	250
13	[1720 2000]	1850	280	**13**	[2000 2250]	2125	250
14	[2000 2320]	2150	320	**14**	[2250 2500]	2375	250
15	[2320 2700]	2500	380	**15**	[2500 3000]	2750	500
16	[2700 3150]	2900	450	**16**	[3000 3500]	3250	500
17	[3150 3700]	3400	550	**17**	[3500 4000]	3750	500
18	[3700 4400]	4000	700	**18**	[4000 5000]	4500	1000
19	[4400 5300]	4800	900	**19**	[5000 6000]	5500	1000
20	[5300 6400]	5800	1100	**20**	[6000 7000]	6500	1000
21	[6400 7700]	7000	1300	**21**	[7000 8000]	7500	1000

Table 3.3 shows the 21 critical bands in the hearing range of 0–8 kHz. Using a six-stage discrete wavelet packet transform (WPT) with a frame length, F_{jmax} of 128 for $j_{max} = 6$, a frequency resolution of 125 Hz can be achieved. The choice of the wavelet basis of the transform influences the separation of the subband signals and determines the maximal frame length. Owing to their regularity, the filters proposed by Daubechies are

the ones that best preserve frequency selectivity as the number of stages j of the WPT increases [STR96]. The temporal resolution of the human ear requires that the analysis windows be limited to 5–10 ms towards the higher frequencies, but they can spread up to 100 ms at lower frequencies. These constraints have led us to the use of the Daubechies basis with length, L = 20. The frame length at stage j is given by $F_j = 2^j$. Once the DWPT is chosen, its time–frequency resolution remains fixed, and each transform coefficient can be expressed as W_{jk}, where k is the coefficient number, j is the transform stage from which W_{jk} is chosen and l is the number of "temporal" coefficients in the critical band. Table 3.4 shows the mapping of the perceptual wavelet packet transform coefficients in each stage.

Table 3.4 Perceptual Wavelet Filter-banks Coefficients

Subband Z_W	l	***Coefficients*** k_a -k_b	***Transform stage j***
1	2	0-1	6
2	2	2-3	6
3	2	4-5	6
4	2	6-7	6
5	2	8-9	6
6	2	10-11	6
7	2	12-13	6
8	2	14-15	6
9	4	16-19	5
10	4	20-23	5
11	4	24-27	5
12	4	28-31	5
13	4	32-35	5
14	4	36-39	5
15	8	40-47	4
16	8	48-55	4
17	8	56-63	4
18	16	64-79	3
19	16	80-95	3
20	16	96-111	3
21	16	112-127	3

*k_a and k_b are the coefficient indices of the first and last transform coefficients within a given critical band.

Fig. 3.3 shows the analysis filter bank and the following equations describe its underlying perceptual wavelet packet decomposition.

$$H_1(z)=H_L(z)H_L(z^2)H_L(z^4)H_L(z^8)H_L(z^{16})H_L(z^{32}) \tag{3.17}$$

$$H_2(z)=H_L(z)H_L(z^2)H_L(z^4)H_L(z^8)H_L(z^{16})H_H(z^{32}) \tag{3.18}$$

$$H_3(z)=H_L(z)H_L(z^2)H_L(z^4)H_L(z^8)H_H(z^{16})H_L(z^{32}) \tag{3.19}$$

$$H_4(z)=H_L(z)H_L(z^2)H_L(z^4)H_L(z^8)H_H(z^{16})H_H(z^{32}) \tag{3.20}$$

$$H_5(z)=H_L(z)H_L(z^2)H_L(z^4)H_H(z^8)H_L(z^{16})H_L(z^{32}) \tag{3.21}$$

$$H_6(z)=H_L(z)H_L(z^2)H_L(z^4)H_H(z^8)H_L(z^{16})H_H(z^{32}) \tag{3.22}$$

$$H_7(z)=H_L(z)H_L(z^2)H_L(z^4)H_H(z^8)H_H(z^{16})H_L(z^{32}) \tag{3.23}$$

$$H_8(z)=H_L(z)H_L(z^2)H_L(z^4)H_H(z^8)H_H(z^{16})H_H(z^{32}) \tag{3.24}$$

$$H_9(z)=H_L(z)H_L(z^2)H_H(z^4)H_L(z^8)H_L(z^{16}) \tag{3.25}$$

$$H_{10}(z)=H_L(z)H_L(z^2)H_H(z^4)H_L(z^8)H_H(z^{16}) \tag{3.26}$$

$$H_{11}(z)=H_L(z)H_L(z^2)H_H(z^4)H_H(z^8)H_L(z^{16}) \tag{3.27}$$

$$H_{12}(z)=H_L(z)H_L(z^2)H_H(z^4)H_H(z^8)H_H(z^{16}) \tag{3.28}$$

$$H_{13}(z)=H_L(z)H_H(z^2)H_L(z^4)H_L(z^8)H_L(z^{16}) \tag{3.29}$$

$$H_{14}(z)=H_L(z)H_H(z^2)H_L(z^4)H_L(z^8)H_H(z^{16}) \tag{3.30}$$

$$H_{15}(z)=H_L(z)H_H(z^2)H_L(z^4)H_H(z^8) \tag{3.31}$$

$$H_{16}(z)=H_L(z)H_H(z^2)H_H(z^4)H_L(z^8) \tag{3.32}$$

$$H_{17}(z)=H_L(z)H_H(z^2)H_H(z^4)H_H(z^8) \tag{3.33}$$

$$H_{18}(z) = H_H(z)H_L(z^2)H_L(z^4) \tag{3.34}$$

$$H_{19}(z) = H_H(z)H_L(z^2)H_H(z^4) \tag{3.35}$$

$$H_{20}(z) = H_H(z)H_H(z^2)H_L(z^4) \tag{3.36}$$

$$H_{21}(z) = H_H(z)H_H(z^2)H_H(z^4) \tag{3.37}$$

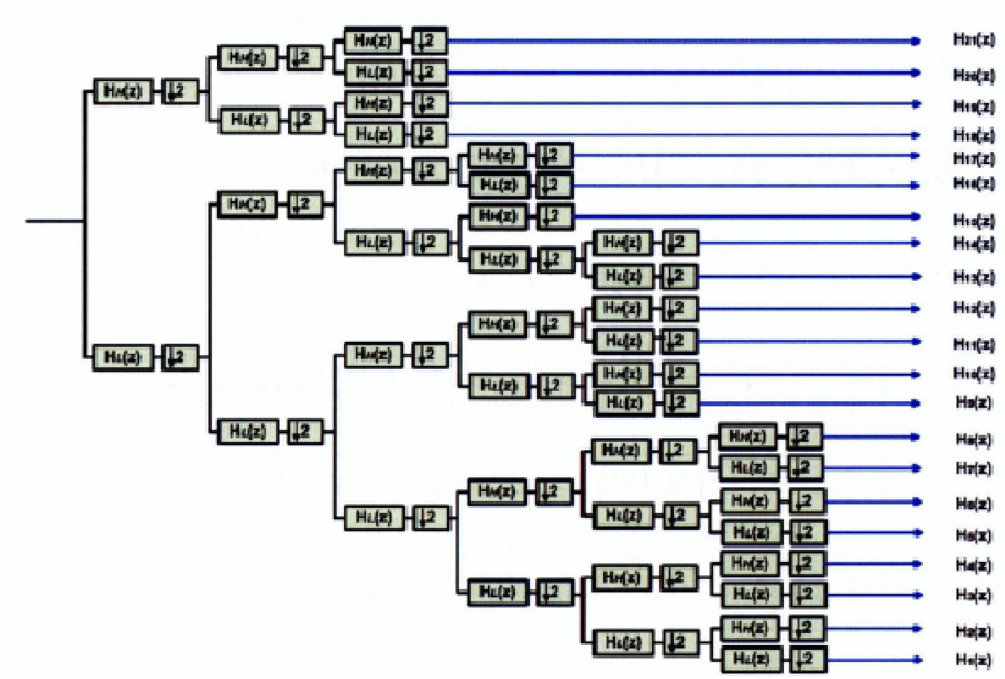


Figure 3.3 Proposed Perceptual Wavelet Packet decomposition tree

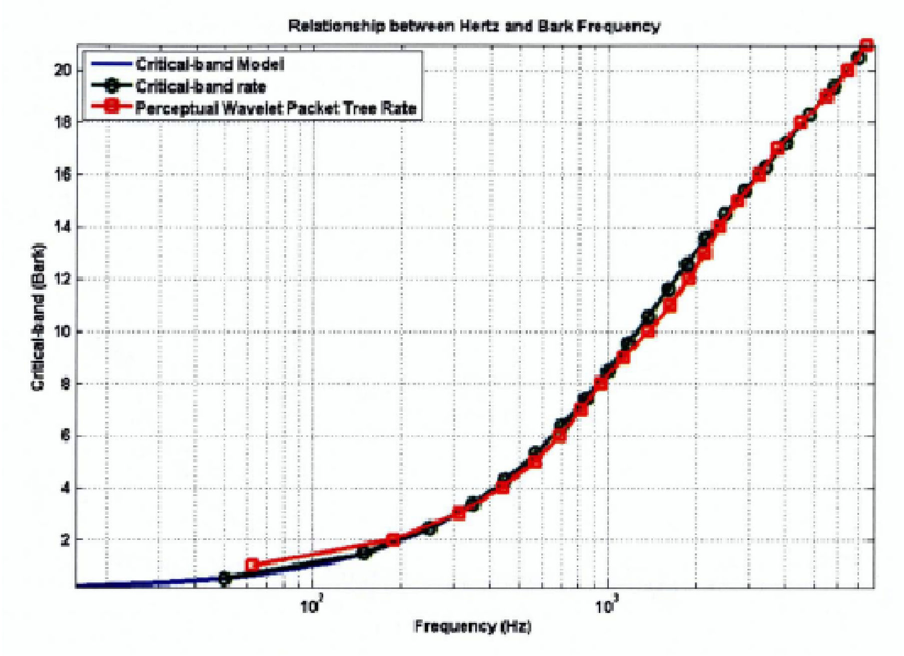


Figure 3.4 Center frequencies of critical bands and perceptual wavelet packet decomposition tree

Fig. 3.4 depicts the difference in the critical band rate between the critical bands and those of the proposed perceptual wavelet packet tree. This figure was obtained from the simulation results of Matlab.

The simulation results of the bandwidths of the critical bands and the perceptual wavelet packet tree are plotted in Fig. 3.5.

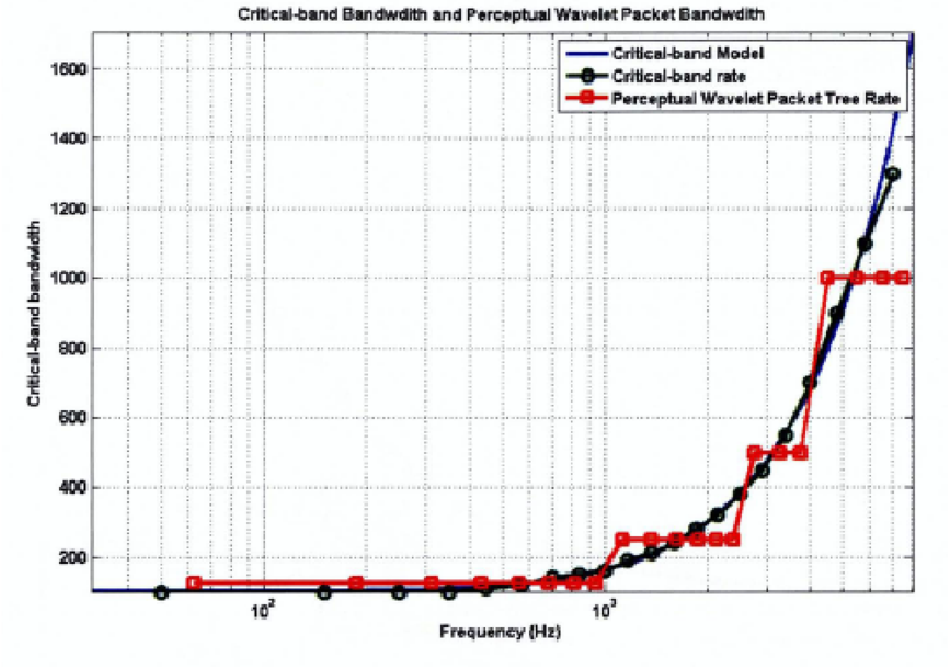


Figure 3.5 Bandwidths of critical bands and perceptual wavelet packet decomposition tree

Fig. 3.6 compares the *absolute threshold of hearing* (ATH) in Hz, critical band scale and perceptual wavelet packet scale. The natural scale of inner ear can be viewed as the critical band scale. The frequency range from 20 to 8000 Hz covered by the human auditory system is approximately 21 Barks. The figure was obtained by Matlab simulation.

From the results tabulated in Table 3.3 and plotted in Figs. 3.4 to 3.6, it is evident that the proposed perceptual wavelet packet tree can closely mimics the experiential critical bands. The parameters of the discrete wavelet packet transform filter used to derive the plots of Figs. 3.4 to 3.6 are determined based on the auditory masking properties and they will be discussed in details in the next section.

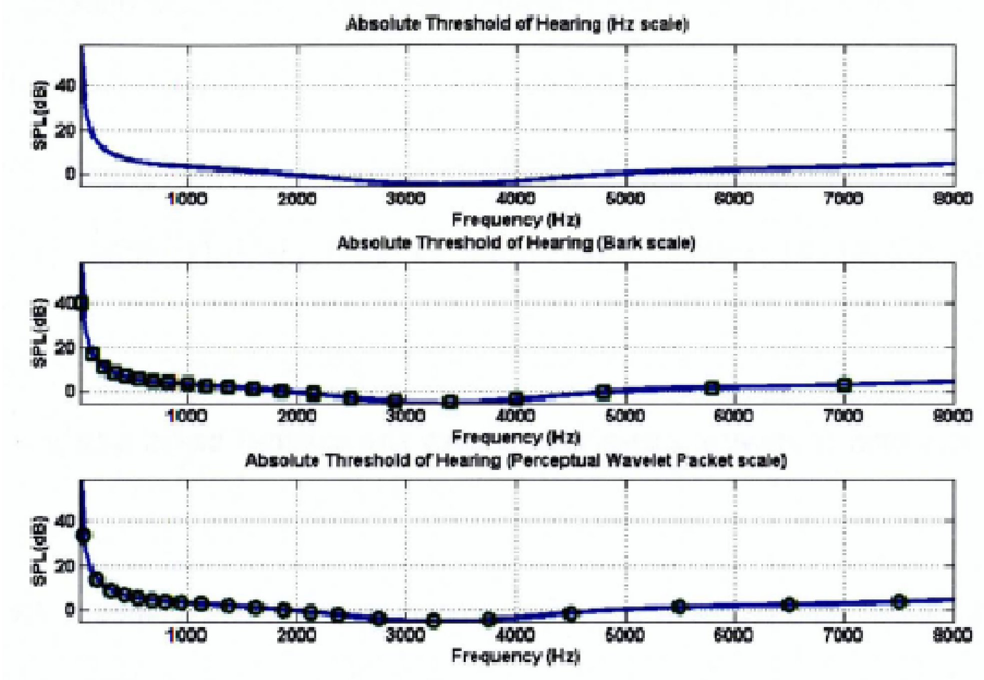


Figure 3.6 Absolute Threshold of Hearing in frequency, Bark, and Perceptual Wavelet Packet Tree Scales

3.4 Unification of Simultaneous and Temporal Maskings in Wavelet Packet Domain

After studying the limitations of traditional auditory model using Fourier transform, we establish a unification of the simultaneous and temporal maskings by means of perceptual wavelet packet transform in this section.

In subtractive-type of speech enhancement, the maskee is the residual noise. To make it inaudible, it should lie below the masking threshold or threshold of just noticeable distortion (JND) [JOH88], [ZWI90]. Clean speech is ideal for the estimation of masking thresholds. When clean speech is not accessible, the enhanced speech processed by the spectral subtraction gives reasonably good estimate of the clean signal. Since masking is manifested by a shift of auditory threshold in signal detectability, and loudness is an important attribute of auditory perception, the noise masking threshold can be determined by the characteristics of the masker and the maskee, the sound pressure level and the frequency of the masker. The noise masking threshold for a residual noise maskee in each critical band can be calculated as follows:

3.4.1 Time and frequency analysis along the critical band scale

Our goals are to approximate the critical band analysis and estimate the auditory masking threshold using the perceptual wavelet packet transform. The resulting 21-band wavelet packet approximation of the critical band rate and the bandwidth approximation have been shown in Figs. 3.4 and 3.5, respectively. The input signal is decomposed into critical bands scale for time-frequency analysis. The outputs of the perceptual wavelet filterbank are squared to obtain the Bark spectrum energy and temporal energy, respectively. The Bark energy spectrum, $E_{Bark}(Z_W)$ with $1 \le Z_W \le 21$ is computed as:

$$E_{Bark}\left(Z_W\right) = \sum_{k=k_a}^{k_b} W_{j,k}^2\left(Z_W\right) \tag{3.38}$$

where k_a and k_b are the coefficient indices of the first and last transform coefficients referring to Table 3.4. The temporal energy $E_{temp}[n]$ with $n = 0, 1, \ldots, l(Z_W) - 1$ in each critical band Z_W can be derived as:

$$E_{temp}\left[n\right] = W_{j,(k_a+n)}^2\left(Z_W\right) \tag{3.39}$$

3.4.2 Convolution with spreading function

In this step, the Bark spectrum energy and temporal energy are convoluted with the spreading function. The spreading of Bark energy $E'_{Bark}\left(Z_W\right)$ is calculated by the convolution of $E_{Bark}\left(Z_W\right)$ with the linear version of *SF*(Z_W) [ZWI90].

$$E'_{Bark}\left(Z_W\right)=\frac{1}{l\left(Z_W\right)}\sum_{m=1}^{21}E_{Bark}\left(m\right)SF\left(Z_W-m\right) \tag{3.40}$$

where the *spreading function* *SF*(Z_W) in dB models the interaction of the masking signals between different critical bands. It can be deduced as follows [SCH79]:

$$SF\left(Z_W\right)=a+\frac{v+u}{2}\left(Z_W-Z_W\left(k\right)+c\right)-\frac{v-u}{2}\sqrt{d+\left(Z_W-Z_W\left(k\right)+c\right)^2} \tag{3.41}$$

where u and v are, respectively, the upper and lower slopes in dB per Bark; d is the peak flatness; a and c are the compensation factors needed to satisfy $SF(0) = 1$. $k \in [1, 21]$ is an integer critical band identifier. The critical band rate distance, $Z_W - Z_W(k)$, varies from –21 to 21 Barks. The parameters of the spreading function, *SF*(Z_W), which are empirically determined by the listening test, were set to v = 30 dB/Bark, u = –25 dB/Bark, d = 0.3, c = 0.05 and a = 27.4.

The temporal masking can be modeled in terms of temporal energy spreading across time. The temporal spreading energy $E'_{temp}\left[n\right]$ with n = 0, 1, .., $l(Z_W) - 1$ can be obtained by the convolution of the linear temporal spreading function *SF(n)* with the temporal energy, whose calculation is similar to (3.40).

$$E'_{temp}\left[n\right]=\sum_{m=0}^{l(Z_W)-1}E_{temp}\left[m\right]SF\left(n-\frac{N}{l\left(Z_W\right)}m\right) \tag{3.42}$$

where *SF*(*n*) is a parameterized spreading function similar to (3.41). It can be deduced as follows:

$$SF\left(n\right)=a+\frac{v+u}{2}\left(n-n\left(k\right)+c\right)-\frac{v-u}{2}\sqrt{d+\left(n-n\left(k\right)+c\right)^2} \tag{3.43}$$

According to the minimum frame duration F_{min}, the parameters of the spreading function, *SF*(*n*) in dB employed were set to v = 20 dB/F_{min} (dB/ms), u = −10 dB/F_{min} (dB/ms), d = 0.3, c = 0.2, a = 7.9 after several listening tests.

3.4.3 Calculation of relative threshold offset

A relative threshold offset $O(Z_w)$ is subtracted from each critical band. It is required to distinguish between tone-like and noise-like components in speech to calculate the relative threshold offset. There are two types of noise masking threshold: the threshold of tone masking noise T_{tmn}, where $T_{tmn}\left(Z_W\right)=E'_{Bark}-14.5-Z_W$ dB, and the threshold of noise masking tone T_{nmt}, where $T_{tmn}\left(Z_W\right)=E'_{Bark}-5.5$ dB. To determine whether the signal is tonelike or noiselike, the spectral flatness measure (SFM) is used. From this value, a tonality coefficient τ [JOH88] is generated:

$$\tau=\min\left(\frac{SFM_{dB}}{SFM_{dB\min}},1\right) \tag{3.44}$$

where τ is a ratio of SFM_{dB} to SFM_{dBmin}. SFM_{dB} is computed at the last stage of the decomposition, SFM_{dBmin} is an entirely tonelike reference over a frame. Using τ, the relative threshold offset $O(Z_w)$ in dB [JOH88] for each critical band is calculated as:

$$O(Z_W) = \tau \cdot (14.5 + Z_W) + (1 - \tau) \cdot 5.5 \tag{3.45}$$

3.4.4 Renormalization and comparison with absolute threshold of hearing

To unify the temporal and simultaneous maskings, a temporal masking factor $\rho_{j,k}(Z_W)$ is introduced as follows:

$$\rho_{jk}(Z_W) = E'_{temp}\left[k - k_a\right] / W^2_{j,k}(Z_W) \tag{3.46}$$

where $\rho_{j,k}(Z_W) \geq 1$ indicates the extent of temporal masking. If $\rho_{j,k}(Z_W) = 1$, no temporal masking is induced. Otherwise, if $\rho_{j,k}(Z_W) > 1$, the temporal masking has occurred. By combining the simultaneous masking threshold, $E'_{Bark}(Z_W)$ and the temporal masking factor, $\rho_{j,k}(Z_W)$, the *time-frequency masking threshold* in the critical band Z_W can be determined as follows:

$$M_{j,k}(Z_W) = 10^{\log\left[E'_{Bark} \cdot \rho_{jk}(Z_W)\right] - \left[O(Z_W)/10\right]} \tag{3.47}$$

Finally, $M_{j,k}(Z_W)$ is compared to the ATH in each critical band, and the maximum of each value is retained as $T_{j,k}(Z_W)$.

$$T_{j,k}(Z_W) = \max\left(M_{j,k}(Z_W), ATH(Z_W)\right) \tag{3.48}$$

3.5 Chapter Summary

A good auditory model should incorporate the physiological components of the human auditory system. The auditory system can be considered as a bank of bandpass filters. Auditory perception is modeled by means of the critical band analysis of frequency transformation occurring along the basilar membrane in the inner ear [ZWI90]. The phenomenon of masking can also be modeled using the critical bands. The masking threshold is determined to quantitatively measure the effect of masking.

This chapter proposes a psychoacoustic model based on a perceptual wavelet filterbank. By incorporating knowledge of human auditory system, this fixed filter bank approximation is simpler to realize than the best basis subband decomposition. The WPT coefficients at the terminal nodes of the WP tree are squared to obtain the energy components at the critical band channels. The time-resolution of the critical band is determined by the respective decomposition depth of the proposed perceptual wavelet filterbank. This improvement advantageously exploits the wavelet multirate signal representation to preserve the critical transient information, and avoids the speech transients from being averaged out over the frame duration in STFT-based models. The models of the different auditory phenomena are reformulated for the multiresolution framework. These include absolute hearing threshold, WP tree based on critical band structure, the model of spectral spreading, and the model of temporal spreading. Furthermore, simultaneous masking and temporal masking of the human auditory system are modeled by the perceptual wavelet packet transform through the frequency and temporal localization of speech components. The wavelet coefficients are used to calculate the Bark spreading energy and temporal spreading energy, from which a time-frequency masking threshold is deduced to adaptively adjust the subtraction parameters of our generalized perceptual time-frequency subtraction (GPTFS) technique and perceptual weighting filter used in the proposed wavelet Kalman filter for speech enhancement. In the next two chapters, we will further elaborate how these two new speech enhancement methods take advantage of the proposed psychoacoustic model.

CHAPTER 4.

SPEECH ENHANCEMENT BASED ON GENERALIZED PERCEPTUAL TIME-FREQUENCY SUBTRACTION

4.1 Introduction

In this Chapter, a subtractive-type algorithm, which estimates the short-time spectral magnitude of speech by subtracting a noise estimate from the noisy speech, will be investigated. The main attractions of spectral subtraction are its relative simplicity, in that it only requires an estimate of the noise power spectrum, and its high flexibility against subtraction parameters variation. However, the main problem in spectral subtraction is the presence of processing distortions caused by the random variations of noise. Musical residual noise is introduced in the subtractive type denoising methods, which degrades the perceptibility of the processed speech. To reduce the effect of residual noise, a number of methods were considered. These methods have been reviewed in Chapter 2, and their performances can be further improved in adverse environments, especially when the SNR is very low. The reason is that in very low SNR condition, it is difficult to suppress noise without degrading the intelligibility and without introducing residual noise and speech distortion. Therefore, the processed speech can be even less perceptible than the original noisy speech. Methods that adopt the masking property of human auditory system were found to be capable of reducing the effect of residual noise, but the drawback is the large computational effort associated with the subband decomposition and the additional fast Fourier transform (FFT) analyzer required for psychoacoustic modeling.

FFT based spectral subtraction methods generally do not make use of the temporal masking models. In order to simplify the mapping of audio signals into a scale that could well preserve the time-frequency related information, the wavelet packet transform has been incorporated recently in some proposed speech and audio codings. As discussed in Chapter 2, many of these methods have the high frequency speech signals grievously reduced by wavelet transform that causes the quality of their processed speech to be degraded.

This chapter introduces a versatile speech enhancement method by leveraging on the human auditory model presented in the previous chapter. The emphases are on the reduction of the effect of residual noise and speech distortion in the denoising process and the enhancement of the denoised speech in high frequency to improve its intelligibility. The proposed speech enhancement system consists of two main functions. One is a generalized perceptual time-frequency subtraction (GPTFS) method based on the masking properties of human auditory system. The GPTFS works in conjunction with the perceptual wavelet packet transform (PWPT) described in Chapter 3 to reduce the effect of noise contamination. The proposed approach is a subtractive-type enhancement process based on the auditory model. A new multiresolution auditory model is introduced. It uses perceptual wavelet packet transform for time-frequency decomposition of the input signal. The selection of the wavelet packet tree is based on a critical band structure of human auditory system. The models of different auditory phenomena, e.g. absolute hearing threshold, spectral spreading and temporal smearing, are reformulated for the multiresolution framework. Then, the unified simultaneous and temporal masking is derived based on multiresolution spectral spreading and temporal smearing. This single channel subtractive-type algorithm is characterized by a tradeoff between the amount of noise reduction, the speech distortion, and the level of musical residual noise. Unlike the classical algorithms, which are usually limited to the use of fixed optimized parameters; our proposed GPTFS allows an automatic parametric adaptation in time and frequency for all speech and noise conditions. An optimal tradeoff of the subtraction parameters can be determined based on the unified simultaneous and temporal masking threshold. The other is an unvoiced speech enhancement (USE), which tunes a set of weights in high frequency subbands to improve the intelligibility of the processed speech.

The rest of this Chapter is structured as follows. The architecture of the proposed system is briefly described in Section 4.2. Section 4.3 presents the proposed generalized perceptual time-frequency subtraction method based on the perceptual wavelet filterbank in critical band scale. The subtractive parameters of GPTFS method is adaptively tuned by a unification of simultaneous and temporal masking thresholds in wavelet domain, which is introduced in Chapter 3. Section 4.4 introduces an unvoiced speech enhancement subsystem which is integrated with GPTFS to improve the intelligibility of the processed speech. To evaluate the performance of the proposed method, the objective and subjective measurement results of the proposed system and other competitive methods are shown in Section 4.5. Section 4.6 summarizes this chapter. A part of the work in Sections 4.3 and 4.4 have been presented at the 2005 and 2006 International Symposiums on Circuits and Systems [SHA05a], [SHA06c]. A large portion of the work presented in this chapter has been published in the *IEEE Transactions on System, Man and Cybernetics, Part B* [SHA07a].

4.2 Overview of Proposed Speech Enhancement System

Variants of spectral subtraction method with varying subtraction rules have been derived from different criteria [VAS00]. Most of these algorithms possess parametric forms to improve the agility of their estimators. The focus is mainly on the error function used to reduce the musical residual noise in spectral domain. These methods suffer from the inevitable loss of some vital information that degrades the auditory perception of enhanced speech. Our proposed adaptive speech enhancement system is based on a new GPTFS algorithm that has incorporated the human auditory perception properties and most of the basic subtraction rules. It realizes the subtraction in a broader time-frequency domain. An unvoiced speech enhancement (USE) technique is used to further improve the intelligibility of speech by making the discrimination of the original speech in high frequency bands from noise palpable.

The block diagram of the proposed speech enhancement system is shown in Fig. 4.1. After the noisy signal is decomposed by PWPT, the transform sequence is enhanced by a subtractive-type algorithm to produce the rough speech estimate. This estimate is used to calculate a time–frequency masking threshold. Using this masking threshold, a new subtraction rule, which is masking dependent, is designed to compute an estimation of the original speech. This approach assumes that the high-energy frames of speech will partially mask the input noise (high masking threshold), hence reducing the need for a strong enhancement mechanism. On the other hand, frames containing less speech (low masking threshold) will undergo an overestimated subtraction. To further improve the intelligibility of the processed speech, an unvoiced speech enhancement is applied. Finally, the processed speech is reconstructed by the inverse PWPT. The noise estimation is assumed to be available and is performed during speech pauses. The speech-pause-detection algorithm from Marzinzik and Kollmeier [MAR02] is adopted for the noise spectrum estimation by tracking the power envelope dynamics. The speech-pause-detection algorithm has been extended for subband processing on our proposed system.

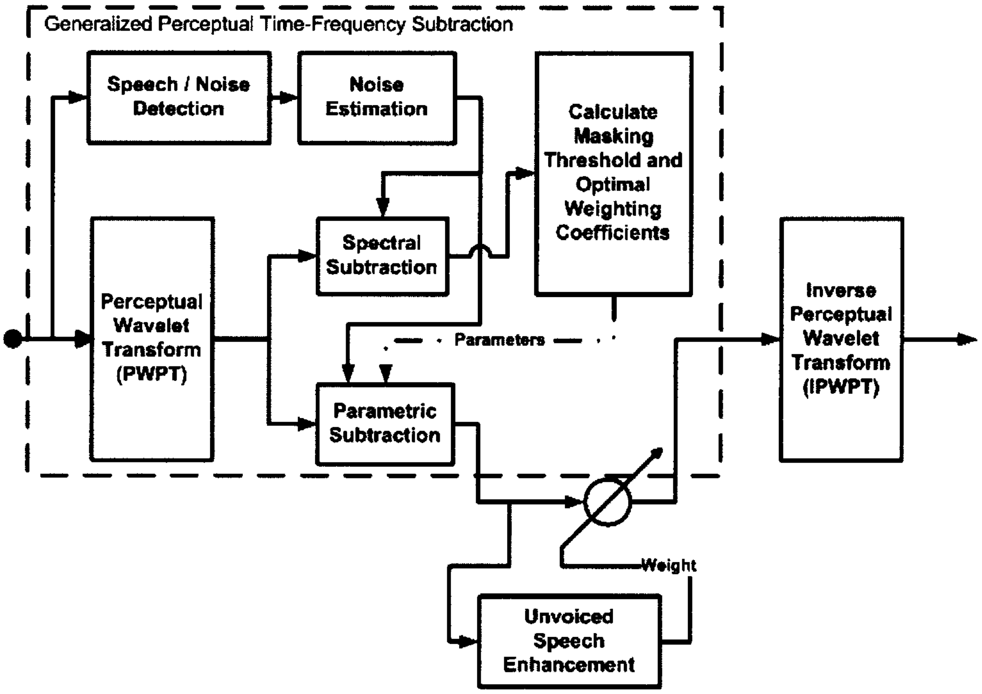

Generalized Perceptual Time-Frequency Subtraction
Speech / Noise Detection
Noise Estimation
Calculate Masking Threshold and Optimal Weighting Coefficients
Spectral Subtraction
Perceptual Wavelet Transform (PWPT)
Parameters
Parametric Subtraction
Inverse Perceptual Wavelet Transform (IPWPT)
Weight
Unvoiced Speech Enhancement

Figure 4.1 Architecture of the proposed speech enhancement system

In what follows, the parametric formulation of the subtractive noise reduction will be derived from the basis of the generalized Fourier spectral subtraction algorithm of [VIR99]. Close form expressions will be derived for the subtraction factor and noise flooring factor to optimize the tradeoff in the simultaneous reduction of background noise, residual noise and speech distortion. These parameters of GPTFS can then be adaptively tuned according to the noise level and the masking thresholds derived from the human auditory model in wavelet domain. Consequently, more crucial information than in Fourier transform domain is preserved, which leads to better perceptual outputs than existing subtraction algorithms.

4.3 Generalized Perceptual Time-Frequency Subtraction

While most speech enhancement algorithms based on mathematical and statistical models of speech/noise signals could work well in some specific conditions, they have made some assumptions about the speech and noise characteristics. However, real-world noise is highly random in nature and the spectral content of speech is very complex due to its variation from speaker to speaker. This problem has inspired us to design a generalized solution that can exploit as much of the palpable properties of the speech and noise signals as possible.

Consider a noisy speech signal consisting of the clean speech additively degraded by noise.

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

where $x[n]$ and $s[n]$ are the noisy and original speech, respectively. $v[n]$is the additive noise, which can be white or colored. Original speech and noise are assumed to be uncorrelated.

Because the real world noise affects the speech signal in a non-uniformly manner over the entire spectrum, our proposed perceptual wavelet packet transform decomposes the noisy input signal in order to capture the localized information of the transient signal.

$$\begin{aligned} &\{w_{j,k}(x)\} = PWPT(x[n]) \\ &w_{j,k}(x) = w_{j,k}(s) + w_{j,k}(v) \quad \forall j = 0,1,..,j_{\max} - 1,\, k = 0,1,...,2^j - 1 \end{aligned} \tag{4.2}$$

where $w_{j,k}(x)$, $w_{j,k}(s)$, and $w_{j,k}(v)$ are the wavelet transform coefficients of noisy signal, clean signal and noise, respectively. (j, k) in the subscript of w corresponds to its scale and translation indices. j_{max} is the maximum number of levels of wavelet decomposition. Unlike the convenient methods, the proposed algorithm can address the frequency components that are more adversely affected by noise through the PWPT analysis. The energy of the noisy input signal is decomposed into bands that closely match the critical band structure.

Next, the subtractive-type algorithm is used to filter the noise. By applying the power spectral subtraction method [VAS00] in the wavelet domain, the estimated power of the enhanced speech is given by:

$$\left|\tilde{w}_{j,k}(s)\right|^2 = \begin{cases} \left|w_{j,k}(x)\right|^2 - \left|\tilde{w}_{j,k}(v)\right|^2, & \text{if } \left|w_{j,k}(x)\right|^2 > \left|\tilde{w}_{j,k}(v)\right|^2 \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

where $\left|\tilde{w}_{j,k}(s)\right|^2$, $\left|w_{j,k}(x)\right|^2$, and $\left|\tilde{w}_{j,k}(v)\right|^2$ are the estimates of the power wavelet coefficients of the noise-suppressed speech signal, the noisy speech and the estimated noise, respectively in the wavelet domain.

Since there is no prior knowledge of the parameters of $s[n]$ and $v[n]$, and only $x[n]$ is accessible, the noise power spectrum in (4.3) is to be estimated from the time-average noise magnitude during the period in which the speech signal is absent. The time-average noise spectrum can be obtained by [VAS00]:

$$\left|\tilde{w}_{i,j,k}(v)\right|^2 = 0.9 \cdot \left|w_{i-1,j,k}(v)\right|^2 + 0.1 \cdot \left|w_{i-1,j,k}(x)\right|^2 \tag{4.4}$$

where the subscript, i is the frame index.

The design takes into account that human speech is based on mixing voiced and unvoiced phonemes over time. Hence, $s[n]$ is non-stationary over the time interval of above 250 ms, and the noise $v[n]$ is assumed to be piecewise-stationary or more stationary than $s[n]$, which is valid for most environmental noises encountered. With the above assumption, the input signal $x[n]$ is divided into 128 samples.

To analysis each component in the estimated speech signal, substituting (4.2b) into (4.3), we have

$$\begin{aligned} \left|\tilde{w}_{j,k}(s)\right|^2 &= \left|w_{j,k}(s) + w_{j,k}(v)\right|^2 - \left|\tilde{w}_{j,k}(v)\right|^2 \\ &= \left|w_{j,k}(s)\right|^2 + \underbrace{\left(\left|w_{j,k}(v)\right|^2 - \left|\tilde{w}_{j,k}(v)\right|^2\right)}_{\text{Noise Variations}} + \underbrace{w^*_{j,k}(s)\,w_{j,k}(v) + w_{j,k}(s)\,w^*_{j,k}(v)}_{\text{Cross Products}} \end{aligned} \tag{4.5}$$

The wavelet power spectrum of the estimated speech signal includes error terms due to the non-linear mapping of the spectral estimates, the variations of the instantaneous noise power about the estimated mean noise power, and the cross products of the original speech and noise. Since signal and noise are assumed to be uncorrelated, and the noise wavelet coefficients are assumed to have zero mean, the expected value of the power spectrum of estimated speech signal in wavelet domain is given by:

$$\begin{aligned} E\left[\left|\tilde{w}_{j,k}(s)\right|^2\right] &= E\left[\left|w_{j,k}(s)\right|^2 + \underbrace{\left(\left|w_{j,k}(v)\right|^2 - \left|\tilde{w}_{j,k}(v)\right|^2\right)}_{\text{Noise Variations}} + \underbrace{w^*_{j,k}(s)\,w_{j,k}(v) + w_{j,k}(s)\,w^*_{j,k}(v)}_{\text{Cross Products}}\right] \\ &= E\left[\left|w_{j,k}(s)\right|^2\right] \end{aligned} \tag{4.6}$$

(4.6) shows that the expected value of the instantaneous power spectrum of estimated speech signal converges to that of the noise-free signal. However, it is noted that for non-stationary signals, such as speech, the objective is to recover the instantaneous or short-time signal, and only a relatively small amount of averaging can be applied. Too much averaging will obscure the temporal evaluation of the signal. The spectral subtraction in (4.3) can be deemed as an adaptive attenuation of noisy speech according to the *posteriori* signal to noise (SNR) ratio of the noisy speech signal to the estimated noise. An approach of filtering noisy speech could be derived. A gain function, $Gain(j,k)$ for the zero-phase wavelet spectral subtraction filter can be defined such that its magnitude response is in the range between 0 and 1, i.e.,

$$\tilde{w}_{j,k}(s) = Gain(j,k)\cdot w_{j,k}(x), \qquad 0 \le Gain(j,k) \le 1 \tag{4.7}$$

Comparing (4.3) and (4.7), we have

$$Gain(j,k) = \sqrt{1 - \frac{\left|\tilde{w}_{j,k}(v)\right|^2}{\left|w_{j,k}(x)\right|^2}} = \sqrt{1 - \frac{1}{SNR_{post}(j,k)}} \tag{4.8}$$

where $SNR_{post}(j,k) = \left|w_{j,k}(x)\right|^2 / \left|\tilde{w}_{j,k}(v)\right|^2$ is the *posteriori* SNR, which is defined as the ratio of the wavelet power spectrum of the noisy speech signal to that of the estimated noise. In the low SNR case, the gain is set to zero when the wavelet power spectrum of the noisy speech is less than that of the estimated noise.

The gain function in (4.8) acts as *a posteriori* SNR-dependent attenuator in the time-frequency domain. The input noisy speech is attenuated more heavily with decreasing *posteriori* SNR and vice versa with increasing *posteriori* SNR. This filtering function can be readily implemented since the *posteriori SNR* can be easily measured on the noisy speech. Although it can effectively filter out the background noise from the noisy speech, the processing itself introduces musical residual noise in the enhanced speech. This processing distortion is due to the random variation of noise spectrum and the non-linear mapping of the negative or small-valued spectral estimate. To make this residual noise “perceptually white”, Virag [VIR99] introduced several flexible subtraction parameters into the spectral subtraction method. These subtraction parameters are chosen to adapt to a criterion associated with the human auditory perception. The generalized spectral subtraction method in [VIR99] is based on the Fourier transform. It considers only the simultaneous masking property of human auditory system, and the adaptation parameters are empirically determined. In what follows, we will state the generalized time-frequency subtraction method in the perceptual wavelet packet transform domain and derive the close form expressions for its optimal adaptation parameters. Unlike the conventional speech enhancement methods [VIR99], [HU04], the proposed method adopts a multiresolution human auditory model. This wavelet packet based transformation closely models the signal processing in the periphery of the auditory system.

From (4.8), the gain function of Virag's generalized spectral subtraction algorithm [VIR99] can be re-expressed in terms of perceptual wavelet spectral coefficients as follows:

$$Gain'(j,k)=\begin{cases}\left(1-\alpha\cdot\left[\dfrac{\tilde{w}_{j,k}(v)}{w_{j,k}(x)}\right]^{\gamma_1}\right)^{\gamma_2} & if\left[\dfrac{\tilde{w}_{j,k}(v)}{w_{j,k}(x)}\right]^{\gamma_1}<\dfrac{1}{\alpha+\beta}\\ \left(\beta\cdot\left[\dfrac{\tilde{w}_{j,k}(v)}{w_{j,k}(x)}\right]^{\gamma_1}\right)^{\gamma_2} & otherwise\end{cases} \tag{4.9}$$

The PWPT based gain function of (4.9) possesses the same set of free parameters as that of [VIR99], and their effects are summarized as follows: (1) The *subtraction factor* α ($\alpha \geq 1$) controls the amount of noise subtracted from the noisy signal. For full noise reduction, $\alpha = 1$ whereas for over-subtraction $\alpha > 1$. Over-subtraction allows the time-frequency spectrum to be attenuated more than necessary. This factor should be appropriately selected to provide the best tradeoff between residual noise peaks and audible distortion. (2) The *noise flooring factor* β ($0 \leq \beta < 1$) makes use of the addition of background noise to mask the residual noise. It determines the minimum value of the gain function. If this factor is increased, parts of residual noise can be masked, but the level of background noise retained in the enhanced speech increases. (3) The *exponent* $\gamma = \gamma_1 = 1/\gamma_2$ determines the abruptness of the transition from pure clean speech to pure noise in noisy speech. For magnitude subtraction, $\gamma_1 = \gamma_2 = 1$ whereas for power subtraction, $\gamma_1 = 2$, $\gamma_2 = 0.5$.

To formally select these adaptation parameters, we will optimize the fixed gain function of (4.8) by minimizing the speech distortion and residue noise based on the masking threshold determined in Chapter 3, Section 3.4. To segregate the residual noise from the speech distortion, we define a differential wavelet coefficient as the difference between the wavelet coefficients of the clean speech and the enhanced speech. The differential wavelet coefficients can be expressed as follows:

$$\begin{aligned}\partial w_{j,k} &= \tilde{w}_{j,k}(s) - w_{j,k}(s) = Gain(j,k)\cdot w_{j,k}(x) - w_{j,k}(s) \\ &= w_{j,k}(s)\left(Gain(j,k)-1\right) + Gain(j,k)\cdot w_{j,k}(v)\end{aligned} \tag{4.10}$$

The optimal gain function will be derived by following the same approach as in [LU04]. The difference is that our gain function will be optimized based on the thresholding criterion that correlates with both temporal and simultaneous maskings. The auditory perception is better approximated by using more complex wavelet basis and efficient filter bank structure. The differential wavelet coefficients can be viewed as a linear superposition of speech distortion, $\xi_{j,k}(s)$ and residual noise, $\xi_{j,k}(v)$ in the wavelet domain, where the speech distortion, $\xi_{j,k}(s)$ and residual noise, $\xi_{j,k}(v)$ are defined as:

$$\xi_{j,k}(s) = w_{j,k}(s)\left(Gain(j,k)-1\right) \tag{4.11}$$

$$\xi_{j,k}(v) = w_{j,k}(v)\,Gain(j,k) \tag{4.12}$$

The energies of the speech distortion and the residual noise are respectively given by:

$$E_{\xi_{j,k}(s)} = \sum_{k=0}^{2^{-j}-1} \left|w_{j,k}(s)\right|^2 \cdot \left|\left(Gain(j,k)-1\right)\right|^2 \tag{4.13}$$

$$E_{\xi_{j,k}(v)} = \sum_{k=0}^{2^{-j}-1} \left|w_{j,k}(v)\right|^2 \cdot \left|\left(Gain(j,k)\right)\right|^2 \tag{4.14}$$

Since the residual noise is not audible when its energy is smaller than the masking threshold, we will suppress only the wavelet coefficients of noisy speech when the noise level is greater than the masking threshold as in [LU04]. To optimize the gain function, we minimize the energy of speech distortion in wavelet domain subjected to the constraint that the energy of residual noise is kept below the masking threshold. A cost function *J*(*j*, *k*) can be formulated according to the energies of the speech distortion and the residual noise expressed in terms of the wavelet coefficients.

$$J(j,k)=E_{\xi_{j,k}(s)}+\eta_{j,k}(Z_W)\cdot\left\{E_{\xi_{j,k}(s)}-T_{j,k}(Z_W)\right\} \tag{4.15}$$

where the Largrangian multiplier of each subband, $\eta_{j,k}(Z_W)$ serves as a weighting factor to the residue noise. Substituting (4.13) and (4.14) into (4.15), we have

$$\begin{aligned}J(j,k)&=\sum_{k=0}^{2^{-j}-1}\left|w_{j,k}(s)\right|^2\cdot\left|\left(Gain(j,k)-1\right)\right|^2\\&+\eta_{j,k}(Z_W)\cdot\left\{\sum_{k=0}^{2^{-j}-1}\left|w_{j,k}(v)\right|^2\cdot\left|\left(Gain(j,k)\right)\right|^2-T_{j,k}(Z_W)\right\}\end{aligned} \tag{4.16}$$

The optimal gain function is obtained by minimizing the cost function *J*(*j*, *k*). Taking the derivative of the cost function *J*(*j*, *k*) with respect the Largrangian multiplier, $\eta_{j,k}(Z_W)$ and set the result to zero, the optimal gain function is determined.

$$\begin{aligned}&\frac{dJ(j,k)}{d\eta_{j,k}(Z_W)}=\sum_{k=0}^{2^{-j}-1}\left|w_{j,k}(v)\right|^2\cdot\left|\left(Gain(j,k)\right)\right|^2-T_{j,k}(Z_W)=0\\&Gain_{opt}(j,k)=\sqrt{\frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-j}-1}\left|w_{j,k}(v)\right|^2}},\qquad 0\le Gain(j,k)\le 1\end{aligned} \tag{4.17}$$

By equating the adaptive gain function of (4.9) to the optimal gain function of (4.17), and considering the power subtraction, i.e.,γ_1 = 2 and γ_2 = 0.5, the closed form expressions for the subtraction parameters, α and β can be derived as follows:

$$1-\alpha\cdot\frac{1}{SNR(j,k)}=\frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-J}-1}\left|w_{j,k}(v)\right|^2}\quad\Rightarrow\quad\alpha=SNR(j,k)\cdot\left(1-\frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-J}-1}\left|w_{j,k}(v)\right|^2}\right)\tag{4.18}$$

$$\beta\cdot\frac{1}{SNR(j,k)}=\frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-J}-1}\left|w_{j,k}(v)\right|^2}\quad\Rightarrow\quad\beta=SNR(j,k)\cdot\frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-J}-1}\left|w_{j,k}(v)\right|^2}\tag{4.19}$$

(4.18) and (4.19) ensure that the subtraction parameters α and β are adapted to the masking threshold of human auditory system to achieve a good tradeoff between residual noise, speech distortion and background noise. In high SNR condition, the parameter α is increased to reduce the residual noise at the expense of introducing more speech distortion. On the contrary, in low SNR condition, the parameter β is increased to trade the residual noise reduction for an increased background noise in the enhanced speech. To make the residual noise inaudible, the subtraction parameters are set such that the residual noise stays just below the masking threshold, $T_{j,k}(Z_W)$. If the masking threshold is low, the subtraction parameters will be increased to reduce the effect of the residual noise. If the masking threshold is high, the residual noise will naturally be masked and become inaudible. Therefore, the subtraction parameters can be kept at their minimal values to minimize the speech distortion. Different from the traditional methods, the proposed method considers the SNR variation in different frequency bands of the speech corrupted with noise. To make use of the multiresolution auditory model, the optimal subtractive-factors are estimated by the proposed method to subtract just the necessary amount of the noise spectrum from each frequency band. This will prevent damaging over-subtraction of the original speech while removing most of the residual noise.

4.4 Unvoiced Speech Enhancement

Although the GPTFS subsystem is useful for enhancing the portion of the speech signal that contains most of the signal energy, they sometimes degrade the high frequency bands of low energy. To alleviate this problem, we apply a soft thresholding method to enhance the portion of speech signal that lies in the high frequency bands [SHA05a].

Despite the fact that most energy of the speech signal is concentrated around the lower frequency bands, the speech energy in the high frequency range may possess considerably high peak and carry crucial perception information. To enhance the portion of the speech signal that lies in the high frequency range without degrading the performance of the overall system, the unvoiced speech enhancement unequally weights the frequency bands to amplify only those components with detectable peaks in the high frequency range. The time-frequency energy (TFE), which is estimated using the wavelet coefficients, is applied in this subsystem. The TFE is computed as:

$$\Gamma_{j,k} = \sum_{F_j} \left(\tilde{w}_{j,k}(x) \right)^2 \Big/ \sum_{n=1}^{F_{j\max}} \left\| x[n] \right\|^2 \tag{4.20}$$

where $\|.\|$ is the norm operator. F_{jmax} is the total frame length. F_j is the frame length of each subband. The TFE mapping, Γ is a normalized energy function of wavelet coefficients at each position of the binary tree. To estimate the enhanced original speech, we assume that the TFE of noise does not change much over time. Since only noisy observation is accessible, we can only estimate the noise TFE by the minimum TFE of the observation. The high-frequency time-frequency energy peak can be deduced as:

$$\tilde{\Gamma}_{j,k}(s) = \Gamma_{j,k}(x) - \tilde{\Gamma}_{j,k}(v) = \Gamma_{j,k}(x) - \min_{1 \le m \le M} \Gamma^{m}_{j,k}(x) \tag{4.21}$$

where $\Gamma_{j,k}(x)$ denote the TFE of the noisy speech which is composed of the TFEs of the noise, $\Gamma_{j,k}(v)$ and the original speech, $\Gamma_{j,k}(s)$. The superscript, m denotes the frame index. Unlike the GPTFS, which operates on each time frame, the unvoiced speech enhancement spans over several frames, where M indicates the number of frames covered in this duration.

To amplify those high frequency bands containing components of unvoiced speech without affecting all other high frequency bands, we define a threshold $\delta_{j,k}$. Different wavelet coefficients of the processed speech are then emphasized via their weights, $u_{j,k}$. The weight, $u_{j,k}$ is obtained by:

$$u_{j,k} = \begin{cases} 1 & \widehat{\Gamma}_{j,k}(s) < \delta_{j,k} \\ \sum_{m=1}^{M} \left(\Gamma^{m}_{j,k}(x) - \widehat{\Gamma}_{j,k}(v) \right) \Big/ \widehat{\Gamma}_{j,k}(v), & \widehat{\Gamma}_{j,k}(s) > \delta_{j,k} \end{cases} \tag{4.22}$$

where the threshold, $\delta_{j,k}$ is determined empirically through experimentation, below which the subband is noise-free and $u_{j,k}$ is set to 1. Therefore, when the noise estimate approaches zero and the speech is strong enough, $u_{j,k}$ = 1. The weighted wavelet coefficients are given by:

$$\widehat{w}_{j,k}(s) = u_{j,k} \cdot \tilde{w}_{j,k}(s) \tag{4.23}$$

The GPTFS either amplifies or attenuates a particular frequency band based on the estimated signal energy content in low frequency. When the SNR is high, the unvoiced speech enhancement is effective. In the case of low SNR, the GPTFS have suppressed most energy of noise while significantly reduced the unvoiced speech at the same time. Nevertheless, the succeeding unvoiced speech enhancement can still coarsely estimate the noise and tune a set of weights to enhance the unvoiced speech somewhat. Finally, an inverse wavelet transform [STR96] is applied to re-synthesize the enhanced speech.

$$\hat{s}(n) = IPWPT\left(\hat{w}_{j,k}(x)\right) \tag{4.24}$$

where *IPWPT* means inverse perceptual wavelet packet transform.

4.5 Results and Discussions

In this section, the proposed system is evaluated with speeches produced in various adverse conditions and compared against the following competitive methods: (1) Speech enhancement method using perceptually constrained gain factors in critical-band-wavelet-packet transform (Lu04) [LU04], (2) Speech enhancement method incorporating a psychoacoustical model in frequency domain (Hu04) [HU04], (3) Wavelet speech enhancement method based on the teager energy operator (B&R01) [BAH01], (4) Perceptual time-frequency subtraction algorithm (Li01) [LI01], (5) Single channel speech enhancement based on masking properties of human auditory system (Vir99) [VIR99], (6) Parametric spectral subtraction (Sim98) [SIM98]. The original speeches used for the tests were taken from the TIMIT database [TIM90]. Different background noises with different time-frequency distributions were taken from the Noisex-92 database [NOI92]. The tested noisy environments include white Gaussian noise, pink noise, Volvo engine noise, F16 cockpit noise, factory noise, high frequency channel noise and speech-like noise. Noise has been added to the clean speech signal with different SNR's. The same speech-pause-detection algorithm from Marzinzik and Kollmeier [MAR02] is used for all algorithms being compared that require noise estimation. This speech-pause-detection algorithm has a low hit rate of 0.11 and 0.12, and a low false-alarm rate of 0.085 and 0.08 at SNR of -5dB and -10dB, respectively in the speech-like noise environment.

4.5.1 SNR Improvement and Itakura-Saito distortion

The amount of noise reduction is generally measured in terms of the SNR improvement, given by the difference between the input and output segmental SNRs. The pre-SNR and post-SNR are defined as:

$$SNR_{pre} = \frac{1}{K}\sum_{m=0}^{K-1}\left(20\cdot\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}s(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}v(n+Nm)}\right)\right) \tag{4.25}$$

$$SNR_{post} = \frac{1}{K}\sum_{m=0}^{K-1}\left(20\cdot\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}s(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}\left(s(n+Nm)-\widehat{s}(n+Nm)\right)}\right)\right) \tag{4.26}$$

where K represents the number of frames in the signal and N indicates the number of samples per frame. The segment SNR improvement is defined as $SNR_{post} - SNR_{pre}$:

$$SNR_{imp} = \frac{1}{K}\sum_{m=0}^{K-1}\left(20\cdot\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}v(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}\left(s(n+Nm)-\widehat{s}(n+Nm)\right)}\right)\right) \tag{4.27}$$

This equation takes into account both residual noise and speech distortion. Figs. 4.2(a) to 4.5(a) compare the SNR improvement of various speech enhancement methods in white noise, Volvo engine noise, factory noise and speech-like noise of different noise levels. Our proposed method produces higher SNR improvement than other methods, particularly for low input SNR's. In the proposed algorithm, the musical structure of the residual noise has been reduced more than the algorithms of [HU04] and [VIR99] which have been known to achieve the best results among the competitive algorithms. Speech enhanced with the proposed method is more pleasant. The residual noise has a "perceptually white" quality and the distortion remains acceptably low. From Figs. 4.3 - 4.5, we can see that the musical structure of residual noise has also been greatly reduced, as long as the noise remains stationary. As the nonstationarity of noise increases, the results become poorer. This is because the noise estimate cannot follow the variations of the background noise. Thus, the best noise reduction of the proposed method is obtained from white Gaussian noise. For other types of noise, the improvement is less prominent.

It has been shown by experiments that even though the SNR's are very similar at the output of the enhancement system, the listening test and speech spectrograms can produce very divergent results. Therefore, segmental SNR alone is not sufficient to indicate the speech quality and further objective measure is needed to attest the performance ascendancy. Itakura-Saito distortion (IS) [NOC85] provides a higher correlation with subjective results than the SNR for speech processing systems. It is derived from the Linear Predictive (LP) coefficient vector, $\alpha_s(m)$ of the original clean speech frame and the processed speech coefficient vector, $\alpha_{\hat{s}}(m)$ as follows:

$$IS(m)=\frac{\sigma_{s}^{2}(m)}{\sigma_{\hat{s}}^{2}(m)}\cdot\frac{\alpha_{\hat{s}}(m)R_{s}(m)\alpha_{\hat{s}}^{T}(m)}{\alpha_{s}(m)R_{s}(m)\alpha_{s}^{T}(m)}+\log\left(\frac{\sigma_{\hat{s}}^{2}(m)}{\sigma_{s}^{2}(m)}\right)-1 \tag{4.28}$$

where $\sigma_{\hat{s}}^{2}(m)$ and $\sigma_{s}^{2}(m)$ are the all-pole gains for the processed and clean speeches, respectively, and $R_s(m)$ denotes the clean speech signal correlation matrix. The IS distortion measure gives a zero response when the distortion of two signals to be estimated have equal LP coefficients and gain. Smaller value of IS implies better speech quality.

Figs. 4.2(b) to 4.5(b) show the IS distortion of various methods in different noise environments at varying noise levels. All figures show that the proposed method outperforms other methods with significant margins. It is evident that the GPTFS algorithm provides the best performance when the subtraction process is preceded by the generalized time-frequency subtraction incorporated with perceptual wavelet filterbank. This is attributed to the holistic reduction of the background noise, residual noise and speech distortion. In low SNR, GPTFS can suppress most of the noise by appropriately tuning the subtraction parameters according to the time-frequency masking threshold. This tradeoff control helps to minimize the processing distortion in the enhanced speech. The USE of the proposed system compensates for the speech in the high frequency range. It further improves the intelligibility and reduces the distortion. When the SNR is high, the proposed system has excellent performance. Our proposed system has successfully overcome the major drawback of traditional de-noising methods, which is the tendency to deteriorate some useful components of speech, resulting in the weakening of its intelligibility. However, the worst results are obtained with speechlike noise. Indeed, this noise is particularly difficult to handle, because it has the same frequency distribution as long-term speech. In summary, for both measures of SNR and IS, the proposed algorithm achieves a significant improvement over classical subtractive-type algorithms.

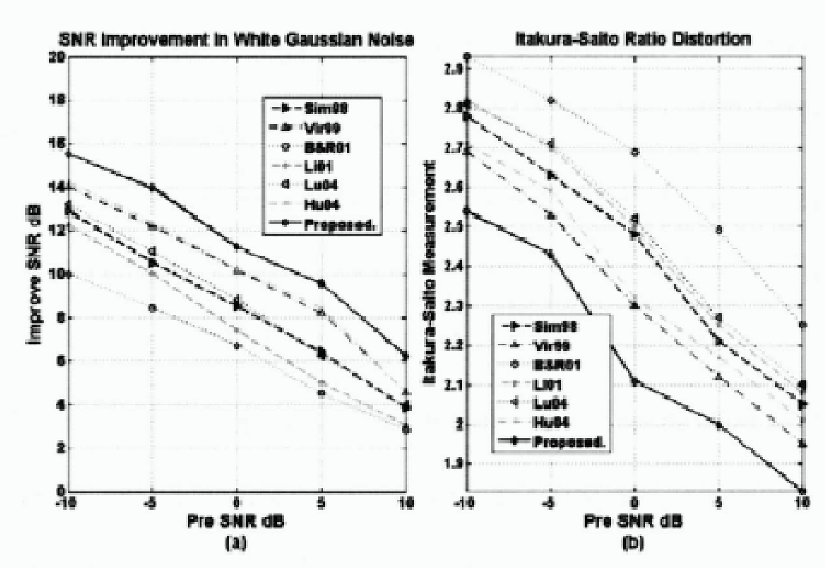


Figure 4.2 Comparison of different speech enhancement methods in white Gaussian noise by (a) SNR improvement (b) Itakura-Saito (IS) distortion

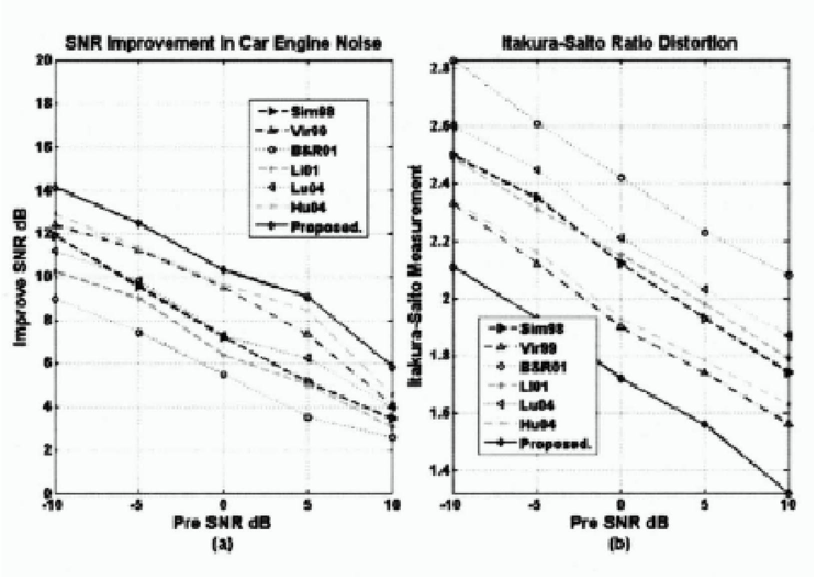


Figure 4.3 Comparison of different speech enhancement methods in Volvo engine noise by (a) SNR improvement (b) Itakura-Saito (IS) distortion

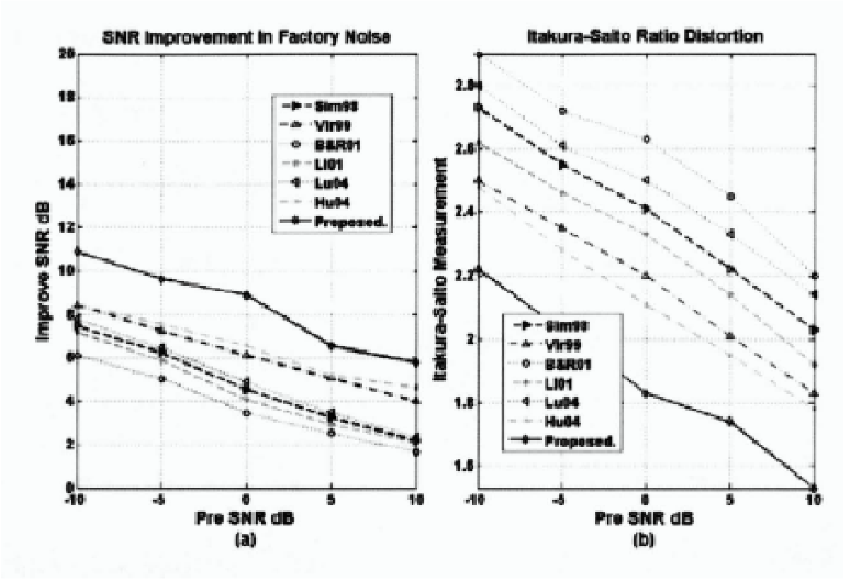


Figure 4.4 Comparison of different speech enhancement methods in Factory noise by (a) SNR improvement (b) Itakura-Saito (IS) distortion

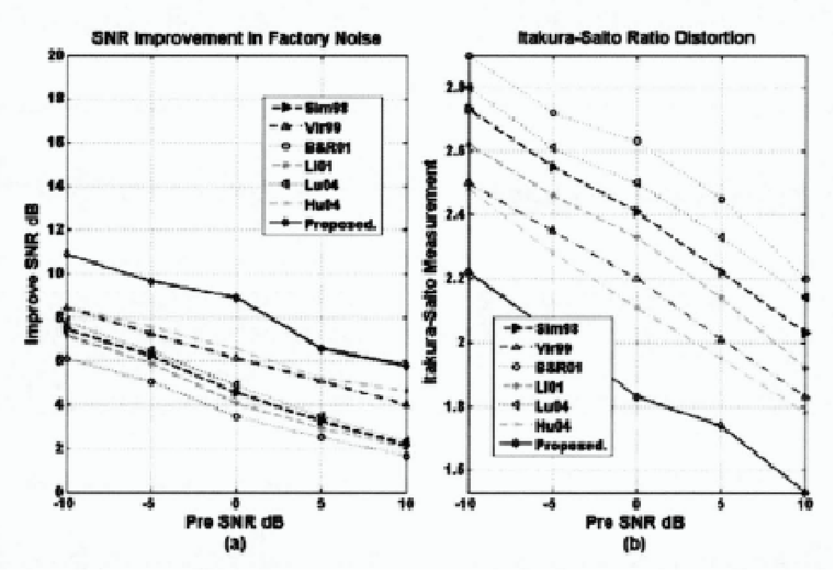
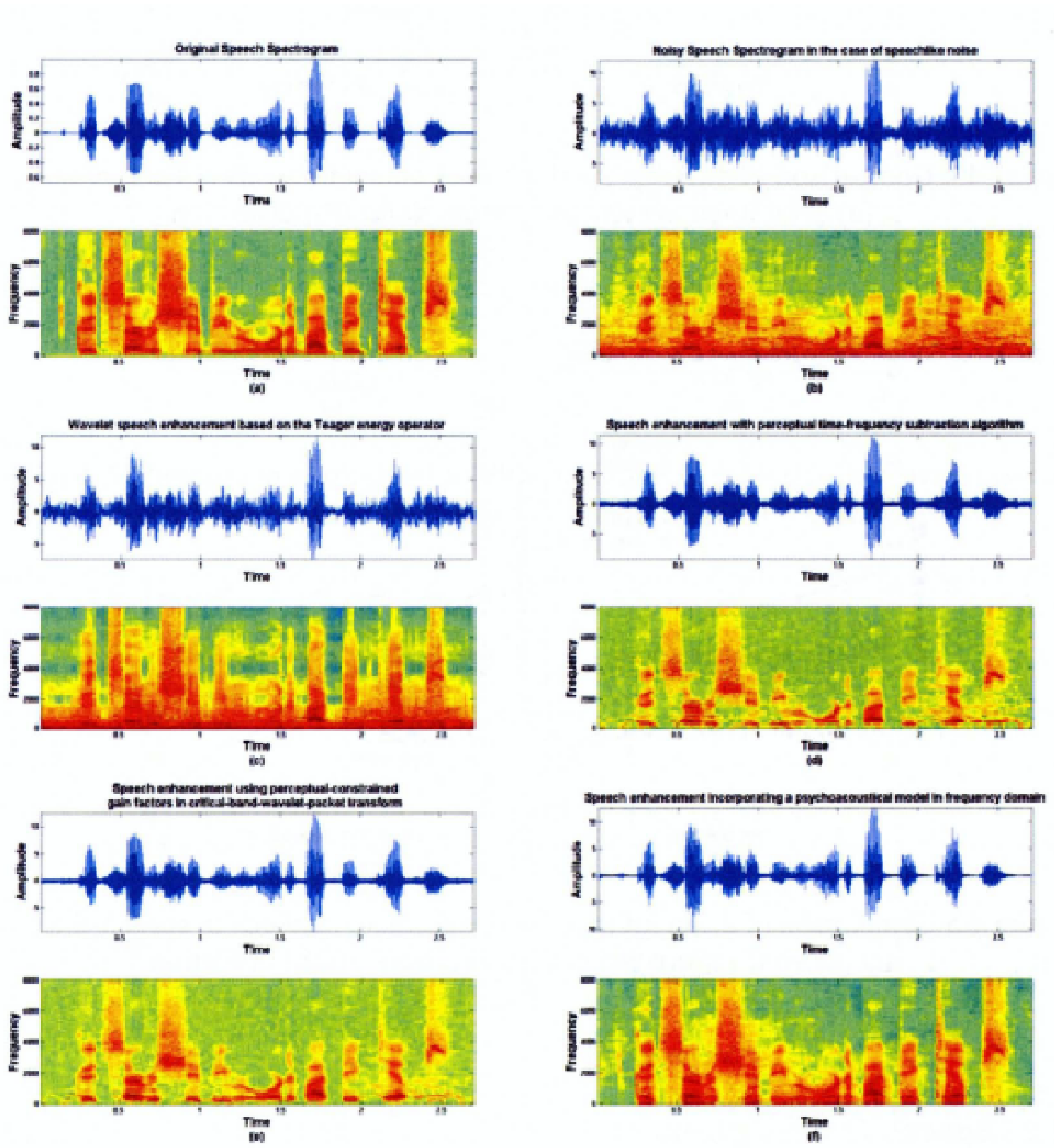


Figure 4.5 Comparison of different speech enhancement methods in speech-like noise by (a) SNR improvement (b) Itakura-Saito (IS) distortion

4.5.2 Speech Spectrograms

The above objective measures do not provide information about how speech and noise are distributed across frequencies. In this perspective, the speech spectrograms, which yields more accurate information about the residual noise and speech distortion than the corresponding timing waveforms are analyzed. All the speech spectrograms presented in this section used Kaiser Window of 500 samples with an overlap of 475 samples. Experiments show that the noisy phase is not perceived as long as the local SNR is greater than about 6 dB. However, at SNR = 0 dB, the effect of the noisy phase is audible and is perceived as an increased roughness.



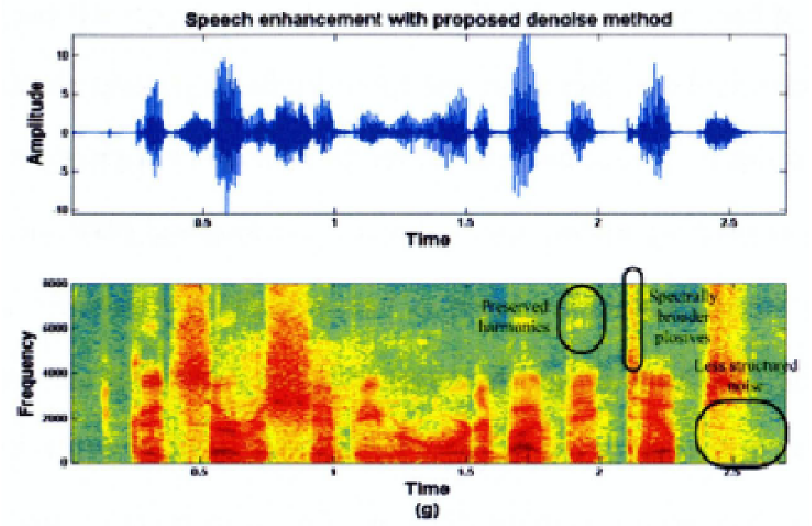


Figure 4.6 Speech spectrograms (a) Original clean speech. (b) Noisy signal (additive Speechlike noise at a SNR = 0 dB). (c) Speech enhanced by B&R01 [BAH01] (d) Speech enhanced by Li01 [LI01] (e) Speech enhanced by Lu04 [LU04] (f) Speech enhanced by Hu04 [HU04] (g) Speech enhanced by the proposed method

The spectrograms of noisy and enhanced speeches obtained with the proposed method in various noise environments show that the residual noise has been greatly reduced, as long as the noise remains stationary or piecewise stationary. The worst results of the proposed method are obtained with speech-like noise. This noise is particularly difficult to handle by any method, because it has the same frequency distribution as the long-term speech. Fig. 4.6 shows the spectrograms of a noisy speech sample filtered with the proposed GPTFS method and other conventional subtractive-type approaches, respectively. The noisy signal is perturbed by speech-like noise, which has been recorded inside a cafeteria. It consists of nonstationary speech bursts and a relatively stationary floor due to reverberation. As shown in the figure, the conventional filter could only suppress the stationary portion but not the voiced bursts in babble. The filtered babble contains unnatural tonal residuals. The proposed GPTFS method not only effectively suppresses the stationary portions of the speech-like noise, but also prevents tonal contents from being emphasized. In particular, the proposed method can better preserve important structures of speech. In the spectrograms, it can be seen that pitch harmonics are better preserved, and spectrally broad plosives are less distorted. As the non-stationary part of noise increases, the results of our proposed method are still better than other algorithms. This is because the proposed PWPT filter bank has closely approximated the critical bands of human auditory system. By appropriately segregating the speech and noise components of the noisy speech in frequency and time, the subtraction parameters of our proposed GPTFS method adapt well to the combined temporal and simultaneous masking threshold. The supplement of unvoiced speech enhancement also improves the performance of our proposed system in high frequency noise.

4.5.3 Subjective Measures

In order to validate the performances of objective evaluation, subjective listening tests are performed by subjecting the speech to varying noise at SNR = 0 and 10 dB before they are processed by different algorithms. An informal subjective test was performed with ten listeners. The testing has been performed with headphone listener in laboratory environment. Each listener gives a score between one and five to each test signal. The score represents his global perception of the residual noise, background noise and speech distortion. The scale used for these tests corresponds to the mean opinion score (MOS) scale. MOS is a live listener test designed to yield a single numeric score that rates the perceived quality of speech of the analyzed audio sample [DEL93]. For each tester, the following testing procedure has been applied: first, the clean and noisy speeches are played and repeated twice; second, each test signal, which is repeated twice for each score, is played three times in a random order. This leads to 30 scores for each test signal. The MOS score is the average of all the ranks voted by the different listeners of the different voice files used in the experiment. The results are presented in Table 4.2. This test has been conducted in a carefully-controlled listener environment and the raw test scores have been carefully analyzed for validity. The subjective listening tests confirm that the proposed enhancement method produces the highest quality speech perceived by the actual human listeners among the algorithms being tested.

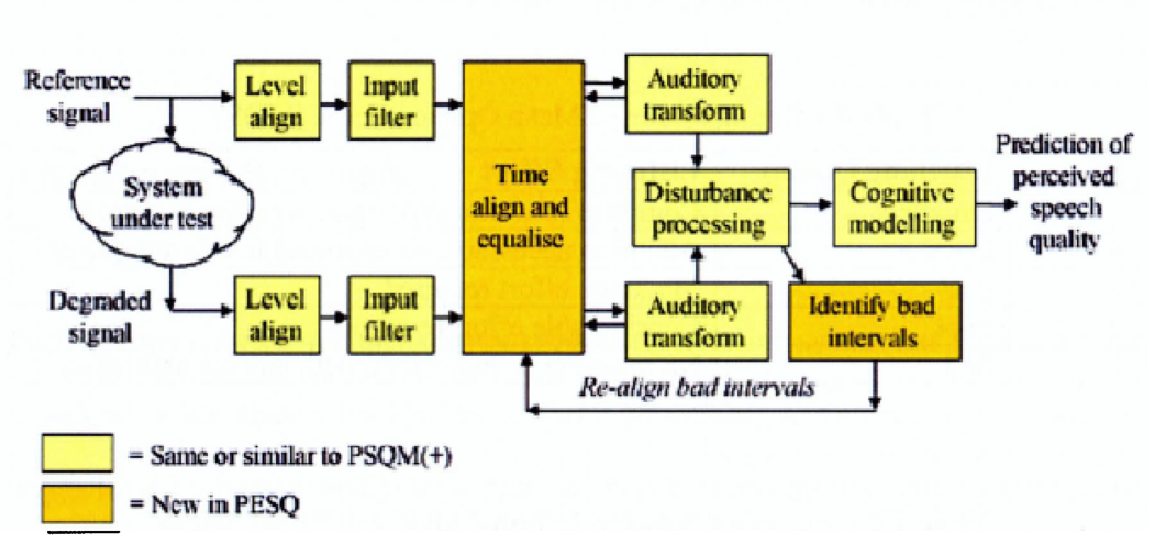


Figure 4.7 The block diagram of PESQ (Figure excerpted from [ITU01])

MOS is considered to be the authoritative way to rate perceived speech quality in communication system. However, it is more time-consuming and costly than automated quality analysis techniques, e.g. perceptual speech quality measure (PSQM) and perceptual evaluation of speech quality (PESQ). The common idea behind perceptual quality measures is to mimic the situation of a subjective test, where human beings would have to score the quality of sound samples in a listening laboratory environment. To prove that the trend of the objective Perceptual Evaluation of Speech Quality measure matches well with the informal MOS scores [ITU01], the perceptual quality measure of processed speech has been performed. The PESQ algorithm is an automated speech quality analysis technique that predicts the subjective opinion score of a degraded audio sample and the listening effort for the collected speech samples. Fig. 4.7 shows the block diagram of PESQ. This repeatable and objective test derives a set of scores by comparing one or more high-quality reference speech samples to the processed (degraded) resulting audio. PESQ is an automated technique that can be performed quickly and it returns a score from 4.5 to –0.5, with higher score indicating better quality. The PESQ listening quality and listening effort scores are based upon a five point category judgment scale as shown in Table 4.1.

Table 4.1 Configuration of Mean Opinion Score (MOS)

Score	**Listening Quality**	**Listening Effort**
5	Excellent	Complete relaxation possible; no effort required
4	Good	Attention necessary; no appreciable effort required
3	Fair	Moderate effort required
2	Poor	Considerable effort required
1	Bad	No meaning is understood with any feasible effort

Table 4.2 Comparison between Informal MOS and PESQ Scores

Informal MOS score				
Noise type, SNR$_i$	*Noisy signal*	*Lu04 [13]*	*Hu04 [10]*	*Proposed*
White, 0 dB	1.35	1.46	1.55	1.78
White, 10 dB	2.53	2.35	2.64	3.00
Car engine, 0 dB	1.59	1.59	1.71	1.89
Car engine, 10 dB	2.22	2.22	2.36	2.60
Factory , 0 dB	1.45	1.50	1.70	2.00
Factory, 10 dB	2.62	2.45	2.77	2.85
Speechlike, 0 dB	1.51	1.43	1.66	1.90
Speechlike, 10 dB	2.43	2.33	2.58	2.71
PESQ score				
White, 0 dB	1.24	1.32	1.48	1.85
White, 10 dB	1.94	2.00	2.17	2.45
Car engine, 0 dB	1.42	1.51	1.65	1.78
Car engine, 10 dB	2.11	2.04	2.25	2.39
Factory , 0 dB	1.51	1.41	1.65	1.95
Factory, 10 dB	2.16	2.11	2.37	2.50
Speechlike, 0 dB	1.22	1.14	1.39	1.70
Speechlike, 10 dB	1.88	1.89	2.05	2.35

The PESQ scores are also indicated in Table 4.2. Form Table 4.2, it is observed that the gap between the enhanced speech and noisy speech scores is similar in both MOS and PESQ scores. It is observed that for all the noise conditions, the performance trend of the three speech enhancement methods remains well correlated in both MOS and PESQ scores. The best scores are consistently obtained by our proposed method for different types of noisy speech. The MOS and PESQ of the proposed method are improved by 9.4% and 11% over [HU04], respectively, under all types of noise and SNR conditions.

4.6 Chapter Summary

In this chapter, a new subtractive-type speech enhancement system has been proposed. It can be anatomized into several powerful processing techniques that exploit the physiology of human auditory system to recover high quality speech from noise contaminated speech. The system consists of two functional stages working cooperatively to perform perceptual time-frequency subtraction by adapting the weights of the perceptual wavelet coefficients. The noisy speech is first decomposed into critical bands by perceptual wavelet transform. The temporal and spectral psychoacoustic model of masking developed in Chapter 3 is used to calculate the threshold to be applied to GPTFS method for noise reduction. Different spectral resolutions of the wavelet representation preserve the energy of the critical transient components so that the background noises, distortion, and residual noise can be adaptively processed by GPTFS method. The unvoiced speech is also enhanced by a soft-thresholding scheme. The effectiveness of the proposed system to extract a clear and intelligible speech from various adverse noisy environments in comparison with other well-known methods has been demonstrated through both objective and subjective measurements. The performance robustness under varying signal-to-noise conditions of the proposed system is attributable to the consideration of both the temporal and simultaneous maskings for the tuning of subtraction parameters in the proposed GPTFS. Together with the unvoiced speech enhancement, they make on average, a SNR improvement of 5.5% over [HU04] by objective measurements, and an average intelligibility improvement of 8% by subjective evaluation.

CHAPTER 5.

A PERCEPTUAL WAVELET KALMAN FILTERING SPEECH ENHANCEMENT METHOD

5.1 Introduction

The speech enhancement method proposed in Chapter 4 utilizes short-time spectral amplitude estimation for parametric subtraction in the time-frequency domain. The parametric formulation applies solely to the salient subtraction parameters, which are optimally tuned according to the masking threshold of human auditory system. Strictly speaking, this method belongs to the non-parametric method. Non-parametric methods [SHA07a], [VIR99], [VAS00] do not rely on any statistical or parametric model as the genesis for speech signal processing. They consider only models for short-time speech segment by assuming that the speech signals are stationary or quasi-stationary in short period. The main attractions of this type of methods are its relative simplicity, high flexibility, broad applicability and ease of implementation. The problem with the short-time model is the ignorance of the long-term characteristics of noise. Noise induced aggravation causes inaccurate information exchange and lowers the quality of speech. Due to the complexity and subtlety of the problem, producing a natural-like speech signal in the presence of unknown noise poses considerable challenges as speech and noise are not always degradable over the entire gamut of hearing range. In this chapter, we extend the proposed critical band excision of discrete wavelet transform to model-based speech enhancement method based on Kalman filter theory.

When there is no knowledge of the statistical properties of the speech or noise signal, model-based methods, like Kalman filtering, are capable of better results. Their theoretical underpinning is the predictable structures and the observable patterns of the speech production process [VAS00]. Standard Kalman filtering algorithm aims at minimizing the error variance in the clean speech estimation, often without *a priori* knowledge of the environment. To achieve optimal clean speech estimation, it must work in conjunction with a good analytical model of speech [QUA02], [VAS00]. The voiced-unvoiced feature sets of the speech signal are derived from time-frequency distribution of its energy [HAY01], [QUA02]. The voiced speech consists of vowels, semi-vowels and nasals, with vowels being the largest phoneme group. Vowel has a relatively long duration of 40–250 ms. The energy of vowel ranges from 100 Hz to 4500 Hz and its power spectrum is concentrated in the range below 1000Hz. Unvoiced speech consists of the stop, fricative and affricate consonants, and contributes considerably to the intelligibility of speech. The duration of unvoiced phonemes is in the range of 10-50 ms. Its power spectra lies in much higher frequency bands than that of vowel and can extend to about 7 or 8 kHz. In reality, there are various types of phonemes mixed and randomly distributed in frequency and have different duration. An autoregressive (AR) process can be used to model the long-term characteristics of speech and noise. It represents unvoiced speech, which is excited by random noise well but not the voiced speech, which is excited by periodic pulse. This problem was addressed in Goh *et al.* [GOH99] in their proposed voiced-unvoiced speech model for Kalman filtering.

The focus of this paper is on the recovery of clean speech from speech corrupted by slowly varying, non-white additive noise, when only the noisy signal is available. A new multirate system structure for speech enhancement is proposed by amalgamating the voiced and unvoiced speech model, Kalman filter and perceptual weighting filter in the discrete wavelet transform domain. Since the speech and noise signals are intrinsically non-stationary, the subtle time variation of their combined spectrum is better preserved in the wavelet domain. Because the real world noise is not flat in spectrum and affects the speech signal non-uniformly over the entire spectrum, the voiced-unvoiced speech is excised into critical band scale by our proposed perceptual wavelet filter bank described in Chapter 3, to capture the localized information of transient signal. To filter the color noise interference, the subband speech and colored noise are modeled as a white Gaussian noise excited AR process for the state-space formulation of Kalman filter. Different from the conventional methods [POP98], [GOH99], [GAB01], [HAY01], [MA06], our proposed state-space formulation has been established based on the multiresolution framework. The model parameters required by the proposed algorithm are estimated from the noisy speech using an iterative expectation-maximization (EM) procedure [HAY01]. Unlike the wavelet filter bank proposed in Chapter 4, which is designated for desktop speech processing with audio data recorded at sampling rates of 16 kHz/16bits per sample, the perceptual wavelet filterbank here is designed to approximate 21 critical bands of human auditory system up to 8 kHz. Comparing with the application of traditional human auditory masking in Kalman filtering to speech enhancement [GAB01], [MA06], the proposed

Kalman filter method incorporates a new unified simultaneous and temporal masking derived by WP tree-based auditory model described in Section 4 of Chapter 3. The quality of the speech processed by wavelet-Kalman filter is further enhanced by feeding it through the proposed perceptually weighting filter, which is designed based on the proposed masking thresholds.

This chapter is organized as follows: Section 5.2 describes the subband-based autoregressive process of the mixed excitation model for voiced-unvoiced speech. In Section 5.3, the colored noise model is introduced, and extended state-space formulation for speech in colored noise is derived in detail. The proposed Kalman filter based on perceptual wavelet filterbank excision is presented in Section 5.4. In Section 5.5, the performance of the proposed speech enhancement method is evaluated and compared against other methods under a variety of adverse conditions using both objective and subjective measures. The chapter is concluded in Section 5.6 where a brief summary of the proposed method is given. A part of the work presented in Sections 5.4 has been presented at the *2006 IEEE International Symposium on Circuits and Systems* [SHA06b].

5.2 Subband-based Autoregressive Model for Voiced and Unvoiced Speech

Kalman filter is a model based recurrent state estimation technique for stationary signal with finite duration. In this section, the voiced-unvoiced speech model used for the state-space formalism of Kalman filter in the perceptual wavelet transform domain is deduced. The novelty of this proposed subband-based autoregressive model for voiced and unvoiced speech lies in its multiresolution formulation to unveil the composition of noise and speech in different frequency bands and the nonstationary property of speech and noise band.

Typically, the noisy speech is modeled by a linear discrete-time system representation.

$$x[n] = s[n] + v[n] \tag{5.1}$$

where $x[n]$ and $s[n]$ are the discrete-time samples of the noisy and clean speech signals, respectively. $v[n]$ is the additive noise, which can be white or colored. The original clean speech and noise are assumed to be uncorrelated.

If the speech signal at time n is assumed to be well predicted by a linear combination of p previous samples, then it can be modeled by an AR process of an excitation signal, $e[n]$ on the state, $s[n]$ [GRA06], [GOH99], [HAY01].

$$s[n] = \sum_{i=1}^{p} a_i s[n-i] + e[n] \tag{5.2}$$

where p is the order of the linear predictive filter and the weights, a_i, are called the AR coefficients. The predictive error, $e[n]$ is small with respect to $s[n]$ and it can be considered as an uncorrelated Gaussian white noise process with zero mean and constant variance, σ_e^2.

From the perspective of human physiological speech production process [QUA02], [VAS00], the voiced and unvoiced sounds are produced with different excitations. For unvoiced sounds, the air is forced from the lungs directly to the vocal tract. This type of excitation can be modeled by random noise. For voiced sounds, the glottis generates quasi-periodic air pulses with the vibration of the voiced cords. Therefore, the above AR model is not suitable for modeling the voiced speech as it is excited by periodic pulse train. A better excitation model to unify silence, voiced and unvoiced speech was proposed in [GOH99]:

$$e[n] = b(n, p_n) e[n - p_n] + d[n] \tag{5.3}$$

The first term of (5.3) expresses the excitation of voiced speech by pulse train in p_n instantaneous pitch period. The term, $b(n, p_n)$ is used to modulate the periodicity of voiced speech. The second term, $d[n]$ describes the excitation of unvoiced speech which is generated by a zero-mean white Gaussian noise with variance, σ_d^2. For unvoiced speech, $b(n, p_n)$ is set to zero and the excitation $e[n]$ is the white Gaussian noise. For voiced speech, σ_d^2 approximates zero and $b(n, p_n)$ is set to unity. Thus, $e[n] \approx e[n{-}p_n]$ represents a periodic signal with pitch period p_n. The mixture of voiced-unvoiced characteristics of speech can be obtained by tuning $b(n, p_n)$ and σ_d^2 in the range between zero and one. Silence is represented by $e[n] = 0$ with both $b(n, p_n)$ and σ_d^2 set to zero.

The single speech model of (5.3), which assumes uniform power distribution of voiced and unvoiced speech throughout the spectrum, is still inadequate. This is because vowels are found in the range from 100 Hz to 4500 Hz and their power is concentrated in the spectrum below 1000Hz. The power spectrum of unvoiced speech lies in a much higher frequency range than that of vowels and can extend to about 7 or 8 kHz. The energy is randomly distributed in frequency. To circumvent this deficiency, we unify the modeling of silence, voiced and unvoiced speech excitations in the time-frequency domain.

To support a mixed voiced-unvoiced speech excitation model with non uniform power spectrum, a frequency to bark transformation is needed to model the frequency dependent sensitivity of human ears. This is achieved with the perceptual wavelet packet transform proposed in Chapter 3. The excitation signal is modeled as a combination of M subband signals in the wavelet transform domain as follows:

$$e[n] = e_1[n] + e_2[n] + \cdots + e_M[n] \tag{5.4}$$

$$e_m[n] = b_m(n, p_n)e_m[n - p_n] + d_m[n] \tag{5.5}$$

where $m = 1, 2, \ldots, M$ is the subband index and $d_m[n]$ is generated by a zero mean unity variance white Gaussian process. The excitation is represented as a mixed proportion of periodic signal and noise in each subband. Voiced and unvoiced speech analyses are separately applied in each subband. For voiced speech, the periodic content of each subband excitation, $e_m[n]$ can be calculated either by using the fundamental pitch period or by the smallest harmonic of the pitch period within the band.

If we divide the pitch period into τ intervals, and set the periodicity, $b_m(n, i) = 0$ if $i \neq p_n$ for all $i = 1, 2, \ldots, \tau$ [GOH99], then (5.5) can be expressed in a autoregressive form as follows:

$$e_m[n] = \sum_{i=1}^{\tau} b_m(n,i)e_m[n-i] + d_m[n] \tag{5.6}$$

where τ is the maximum possible pitch period of human speech.

The state vector of each subband can be augmented to generate the full-band excitation but this yields a complex ($M\times\tau$)-th order model. The computational complexity of the full-band excitation model can be greatly simplified by considering the finite impulse response (FIR) of the wavelet filter bank. To obtain the m-th band excitation signal, we use an N order FIR filter such that

$$e_m[n]=c_{m,1}e[n-1]+c_{m,2}e[n-2]+\cdots+c_{m,N}e[n-N] \tag{5.7}$$

where $c_{m,i}$ are the FIR filter coefficients. Substituting (5.7) into (5.4), the subband excitation can be expressed as:

$$e_m[n]=\sum_{i=1}^{N+\tau}f_m(n,i)e[n-i]+d_m[n] \tag{5.8}$$

where $f_m(n, i)$ is defined as:

$$f_m(n,i)=\begin{cases}c_{m,p_n}b_m(n,p_n) & \text{if } i=p_n\\ 0 & \text{if } i\neq p_n\end{cases} \tag{5.9}$$

Finally, the full-band excitation signal can be represented in terms of the subband excitations by substituting (5.9) into (5.4).

$$e[n]=\sum_{i=1}^{N+p}\sum_{m=1}^{M}f_m(n,i)e[n-i]+\sum_{m=1}^{M}d_m[n] \tag{5.10}$$

Since the proposed full-band excitation signal is derived by perceptual wavelet filterbank, it is helpful to reveal the characteristics of voiced and unvoiced speech in different frequency bands with the multiresolution framework.

5.3 Extended State-Space Formalism for Speech in Colored Noise

After the voiced and unvoiced speech model has been deduced, the proposed state-space formulation is derived for Kalman filter in this section. Unlike the conventional methods [GOH99], [MA06], the proposed state-space equation is modeled according to the critical band structure of multiresolution auditory system.

Kalman filter theory is based on a state-space approach whereby a state equation models the dynamics of the signal generation process through the linear prediction model and an observation equation models the noisy and distorted signal [HAY01]. The Kalman filtering method for speech enhancement operates in two steps. First, the parameters of a state-space excitation model are estimated from the noisy observation and then an optimal Kalman filter is constructed to filter out the undesired components dictated by this parametric model. To obtain a minimum mean square error (MMSE) estimate of clean speech from a noisy observation, the subband AR models for voiced-unvoiced speech and colored noise needs to be recast into the state-space formalism required by the Kalman filter [GIB91], [GRA06], [GOH99], [PAL87].

Let $v[n]$ be white and $\mathbf{s}[n] \triangleq \left[s[n]s[n-1]\cdots s[n-p+1]\right]^T$ be a p-dimensional state vector. For notational simplicity, we ignore the subband distinction in the state-space equations and express the input/internal/output states without the subscript, m in a metonymical fashion.

$$\mathbf{s}[n+1] = \mathbf{F}\mathbf{s}[n] + \mathbf{g}e[n] \tag{5.11}$$

$$x[n] = \mathbf{h}^T\mathbf{s}[n] + v[n] \tag{5.12}$$

where the state transition matrix, $\mathbf{F} = \begin{bmatrix} a_1 & a_2 & \cdots & a_{p-1} & a_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{p\times p}$ and $\mathbf{g} = \mathbf{h} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}_{p\times 1}$

Previous subband AR model for voiced-unvoiced speech is constructed based on the white additive noise. In practice, the background noise, $v[n]$ is colored. For colored noise, we assume that the stochastic process is wide-sense stationary. It can be described by a state-dependent Gaussian autoregressive process of order q [WU98].

$$v[n] = \sum_{i=1}^{q} b_i v[n-i] + \eta[n] \tag{5.13}$$

where q is the order of the linear predictive filter and b_i are the AR coefficients [HAY01]. $\eta[n]$ is a predictive error of a white Gaussian process with zero mean and constant variance σ_η^2, which can be estimated during the non-speech interval.

From (5.13), the 'whiten' version of the colored noise in the subbands can be modeled by:

$$v_m[n] = \sum_{i=1}^{q} b_{m,i} v_m[n-i] + \eta_m[n] \tag{5.14}$$

where m = 1, 2, ..., M is the subband index, and $\eta_m[n]$ is defined as a zero mean unity variance white Gaussian noise.

With $\mathbf{v}[n] \triangleq \left[v[n]v[n-1]\cdots v[n-q+1]\right]^T$, the colored noise state-space equations are given by:

$$\mathbf{v}[n+1] = \mathbf{F}_v \cdot \mathbf{v}[n] + \mathbf{g}_v \cdot \eta[n] \tag{5.15}$$

$$v[n] = \mathbf{h}_v^T \cdot \mathbf{v}[n] \tag{5.16}$$

where $\mathbf{F}_v = \begin{bmatrix} b_1 & b_2 & \cdots & b_{q-1} & b_q \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{q\times q}$ and $\mathbf{g}_v = \mathbf{h}_v = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}_{q\times 1}$.

By making use of the subband autoregressive excitation model from (5.8) and (5.14), the state and excitation vectors of (5.11), (5.12), (5.15) and (5.16) can be concatenated and combined as follows.

$$\begin{bmatrix} \mathbf{s}[n] \\ \mathbf{v}[n] \end{bmatrix} = \begin{bmatrix} \mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_v \end{bmatrix} \begin{bmatrix} \mathbf{s}[n-1] \\ \mathbf{v}[n-1] \end{bmatrix} + \begin{bmatrix} \mathbf{g} & \mathbf{0} \\ \mathbf{0} & \mathbf{g}_v \end{bmatrix} \begin{bmatrix} \mathbf{e}[n] \\ \boldsymbol{\eta}[n] \end{bmatrix} \tag{5.17}$$

$$x[n] = \begin{bmatrix} \mathbf{h}^T & \mathbf{h}_v^T \end{bmatrix} \begin{bmatrix} \mathbf{s}[n] \\ \mathbf{v}[n] \end{bmatrix} \tag{5.18}$$

From which the following equivalent state-space equations of (5.1) can be expressed.

$$\overline{\mathbf{s}}[n] = \overline{\mathbf{F}} \cdot \overline{\mathbf{s}}[n-1] + \overline{\mathbf{g}} \cdot \overline{\mathbf{e}}[n] \tag{5.19}$$

$$\overline{x}[n] = \overline{\mathbf{h}}^T \cdot \overline{\mathbf{s}}[n] \tag{5.20}$$

where $\overline{\mathbf{s}}[n]=\begin{bmatrix}\mathbf{s}[n]\\ \mathbf{v}[n]\end{bmatrix},\overline{\mathbf{F}}=\begin{bmatrix}\mathbf{F} & \mathbf{0}\\ \mathbf{0} & \mathbf{F}_{\mathrm{v}}\end{bmatrix},\overline{\mathbf{g}}=\begin{bmatrix}\overline{\mathbf{g}} & \mathbf{0}\\ \mathbf{0} & \overline{\mathbf{g}}_{\mathrm{v}}\end{bmatrix},\overline{\mathbf{e}}[n]=\begin{bmatrix}e[n]\\ \eta[n]\end{bmatrix},\overline{\mathbf{h}}^T=\begin{bmatrix}\mathbf{h}^T & \mathbf{h}_v^T\end{bmatrix}$.

Once the generalized state-space model for Kalman filtering based on the subband voiced-unvoiced speech model in the presence of colored noise has been established, we can aptly implement our proposed perceptual Kalman filter for speech enhancement in the wavelet domain.

5.4 Proposed Speech Enhancement Method based on Perceptual Wavelet Filter Bank and Kalman Filtering

This section presents the proposed Kalman filtering speech enhancement system that incorporates the noise contaminated voiced-unvoiced speech model and the temporal and simultaneous maskings of human auditory system in wavelet domain. The speech is decomposed into subband speech signals by a multichannel analysis filter bank. Each subband speech signal is converted into a sequence of voice frames where a set of low-order Kalman filters is applied. The autoregressive (AR) parameters for each Kalman filter are estimated frame-by-frame by applying the Expectation Maximization (EM) algorithm on each subband speech signal. The filtered subband speech signals are then combined by a multichannel synthesis filter bank and the outputs of the multichannel synthesis filter bank are summed to produce the enhanced full-band speech signal.

Fig. 5.1 shows the architecture of the proposed speech enhancement system. There are three major blocks in this system. The first block is the analysis and synthesis perceptual wavelet filter bank. The second block is the subband Kalman filter deduced from the state-space equation of the autoregressive voiced-unvoiced speech signal in white and colored noises. The third block is the perceptual weighting filter. The time-frequency masking threshold of human auditory system is calculated in this block. Each of these blocks will be further elaborated as follows.

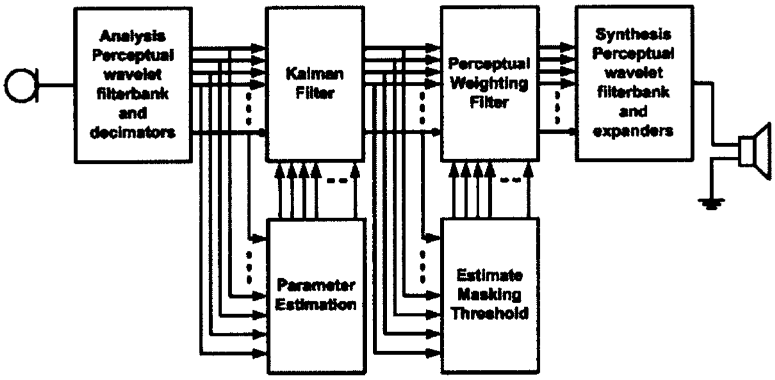

Analysis Perceptual wavelet filterbank and decimators
Kalman Filter
Perceptual Weighting Filter
Synthesis Perceptual wavelet filterbank and expanders
Parameter Estimation
Estimate Masking Threshold

Figure 5.1 Block diagram of the proposed speech enhancement system

5.4.1 Perceptual Wavelet Filterbank

We make use of the perceptual wavelet filter bank proposed in Chapter 3 to achieve a frequency resolution of 125 Hz. The wavelet packet decomposition is accomplished with a six level tree structure of cascaded filter banks. The Daubechies basis of the discrete wavelet packet transform is decided based on the constraints imposed by the temporal resolution of the human ear, which requires that the analysis windows be limited to 5–10 ms towards higher frequencies, but they can spread up to 100 ms at lower frequencies. Once the DWPT is chosen, its time–frequency resolution remains fixed. Each transform coefficient can be expressed as W_{jk}, where k is the coefficient number, j is the transform stage from which W_{jk} is chosen and l is the number of "temporal" coefficients in the critical band. The Kalman filter is divided into 21 subbands by the perceptual wavelet filter bank. A full three-level wavelet packet (WP) decomposition is used to partition the frequency axis into eight bands each of 1 kHz. The lowest band of 0–1 kHz is further decomposed by applying another full three-level WP decomposition. This divides the 0–1 kHz band into eight subbands each of 125 Hz wide. The frequency band of 1–2 kHz is decomposed by applying a two-level WP decomposition, giving four subbands of 250 Hz each. Next, an one-level WP decomposition is performed on the 2–3 kHz band. The frequency band of 2–2.5 kHz is further decomposed into two. This yields six subbands each of 250 Hz wide and one subband of 500 Hz wide. The 3–4 kHz band is decomposed into two bands of 3–3.5 kHz and 3.5–4 kHz, while the frequency bands of 4–5 kHz, 5–6 kHz, 6–7 kHz, and 7–8 kHz are kept intact. Altogether 21 frequency bands are produced, which is the same in number as the critical bands. The analysis filter bank has a maximum frame length of F_L = 128. Thus, the time-frequency analysis is done using this proposed perceptual wavelet filterbank. The noisy input signal is decomposed into frequency bands that closely match the critical band structure. The WPT coefficients at the terminal nodes of the WP tree are fed to the proposed Kalman filter to reduce the interference of adverse noise.

5.4.2 Kalman Filtering in Wavelet Domain

The wavelet transform based Kalman filter has a similar structure as the time-series based Kalman filter [HAY01]. Different from the conventional methods [GOH99], [MA06], the proposed Kalman filter has incorporated the subband voiced and unvoiced speech model in a WP tree-based multiresolution framework. The proposed wavelet

Kalman filtering speech enhancement method requires the knowledge about the parameters of the speech model and additive noise model. Fig. 5.2 shows the block diagram of the proposed wavelet Kalman filter, where only noisy speech is available. To estimate the required parameter of the noise model, a statistical voice activity detector is used to winnow out the noise-only frames. The AR coefficients are first estimated from the noisy speech using an iterative EM procedure before the noise is filtered out by the Kalman filter.

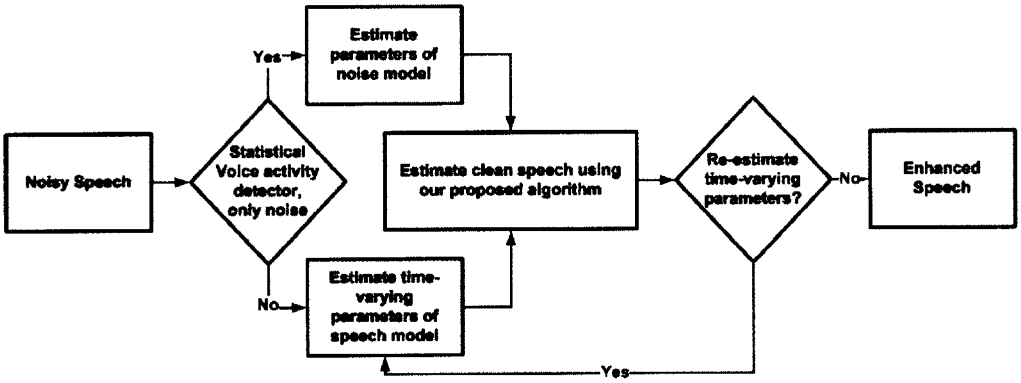

Estimate parameters of noise model
Yes
Noisy Speech
Statistical Voice activity detector, only noise
Estimate clean speech using our proposed algorithm
Re-estimate time-varying parameters?
No
Enhanced Speech
No
Estimate time-varying parameters of speech model
Yes

Figure 5.2 Block diagram of proposed wavelet Kalman filtering speech enhancement method

The speech signal, $s[n]$, is processed on a block transform, and the length of each block is equal to the linear predictive filter order, p. The length of the data to be transformed by wavelet packets is usually a power of 2. In our simulations, the frame length, F_L, was set to 128 and p was set to 8. Thus, there are $B = F_L/p = 16$ blocks within a frame. The state equations and observation equations for block and frame processing after the analysis wavelet filter bank and decimators will be deduced as follows.

Let the column vectors, $\mathbf{S}_{b_i} = \left[s\left[(i-1)\cdot p+1\right] \quad s\left[(i-1)\cdot p+2\right] \quad \dots \quad s\left[(i-1)\cdot p+p\right] \right]^{\mathrm{T}}$ and $\mathbf{E}_{b_i} = \left[e\left[(i-1)\cdot p+1\right] \quad e\left[(i-1)\cdot p+2\right] \quad \dots \quad e\left[(i-1)\cdot p+p\right] \right]^{\mathrm{T}}$, be the clean speech and excitation signals of the i-th block, respectively. From (5.19), the successive blocks of speech can be obtained iteratively by:

$$\mathbf{S}_{\mathbf{b}_{i+1}} = \mathbf{F}_{\mathbf{b}} \cdot \mathbf{S}_{\mathbf{b}_i} + \mathbf{G}_{\mathbf{b}} \cdot \mathbf{E}_{\mathbf{b}_{i+1}} \tag{5.21}$$

where $\mathbf{F}_\mathbf{b}$ and $\mathbf{G}_\mathbf{b}$ are the transition matrices. To obtain $\mathbf{F}_\mathbf{b}$, a pxp upper triangular matrix **A** of AR coefficients is defined.

$$\mathbf{A} = \begin{bmatrix} a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \\ 0 & a_p & a_{p-1} & \cdots & a_2 \\ 0 & 0 & a_p & \cdots & a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_p \end{bmatrix} \tag{5.22}$$

The first row of $\mathbf{F}_\mathbf{b}$ is given by the first row of **A** and the remaining rows of $\mathbf{F}_\mathbf{b}$ are given by:

$$\mathbf{F}_\mathbf{b}(i,j) = \sum_{k=0}^{i-1} a_k \mathbf{A}(i-k,j) \quad \forall j = 1,2,\cdots p \tag{5.23}$$

where $a_0 = 1$.

The matrix $\mathbf{G}_\mathbf{b}$ is a lower triangular matrix. It can be computed in the same manner as $\mathbf{F}_\mathbf{b}$ except that **A** is replaced by a $p \times p$ identity matrix.

Let $\mathbf{S}_{\mathbf{f}_i} = \left[\mathbf{S}_{\mathbf{b}_i}^{\mathrm{T}} \ \mathbf{S}_{\mathbf{b}_{i+1}}^{\mathrm{T}} \ \ldots \ \mathbf{S}_{\mathbf{b}_{i+B-1}}^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\mathbf{E}_{\mathbf{f}_i} = \left[\mathbf{E}_{\mathbf{b}_i}^{\mathrm{T}} \ \mathbf{E}_{\mathbf{b}_{i+1}}^{\mathrm{T}} \ \ldots \ \mathbf{E}_{\mathbf{b}_{i+B-1}}^{\mathrm{T}}\right]^{\mathrm{T}}$ denote the clean speech and excitation signal of the i-th frame, respectively. There is an overlap of $B-1 = 15$ blocks between two adjacent frames and the frame state equation has a similar relationship as (5.21)

$$\mathbf{S}_{\mathbf{f}_{i+1}} = \mathbf{\Lambda} \cdot \mathbf{S}_{\mathbf{f}_i} + \mathbf{\Gamma} \cdot \mathbf{E}_{\mathbf{f}_{i+1}} \tag{5.24}$$

$\mathbf{S}_{\mathbf{f}_{i+1}} = \mathbf{\Lambda} \cdot \mathbf{S}_{\mathbf{f}_i} + \mathbf{\Gamma} \cdot \mathbf{E}_{\mathbf{f}_{i+1}}$

where $\mathbf{\Lambda} = \begin{pmatrix} \mathbf{F}_\mathbf{b} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{F}_\mathbf{b} \end{pmatrix}$ and $\mathbf{\Gamma} = \begin{pmatrix} \mathbf{G}_\mathbf{b} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{G}_\mathbf{b} \end{pmatrix}$.

Based on the above state equation, the observation equation for block and frame processing is given as follows.

$$\mathbf{X}_{\mathbf{f}_i} = \mathbf{S}_{\mathbf{f}_i} + \mathbf{V}_{\mathbf{f}_i} \tag{5.25}$$

where $\mathbf{V}_{\mathbf{f}_i} = \left[\mathbf{V}_{\mathbf{b}_i}^{\mathrm{T}} \quad \mathbf{V}_{\mathbf{b}_{i+1}}^{\mathrm{T}} \quad \cdots \quad \mathbf{V}_{\mathbf{b}_{i+B-1}}^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\mathbf{X}_{\mathbf{f}_i} = \left[\mathbf{X}_{\mathbf{b}_i}^{\mathrm{T}} \quad \mathbf{X}_{\mathbf{b}_{i+1}}^{\mathrm{T}} \quad \cdots \quad \mathbf{X}_{\mathbf{b}_{i+B-1}}^{\mathrm{T}}\right]^{\mathrm{T}}$ are respectively the noise and noisy speech frames. The corresponding block vectors are:

$$\mathbf{V}_{\mathbf{b}_i} = \left[v\left[(i-1)\cdot p+1\right] \quad v\left[(i-1)\cdot p+2\right] \quad \cdots \quad v\left[(i-1)\cdot p+p\right]\right]^{\mathrm{T}}$$
$$\mathbf{X}_{\mathbf{b}_i} = \left[x\left[(i-1)\cdot p+1\right] \quad x\left[(i-1)\cdot p+2\right] \quad \cdots \quad x\left[(i-1)\cdot p+p\right]\right]^{\mathrm{T}}.$$

Let $W(\mathbf{S})$ denotes the perceptual wavelet transform of the time domain function, $s(t)$. By applying the perceptual wavelet transform, $\mathbf{W}$, to the above time domain state-space model, we have

$$W\left(\mathbf{S}_{\mathbf{f}_{i+1}}\right) = \mathbf{W\Lambda S}_{\mathbf{f}_i} + \mathbf{W\Gamma E}_{\mathbf{f}_{i+1}} = \left(\mathbf{W\Lambda W}^{\mathrm{T}}\right)\mathbf{WS}_{\mathbf{f}_i} + \left(\mathbf{W\Gamma W}^{\mathrm{T}}\right)\mathbf{WE}_{\mathbf{f}_{i+1}} \tag{5.26}$$

Let $\mathbf{W\Lambda W}^{\mathrm{T}} = \overline{\mathbf{\Lambda}}$ and $\mathbf{W\Gamma W}^{\mathrm{T}} = \overline{\mathbf{\Gamma}}$,

$$W\left(\mathbf{S}_{\mathbf{f}_{i+1}}\right) = \overline{\mathbf{\Lambda}}W\left(\mathbf{S}_{\mathbf{f}_i}\right) + \overline{\mathbf{\Gamma}}W\left(\mathbf{E}_{\mathbf{f}_{i+1}}\right) \tag{5.27}$$

The corresponding observation equation is given by:

$$\mathbf{X}_{\mathrm{f}_i} = \mathbf{W}^{\mathrm{T}} W\left(\mathbf{S}_{\mathrm{f}_i}\right) + \mathbf{V}_{\mathrm{f}_i} \tag{5.28}$$

Table 5.1 Computations of the Proposed Wavelet-Kalman Filter

Input Vector
Observations: $\mathbf{X}_{\mathrm{f}_i} = \left[\mathbf{X}_{\mathbf{b}_i}^{\mathrm{T}} \ \ \mathbf{X}_{\mathbf{b}_{i+1}}^{\mathrm{T}} \cdots \ \mathbf{X}_{\mathbf{b}_{i+B-1}}^{\mathrm{T}}\right]^{\mathrm{T}}$
Known parameters
Transition matrix in wavelet domain = $\bar{\mathbf{\Lambda}}$.
The covariance matrix associated with $\mathbf{E}_{\mathrm{f}_i} = \mathbf{\Phi}$.
The covariance matrix associated with $\mathbf{V}_{\mathrm{f}_i} = \mathbf{\Psi}$.
Computation for each time instant, i
$\mathbf{G}(i) = \mathbf{K}(i,i-1)\mathbf{W}\left(\mathbf{W}^{\mathrm{T}}\mathbf{K}(i,i-1)\mathbf{W} + \mathbf{\Psi}\right)^{\mathrm{T}}$
$\mathbf{a}(i) = \mathbf{X}_{\mathrm{f}_i} - \mathbf{W}^{T} \cdot W\left(\mathbf{S}_{\mathrm{f}_{i-1}}\right)$
$W\left(\hat{\mathbf{S}}_{\mathrm{f}_i}\right) = \bar{\mathbf{\Lambda}} \cdot W\left(\hat{\mathbf{S}}_{\mathrm{f}_{i-1}}\right) + \mathbf{G}(i)\mathbf{a}(i)$
$\mathbf{K}(i) = \left(\mathbf{I} - \mathbf{G}(i)\mathbf{W}^{\mathrm{T}}\right)\mathbf{K}(i,i-1)$
$\mathbf{K}(i+1,i) = \bar{\mathbf{\Lambda}}\mathbf{K}(i)\bar{\mathbf{\Lambda}}^{\mathrm{T}} + \bar{\mathbf{\Gamma}}\mathbf{\Phi}\bar{\mathbf{\Gamma}}^{\mathrm{T}}$

Based on the state-space model of (5.27) and (5.28), the computation of the Kalman filter algorithm in wavelet domain is presented in Table 5.1. The variables and parameters of the algorithm are defined as follows: **G**(*i*) is the Kalman gain, **a**(*i*) is the innovation vector, $\hat{\mathbf{S}}_{\mathrm{f}_i}$ is the filtered estimate of the state vector, **K**(*i*) and **K**(*i*, *i*–1) are the correlation matrix of the error in the filtered and predicted estimate of the states, respectively, given the observation of one frame.

5.4.3 Perceptual Weighting Filter

Due to the distortion introduced by the estimation error of the proposed Kalman filter, we propose to apply a perceptual weighting filter to minimize the estimation error variance under the constraint that the energy of the estimation error is smaller than a unified temporal and simultaneous masking threshold. Unlike the traditional methods

[MA06], the proposed weighting filter is based on a WP-tree based auditory system. It not only makes distortion inaudible but also improves the intelligibility of the enhanced speech processed by the wavelet Kalman filter. The steps involved in this calculation are motivated by [ZWI90] to incorporate adaptive thresholding of residual noise maskee in each critical band. The input signal is first mapped into critical band scales by the perceptual wavelet filter bank for time-frequency analysis. The spectral energy, $\Theta_{j,k}(Z_W)$ is computed, where j is the transform stage, k is the coefficient number and $Z_W = 1, 2, \ldots, 21$. In addition, the spectral flatness measure (SFM) [JOH88] is used to estimate the tonality coefficient, τ $(0 \leq \tau \leq 1)$ of the signal in each critical band. $\tau = 0$ for a purely white noise random signal and $\tau = 1$ for a purely tonal signal. An adaptive thresholding is performed on the estimated spectral components, $W\left(\widehat{\mathbf{S}}_{\mathbf{f}_i}\right)\left(Z_w\right)$ of the i-th data frame for all values of Z_W.

$$W\left(\widehat{\mathbf{S}}_{\mathbf{f}_i}\right)'\left(Z_w\right)=\begin{cases} W\left(\widehat{\mathbf{S}}_{\mathbf{f}_i}\right)\left(Z_w\right)\times\tau^{\frac{\Theta_{jk}(Z_w)}{T_{jk}(Z_w)}} & \text{if}\,\Theta_{jk}\left(Z_w\right)<T_{jk}\left(Z_w\right) \\ W\left(\widehat{\mathbf{S}}_{\mathbf{f}_i}\right)\left(Z_w\right)\times\dfrac{\Theta_{jk}\left(Z_w\right)}{T_{jk}\left(Z_w\right)} & \text{otherwise} \end{cases} \tag{5.29}$$

Finally, the enhanced speech is recovered from the inverse perceptual wavelet transform.

In (5.29), both the characteristics of speech and masking properties are taken into account in the thresholding procedure. When the spectrum energy, $\Theta_{j,k}(Z_W)$, is smaller than the masking threshold, $T_{j,k}(Z_W)$, it implies that the noise component at that frequency band can be masked. Therefore, the smaller the value of $\Theta_{j,k}(Z_W)$, the more the Kalman filtered signal energy will be retained. On the other hand, if $\Theta_{j,k}(Z_W)$ is larger than $T_{j,k}(Z_W)$, it means that the noise component may not be maskable. In order to reduce the noise component for such cases, a factor is computed from the threshold to attenuate the amplitude of the noisy speech.

The niche of the perceptual weighting filter is the consideration of temporal and simultaneous maskings for the adaptive noise masking threshold, $T_{j,k}(Z_W)$. This adaptive noise masking threshold can be estimated by referring to the derivation and procedure described in Chapter 3.

5.5 Experimental Results and Discussions

In this section, the performances of the proposed perceptual wavelet-Kalman filter (PWKF) speech enhancement method are evaluated and compared against the following competitive methods: (1) SS: spectral subtraction method [BOL79], (2) PSS: Perceptual spectral subtraction method [VIR99], (3) KF: Kalman filtering speech enhancement for colored noise [POP98], (4) KFM: Kalman filtering speech enhancement for colored noise with masking [GAB01], (5) CKF: Constrained Kalman filtering speech enhancement [MA06], and (6) WKFP: Wavelet Kalman filter with post-filter based on masking properties of human auditory system [MA06]. The criteria used for the evaluation includes two objective quality measures, namely segmental SNR and perceptual evaluation of speech quality (PESQ) score. PESQ test is evaluated using ITU-T P.862 Version 1.1 software [P.862]. The speech signals are taken from the TIMIT database [TIM90]. The signals are sampled at 16 kHz. Different background noises with varying time-frequency distributions were taken from the Noisex-92 database [NOI92]. The tested noisy environments include white Gaussian noise, pink noise, Volvo engine noise, F16 cockpit noise, factory noise, high frequency channel noise and speech-like noise. These noises were added to the clean speech signal with different SNR's.

The amount of noise reduction is generally measured in terms of the SNR improvement, given by the difference between the input and output segmental SNRs. The pre-SNR and post-SNR are defined as:

$$SNR_{pre} = \frac{1}{K}\sum_{m=0}^{K-1}\left(20\cdot\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}s(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}v(n+Nm)}\right)\right) \tag{5.30}$$

$$SNR_{post} = \frac{1}{K}\sum_{m=0}^{K-1}\left(20\cdot\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}s(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}\left(s(n+Nm)-\hat{s}(n+Nm)\right)}\right)\right) \tag{5.31}$$

where K represents the number of frames in the signal and N indicates the number of samples per frame.

The segmental SNR improvement is defined as $SNR_{post} - SNR_{pre}$:

$$\Delta SNR = \frac{1}{K}\sum_{m=0}^{K-1}\left(20\cdot\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}v(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}\left(s(n+Nm)-\hat{s}(n+Nm)\right)}\right)\right) \tag{5.32}$$

Table 5.2 show the evaluation results of SNR improvement tested in white Gaussian noise, colored noise, car engine noise, and speech-like noise environments. The results show that the Kalman filter methods improve the SNR and reduce the distortion more than the subtractive-type of speech enhancement methods. This is because Kalman filter is designed based on the model of speech and noise production processes. Comparing with other competitive Kalman filtering speech enhancement methods, it is found that the SNR improvement of our proposed PWKF method rises with increasing filter-band number. The proposed PWKF outperforms other methods with an overall percentage increase in SNR improvement of 9.93%. The ascendancy is due to the better optimal Kalman estimation of clean speech at subband level with good approximation to the critical band of human auditory system provided by the perceptual wavelet filter bank.

Table 5.2 Comparison of Segmental SNR Improvement [dB]

Noise type, SNR$_{in}$	**SS**	**PSS**	**KF**	**KFM**	**CKF**	**WKFP**	**PWKF**
White, 0 dB	6.94	9.21	7.14	7.68	9.10	8.79	**10.41**
White, 10 dB	5.03	5.12	1.87	3.94	5.09	4.65	**5.85**
Colored, 0 dB	5.26	7.72	8.11	6.56	7.61	6.93	**8.90**
Colored, 10 dB	3.22	3.65	2.30	3.47	3.59	3.25	**4.24**
Car engine, 0 dB	7.17	9.56	8.03	8.49	9.44	8.83	**9.97**
Car engine, 10 dB	5.28	5.57	3.05	5.31	5.40	5.08	**6.05**
Speechlike, 0 dB	4.82	6.38	5.31	5.52	6.13	5.80	**6.91**
Speechlike, 10 dB	3.09	3.57	2.09	3.19	3.39	2.92	**4.05**

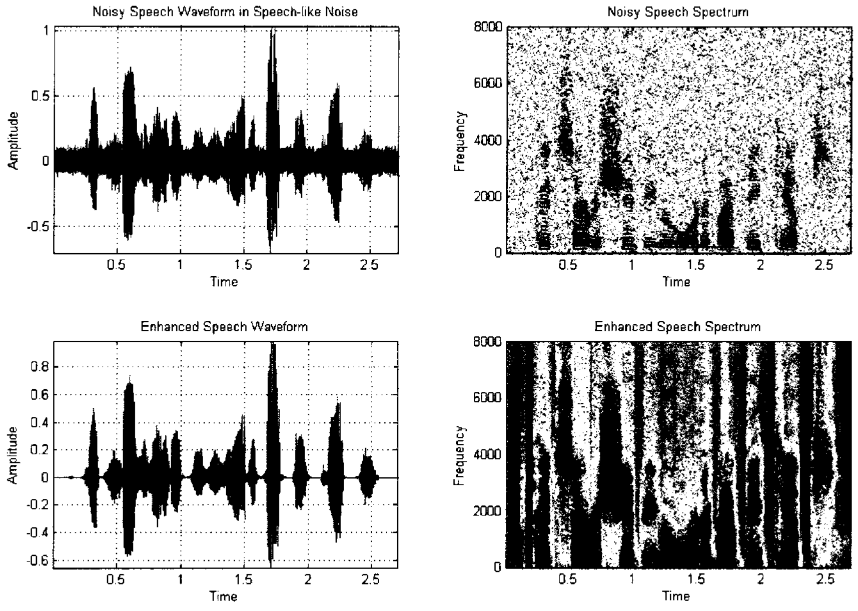


Figure 5.3 Spectrograms of noisy speech and enhanced speech by proposed method in white Gaussian noise at SNR$_{in}$ = 0 dB

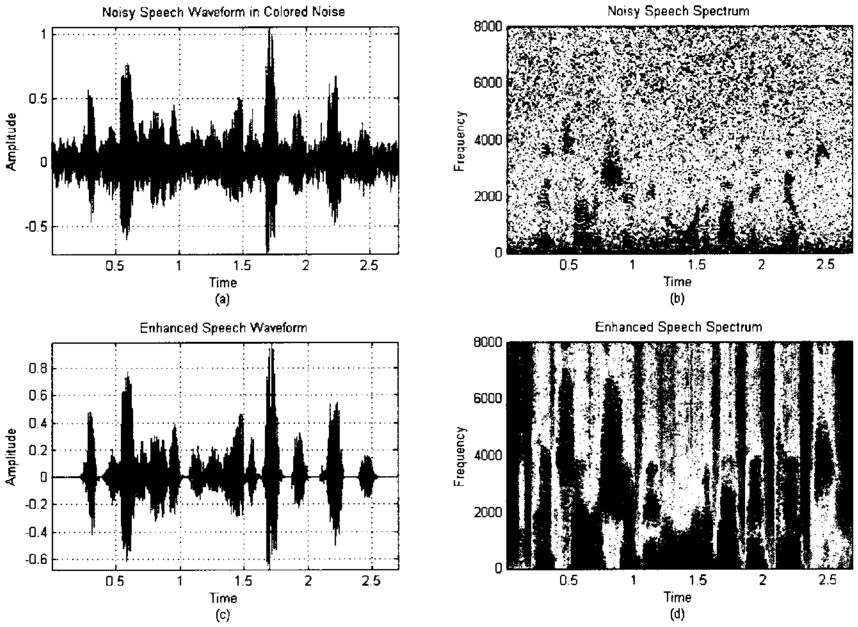


Figure 5.4 Spectrograms of noisy speech and enhanced speech by proposed method in colored noise at SNR$_{in}$ = 0 dB

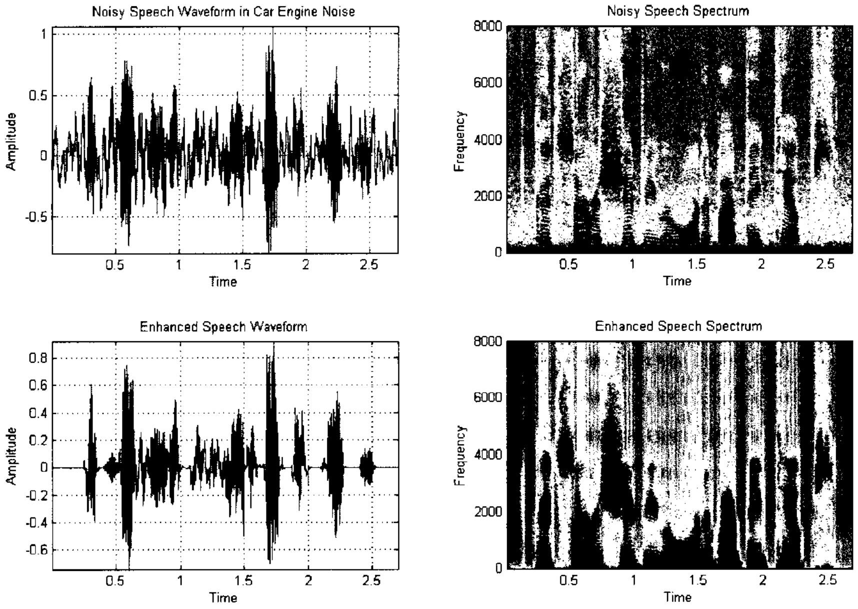


Figure 5.5 Spectrograms of noisy speech and enhanced speech by proposed method in Volvo engine noise at SNR$_{in}$ = 0 dB

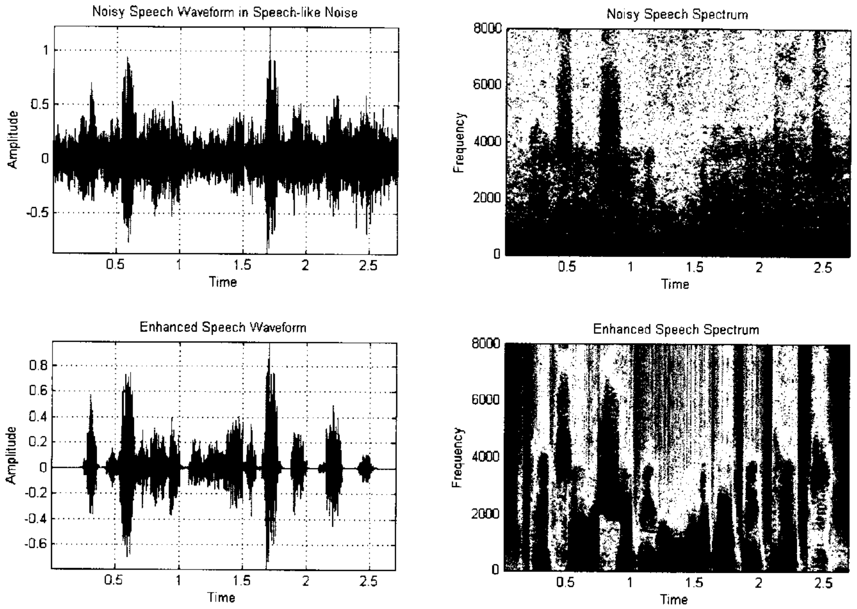


Figure 5.6 Spectrograms of noisy speech and enhanced speech by proposed method in speech-like noise at $\mathrm{SNR}_{\mathrm{in}} = 0$ dB

The time variation of energy in the speech signal can be represented by a spectrogram. Figs. 5.3 – 5.6 show the waveforms and spectrograms of noisy speeches and enhanced speeches in various noise environments of input SNR = 0 dB. In these spectrograms, it can be seen that spectrally broad plosive are less distorted. An example of the spectrograms for the noisy speech and the speech enhanced by the proposed method are presented in Fig. 5.3 for white Gaussian noise at SNR 0 dB. The characteristic of white noise is that it exhibits equal energy in all bandwidth. In the spectrogram of the speech enhanced by the proposed method, very good speech quality is evident while the noise has been clearly diminished. As pitch harmonics are less smooth, low energy pitch harmonics are also better preserved. Fig. 5.4 shows the experiment in color noise. This pink noise exhibits equal energy per 1/3 octave. From the figure, it is clear that the enhanced speech by the proposed method is more pleasant and the distortion remains acceptably low. If we observe the enhanced speech obtained with other noise types at

SNR 0 dB, we can see that the speech quality has also been greatly improved, as long as the noise remains stationary. The spectrograms of Figs. 5.5 and 5.6 illustrate that the proposed method is able to considerably enhance the energy of low frequency speech and at the same time, improve the speech intelligibility at high frequency, with the background noise reduced greatly.

Several listening tests are performed using the Perceptual Evaluation of Speech Quality (PESQ) [PESQ01] score as performance index. PESQ is a speech quality assessment model. The range of PESQ score is from –0.5 to 4.5 with higher score for better speech quality. The PESQ scores of the spectral subtraction methods are typically lower than those of the Kalman filter methods due to the presence of musical noise. Some improvements have been made to suppress the musical noise in spectral subtraction methods [VIR99], but the results for noisy speech with low SNR (<5 dB) are still not satisfactory. Fig. 5.7 shows the PESQ gains obtained by the different speech enhancement methods in various speech-like noise environments. The PESQ gain is the average improvement in PESQ scores of enhanced speech over the PESQ scores of noisy speech for six speech sentences provided by the TIMIT database. The proposed PWKF achieves the highest PESQ gain. Its PESQ gain is 0.436. This is an improvement of 5.83% over the PESQ gain of CKF. The PESQ gains of other methods, in descending order of performance, are CKF (0.412), WKFP (0.388), PSS (0.158), KFM (0.116), and KF (0.082).

An informal Mean Opinion Score (MOS) scale test [DEL93] was also performed with ten listeners. Each listener gives a score between one and five to each test signal. The score represents his global perception of the residual noise, background noise and speech distortion. For each tester, the following procedure has been applied. First, the clean and noisy speeches are played and repeated twice. Next, each test signal, which is repeated twice for each score, is played three times in a random order. This leads to 30 scores for each test signal. Table 5.3 shows the informal MOS obtained by averaging the results obtained from the listening test. From Table 5.3, it is observed that for all the noise conditions, the performance trend of the six speech enhancement methods remains well correlated in both MOS and PESQ scores. The best scores, highlighted in bold print, are consistently obtained by our proposed method for different types of noisy speech. The MOS and PESQ of the proposed method are improved by 3.23% and 3.38% over CKF, respectively, under all types of noise and SNR conditions. The enhancements are more apparent at low input SNR. The improvement in speech intelligibility is attributed to the good estimation of clean speech by the perceptual wavelet-Kalman filtering technique and the limitation of perceivable distortion by the adaptive noise thresholding of the perceptual weighting filter.

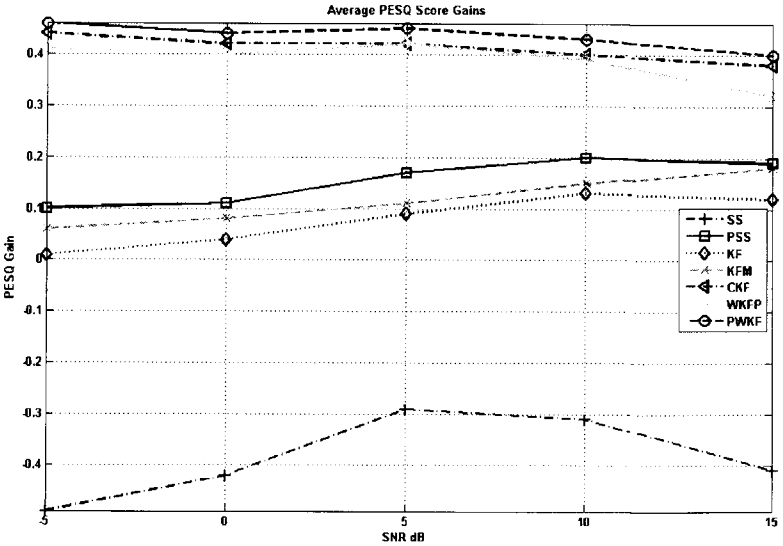


Figure 5.7 PESQ gains obtained by different algorithms using input noisy speech with different SNR in speech-like noise

Table 5.3 Comparison of Informal MOS and PESQ Scores

Noise type, *SNR_i*	**Informal MOS score**						
	Noisy	**PSS**	**KF**	**KFM**	**CKF**	**WKFP**	**PWKF**
White, 0 dB	1.35	1.55	1.40	1.42	1.84	1.82	**1.91**
White, 10 dB	2.53	2.64	2.60	2.61	3.04	3.02	**3.12**
Colored, 0 dB	1.45	1.70	1.50	1.52	2.03	2.01	**2.12**
Colored, 10 dB	2.62	2.77	2.72	2.72	2.90	2.89	**2.98**
Car engine, 0 dB	1.59	1.71	1.63	1.64	1.90	1.89	**2.00**
Car engine, 10 dB	2.22	2.36	2.30	2.31	2.67	2.65	**2.73**
Speechlike, 0 dB	1.51	1.66	1.56	1.57	1.96	1.94	**2.03**
Speechlike, 10 dB	2.43	2.58	2.51	2.52	2.74	2.72	**2.79**
Noise type, *SNR_i*	**PESQ score**						
	Noisy	**PSS**	**KF**	**KFM**	**CKF**	**WKFP**	**PWKF**
White, 0 dB	1.24	1.48	1.28	1.29	1.95	1.92	**2.01**
White, 10 dB	1.94	2.17	2.07	2.08	2.51	2.50	**2.59**
Colored, 0 dB	1.51	1.76	1.56	1.57	2.00	1.98	**2.09**
Colored, 10 dB	2.16	2.37	2.29	2.30	2.54	2.52	**2.63**
Car engine, 0 dB	1.42	1.65	1.46	1.47	1.88	1.85	**1.94**
Car engine, 10 dB	2.11	2.25	2.15	2.18	2.42	2.40	**2.50**
Speechlike, 0 dB	1.22	1.39	1.26	1.29	1.77	1.75	**1.83**
Speechlike, 10 dB	1.88	2.05	1.95	1.99	2.41	2.39	**2.49**

5.6 Chapter Summary

This chapter presents an effective speech enhancement technique that bridges the discrete wavelet packet transform, which is a flexible multi-resolution time-frequency excision of speech features and Kalman filter, which is a powerful model based estimator of signal in noise. A perceptual wavelet filter bank is realized by a six-level dyadic subband tree decomposition to approximate the critical band of human auditory system. The compaction of the energy of the critical transient components into different spectral resolutions in wavelet representation allows the background noise and distortion to be efficiently excised by the Kalman filter. The state-space equations of the proposed wavelet Kalman filter is formulated based on a subband voiced-unvoiced speech model. The residue noise of the clean speech estimated by the wavelet Kalman filter is further attenuated by a perceptually weighted filter. Comparing with other speech enhancement methods, the effectiveness of the proposed wavelet Kalman filter to extract a clear and intelligible speech from adverse noise environments is largely due to the incorporation of statistical and perceptual models into the wavelet transform excision process for optimal estimation of clean speech.

CHAPTER 6.

SPEECH PROCESSING TOOLBOX FOR PERFORMANCE ANALYSIS, ENHANCEMENT AND EVALUATION OF HYBRID SYSTEMS

6.1 Introduction

Several speech processing toolboxes have been developed to facilitate speech processing research [BRO], [LOI99], [SLA98]. However, these toolboxes are mostly targeted for specific speech analysis and tailored to some dedicated applications. If the researchers want to analyze and modify their recorded speech and make measurements on the modified speech, they have to extend the current speech processing toolboxes or integrate it with compatible functions of others. This is not very convenient. To solve this problem, we have consolidated a good number of state-of-the-art algorithms for speech signal processing research, including those that we have developed, investigated and compared in the earlier chapters. They are integrated into a flexible and friendly graphical user interface to build a versatile software platform for some common and widely studied applications based on the principles of discrete-time speech signal processing. Fig. 6.1 shows an overview of the discrete-time speech signal processing.

With the advent of digital technology, the discrete-time speech signal can be easily obtained by passing the continuous-time waveform through an analog-to-digital (A/D) converter. Based on the discrete-time model of speech production, the speech analysis and synthesis systems are designed. In analysis, the speech model parameters are measured with the temporal and spectral resolution. In synthesis, the waveform is reconstructed based on these parameter estimates. The analysis/synthesis methods are the backbone for applications such as speech modification, enhancement, coding, and recognition. (1) Modification - the goal of speech modification is to alter the speech signal to have some desired property. Modifications of interest include time-scale, pitch, and spectral changes. (2) Coding - the goal of speech coding is to reduce the information rate, measured in bits per second, while maintaining the quality of the original speech waveform. (3) Enhancement - the goal of speech enhancement is to improve the quality of degraded speech. (4) Recognition - the goal of speech recognition is to use the speech features provided by feature extractor to assign the object to a category.

Written for digital speech processing researchers, this speech toolbox introduces speech analysis and synthesis through user-computer interaction. It provides a means to study the features and properties of speech as a signal without having to record data or write software to analyze the data. The toolbox is completed with data collection and measurement procedures, the theory of speech data processing, and the application of digital signal processing procedures, such as speech modification, enhancement, feature extraction and classification.

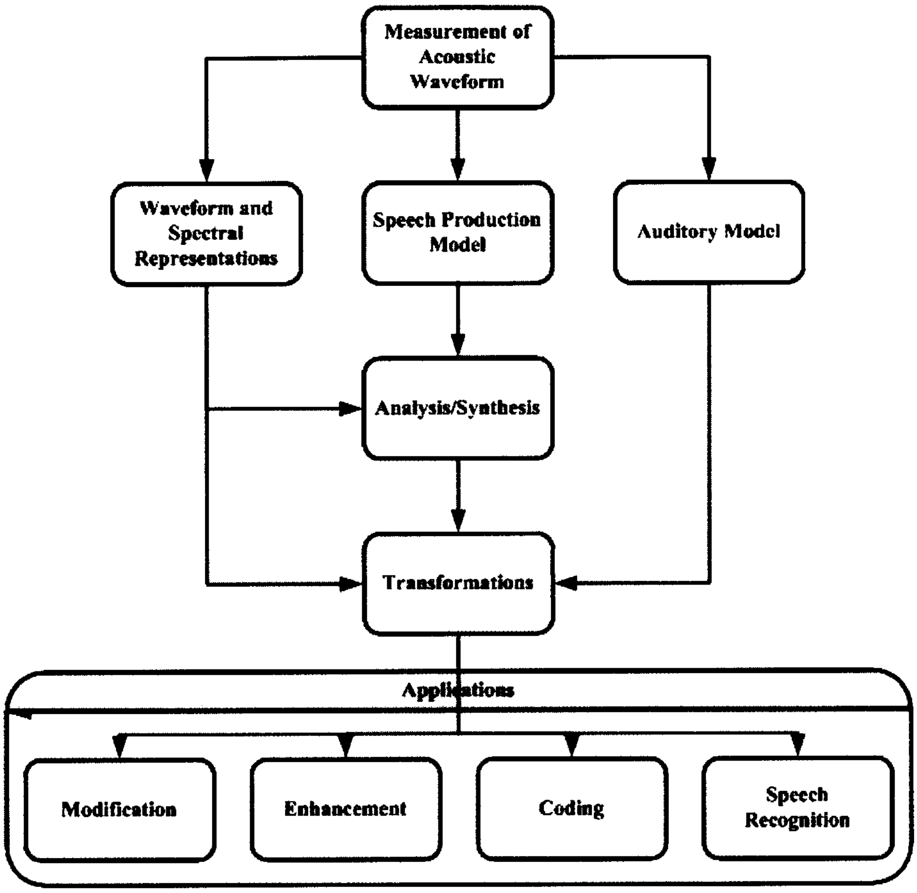


Figure 6.1 Discrete-time speech signal processing overview

Besides easing the analyses and comparison of experimentation results, this toolbox also facilitates speciation of a framework of wavelet based techniques to harness automatic speech recognition (ASR) performance in the presence of background noise. The proposed robust speech recognition system is realized by cascading speech enhancement preprocessing, feature extraction, and hybrid speech recognizer in time-frequency space. This hybrid and hierarchical design paradigm improves the recognition performance by combining the advantages of different integral methods within a single system. Speech enhancement preprocessing is applied to minimize the mismatch between the training and testing conditions of the classifier. The effect of speech enhancement on classifier performances are then evaluated using the enhanced speech estimated by the generalized perceptual time-frequency subtraction (GPTFS) and the perceptual wavelet Kalman filtering (PWKF) algorithms. The non-phonetic information is discarded while the more critical speech features are extracted and represented by the wavelet coefficients via feature extraction. Two methods, wavelet packet transform and local discriminant bases (WPT/LDB) and perceptual wavelet filterbank (PWF), are evaluated. The former is designed to capture the most discriminative information in the time-frequency plane from a dictionary of orthonormal bases, while the later uses a fixed base to imitate the human perceptual modus of speech. The denoised wavelet features are fed to the hybrid classifier founded on Hidden Markov Model (HMM). The intrinsic limitation of HMM is overcome by augmenting it with either Multilayer Perceptron (MLP) or Support Vector Machine (SVM). All the above mentioned functions are available in the speech processing toolbox. The connected digit recognition experiments conducted on the proposed framework show encouraging results of various ASR configurations with the denoised wavelet features.

The rest of this chapter is organized as follows. Section 6.2 presents the graphical user interface (GUI) of the speech processing toolbox. A brief description of the major functions is given and illustrated by some graphical results on the integrated environment. Section 6.3 depicts an evaluation platform of robust speech recognition for denoised features. Each processing stage, except those that have been presented earlier, is detailed. Different configurations of front-end speech enhancement processing, feature extractor and hybrid recognizer modules for automatic speech recognition (ASR) are analyzed and evaluated in Section 6.4. The toolbox is available on a CD bound in the thesis. A part of the work in Section 6.3.3 has been presented at the 2005 and 2006 International Symposiums on Circuits and Systems [SHA05b], [SHA06a]. A large portion of the work in Section 6.3 presented in this chapter has been submitted for review as a regular paper in the *IEEE Transactions on System, Man and Cybernetics, Part A* [SHA07b].

6.2 Speech Processing Toolbox based on Matlab

This section describes the GUI of a self-contained speech processing application toolbox, which is a collection of algorithms and tools that enable several popular speech analysis and processing functions. An integrated environment is built to demonstrate the effects of speech models and speech analysis procedures on the quality of synthesized speech. The software environment was developed with MATLAB program and the Signal Processing Toolbox deployed under the MATLAB Runtime Server. An extensive speech database and a set of MATLAB M-files are also included. This toolbox is particularly useful to researchers that are interested to perform speech analysis, understand the effect of speech enhancement, and configure, compare and test the performances of some speech processing system. This toolbox is also useful to speech and auditory engineers who want to see how the sounds are modeled and perceived by the human auditory system. The flowchart of this software toolbox is shown in Fig. 6.2. There are 4 primary modules and 5 auxiliary modules making up of the speech processing toolbox. The primary modules consist of software interface, auditory model, speech model, and speech enhancement. The input signal is sampled and preprocessed by software interface module. The modeling of critical band and absolute threshold of hearing is implemented using auditory model. The speech model module provides a list of functions for speech analysis, such as fundamental frequency estimate, formant track, power spectral density estimation, and AR analysis. The speech enhancement module encompasses a list of subtraction methods, wavelet based methods, and adaptive filtering methods. The auxiliary modules compose of frequency scale convertor, AR parameter estimation, transformations, objective measurement, and utility functions. These modules provide supplementary functions for the speech processing toolbox.

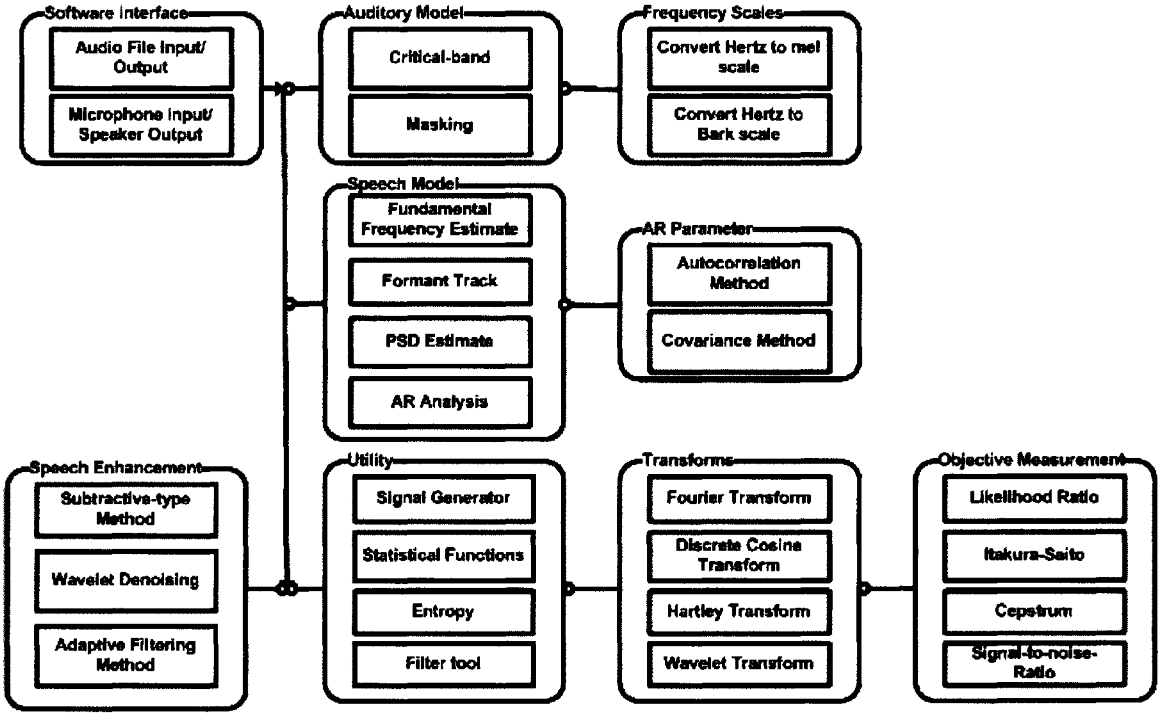
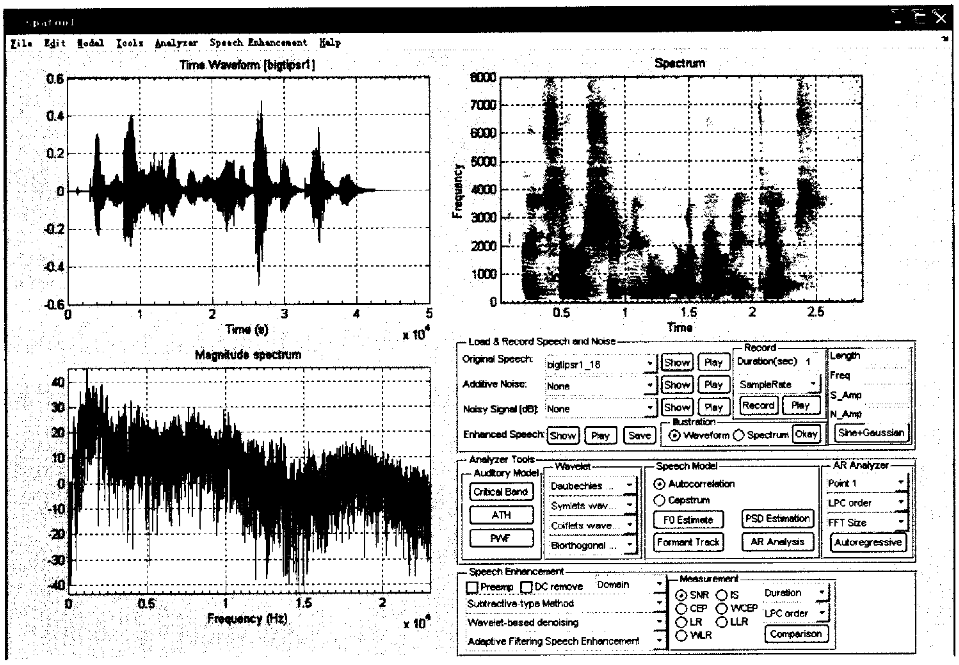


Figure 6.2 Categories of features and modules provided by the speech processing toolbox

Four major types of functions are implemented in this toolbox:

(1) Software interface

The toolbox provides two input/output interfaces for the acquisition of speech signal. The speech signal can be acquired and output by reading from and writing to an audio file. The audio signal can also be acquired from a microphone and played back by a speaker. There are many ways to describe and represent sounds. Fig. 6.3 shows a taxonomy based on the signal dimensionality. A simple waveform is a one-dimensional representation of sound. The two-dimensional representation describes the acoustic signal as a time-frequency image. This is the typical approach for sound and speech analysis.

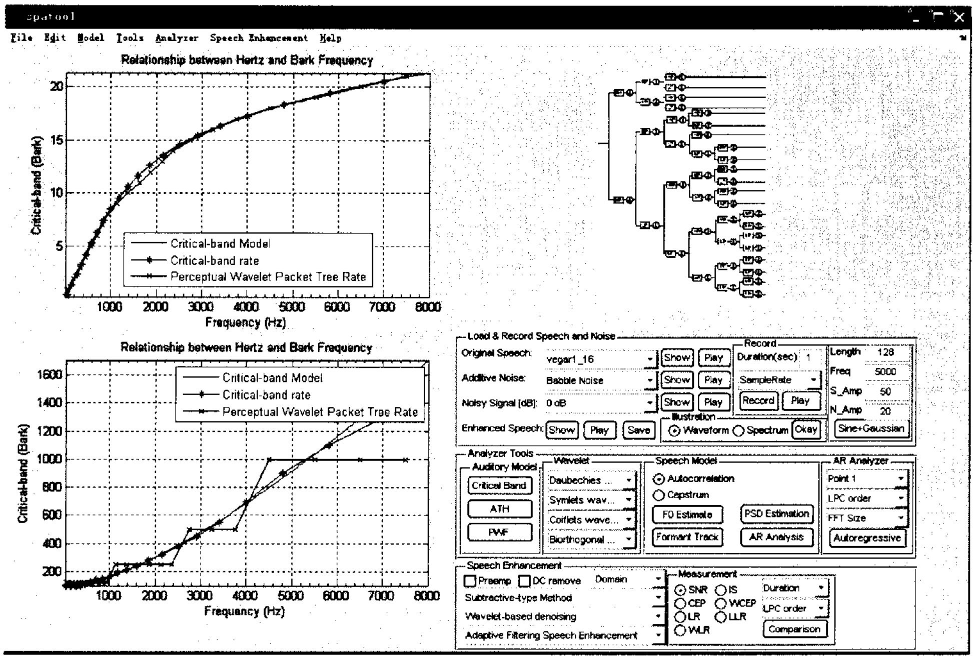


Original Speech: bigtipsr1_16
Additive Noise: None
Noisy Signal [dB]: None
Duration(sec) 1
Illustration ● Waveform ○ Spectrum
● Autocorrelation ○ Cepstrum
● SNR ○ IS ○ CEP ○ WCEP ○ LR ○ LLR ○ WLR

Figure 6.3 GUI for different analytical representations of acoustic signal

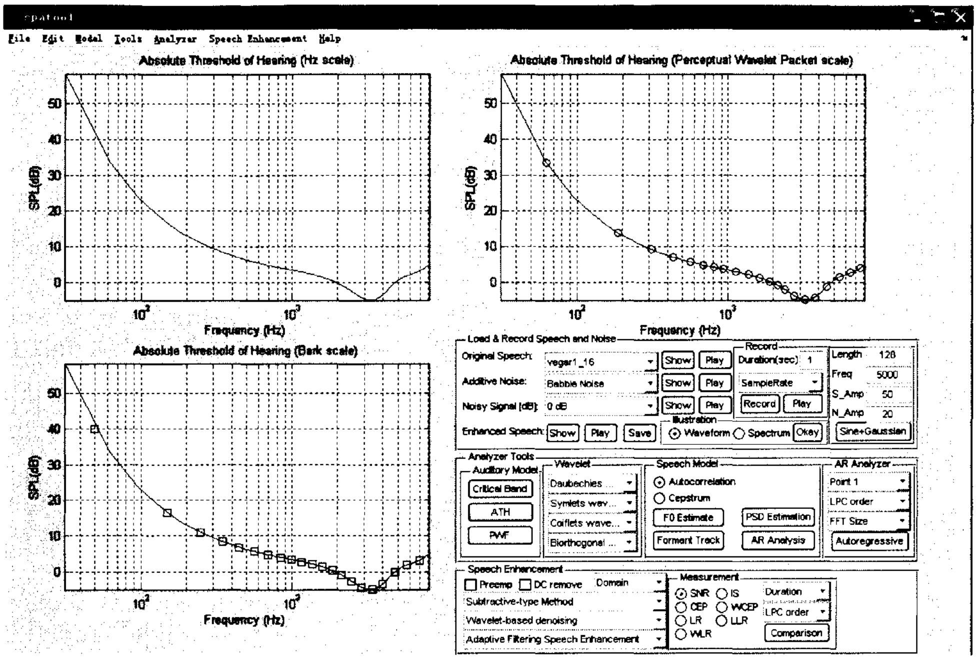
(2) Auditory model

This functionality provides an extension to Matlab speech processing toolbox. It is useful for the auditory modeling of speech signals. The term critical band refers to the frequency bandwidth of the loosely defined auditory filter. Literally, it refers to the specific area on the basilar membrane that goes into vibration in response to an incoming wave. Critical band analyses can be used to explain the auditory masking phenomena. The absolute threshold of hearing (ATH) is the minimum sound level of a pure tone that an average ear with normal hearing can hear in a noiseless environment. In this toolbox, the modeling of critical band and absolute threshold of hearing are investigated and implemented using the Matlab program. The top-right pane of Fig. 6.4 illustrates the wavelet filterbank structure to realize the critical band. The top-left pane of Fig. 6.4 shows the comparison of the critical band rate in Bark scale and the proposed perceptual wavelet packet scale in 0 to 8 kHz range. The bottom-left pane of Fig. 6.4 shows the comparison of the critical bandwidth in Bark scale and the perceptual wavelet packet scale. Fig. 6.5 shows the comparison of ATH in the Bark scale and the perceptual wavelet packet scale. These vivid graphical manifestations have been used in Chapter 3 to show the good match between the proposed perceptual wavelet filterbank and the original critical bank modeling of human auditory system.



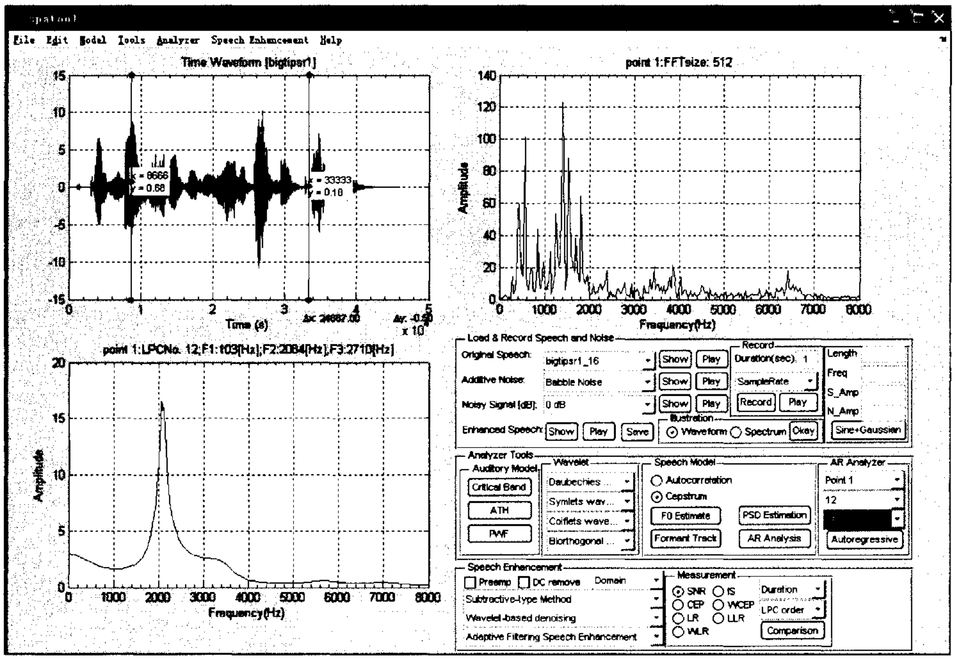
Original Speech: vegar1_16
Additive Noise: Babble Noise
Noisy Signal [dB]: 0 dB
Record Duration(sec) 1
SampleRate
Length 128
Freq 5000
S_Amp 50
N_Amp 20
Illustration: ● Waveform ○ Spectrum
Speech Model: ● Autocorrelation
□ Preemp □ DC remove
Measurement: ● SNR

Figure 6.4 GUI for critical bank rate analyses in Bark scale and perceptual wavelet packet scale


File Edit Model Tools Analyzer Speech Enhancement Help
Absolute Threshold of Hearing (Hz scale)
Absolute Threshold of Hearing (Perceptual Wavelet Packet scale)
Absolute Threshold of Hearing (Bark scale)
SPL(dB)
Frequency (Hz)
Load & Record Speech and Noise
Show Play Record Save
Illustration
○ Waveform ○ Spectrum Okey
Sine+Gaussian
Analyzer Tools
Auditory Model
Critical Band ATH PWF
Wavelet
Daubechies Symlets wav... Coiflets wave... Biorthogonal ...
Speech Model
○ Autocorrelation ○ Cepstrum
F0 Estimate PSD Estimation Formant Track AR Analysis
AR Analyzer
Autoregressive
Speech Enhancement
□ Preemp □ DC remove Domain
Subtractive-type Method
Wavelet-based denoising
Adaptive Filtering Speech Enhancement
Measurement
○ SNR ○ IS ○ CEP ○ WCEP ○ LR ○ LLR ○ WLR
Duration LPC order Comparison

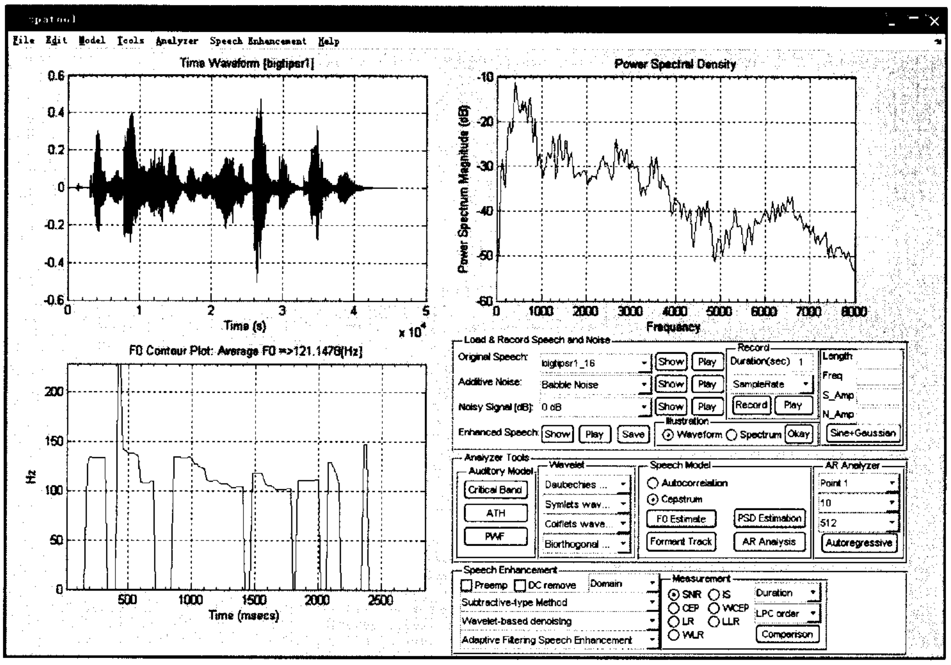
Original Speech: vegar1_16
Additive Noise: Babble Noise
Noisy Signal [dB]: 0 dB
Duration(sec): 1
SampleRate
Length 128
Freq 5000
S_Amp 50
N_Amp 20
Point 1
LPC order
FFT Size

Figure 6.5 GUI display of ATH in Bark scale and perceptual wavelet packet scale


File Edit Model Tools Analyzer Speech Enhancement Help
Time Waveform [bigtipsr1]
x = 8666
y = 0.68
x = 33333
y = 0.18
Time (s)
x 10^4
point 1:FFTsize: 512
Amplitude
Frequency(Hz)
point 1:LPCNo. 12;F1:1031[Hz];F2:2084[Hz];F3:2710[Hz]
Load & Record Speech and Noise
Show Play Record Save
Illustration
○ Waveform ○ Spectrum Okey
Sine+Gaussian
Analyzer Tools
Auditory Model
Critical Band ATH PWF
Wavelet
Daubechies Symlets wav... Coiflets wave... Biorthogonal ...
Speech Model
○ Autocorrelation ○ Cepstrum
F0 Estimate PSD Estimation Formant Track AR Analysis
AR Analyzer
Autoregressive
Speech Enhancement
□ Preemp □ DC remove Domain
Subtractive-type Method
Wavelet-based denoising
Adaptive Filtering Speech Enhancement
Measurement
○ SNR ○ IS ○ CEP ○ WCEP ○ LR ○ LLR ○ WLR
Duration LPC order Comparison

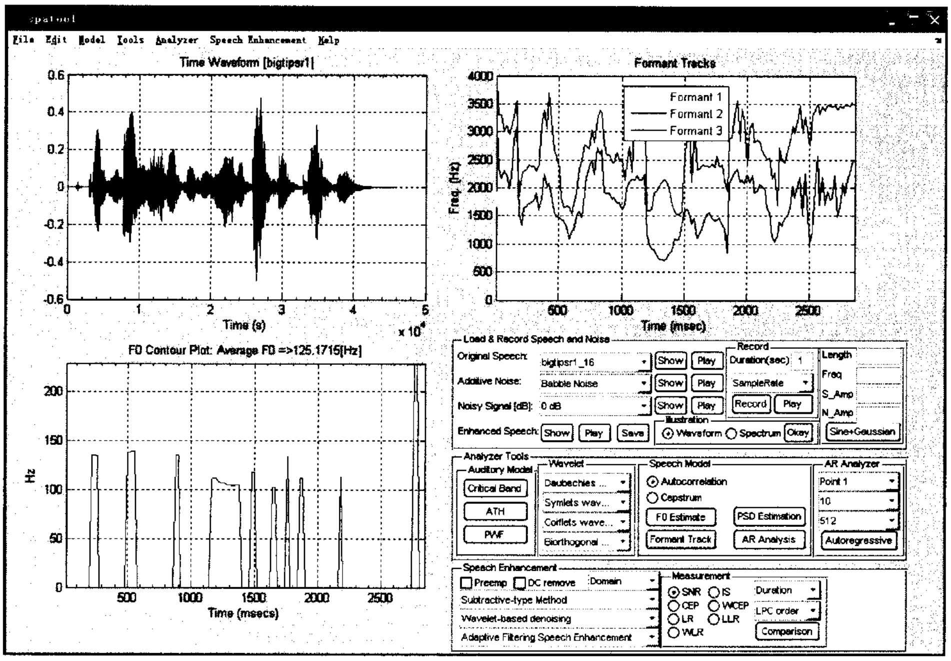
Original Speech: bigtipsr1_16
Additive Noise: Babble Noise
Noisy Signal [dB]: 0 dB
Duration(sec): 1
SampleRate
Length
Freq
S_Amp
N_Amp
Point 1
12

Figure 6.6 GUI for analysis of linear predictive coefficients and frame-based speech transform



Original Speech: bigtipsr1_16
Additive Noise: Babble Noise
Noisy Signal [dB]: 0 dB
Enhanced Speech:
Duration(sec): 1
SampleRate
Illustration: ● Waveform ○ Spectrum
○ Autocorrelation
● Cepstrum
Point 1
10
512
□ Preemp □ DC remove Domain
Subtractive-type Method
Wavelet-based denoising
Adaptive Filtering Speech Enhancement
● SNR ○ IS
○ CEP ○ WCEP
○ LR ○ LLR
○ WLR
Duration
LPC order

Figure 6.7 GUI display of power spectral density and F0 contour plots



Original Speech: bigtipsr1_16
Additive Noise: Babble Noise
Noisy Signal [dB]: 0 dB
Enhanced Speech:
Duration(sec): 1
SampleRate
Illustration: ● Waveform ○ Spectrum
● Autocorrelation
○ Cepstrum
Point 1
10
512
□ Preemp □ DC remove Domain
Subtractive-type Method
Wavelet-based denoising
Adaptive Filtering Speech Enhancement
● SNR ○ IS
○ CEP ○ WCEP
○ LR ○ LLR
○ WLR
Duration
LPC order

Figure 6.8 GUI display for formant tracking

(3) Speech model

This toolbox provides a list of functions for speech analysis and synthesis. Fig. 6.6 shows the linear predictive coefficient analysis and the representation of frame-based speech signal. The filter order and fast Fourier transform (FFT) point can be set according to the user's specification. The top-right pane of Fig. 6.7 shows the power spectral density (PSD) of the input speech. The bottom-left pane of Fig. 6.7 shows the contour plot of the fundamental frequency (F0). Two methods can be selected for the estimation of F0. One method is based on the autocorrelation and the other is based on the cepstrum. The top-right pane of Fig. 6.8 shows the formant tracking of the input speech signal. These functions can be used for speech modification and speech enhancement applications.

(4) Speech enhancement functions

The genesis of the development of this toolbox is the evaluation of speech enhancement algorithms. The toolbox now encompasses a list of subtraction methods (Amplitude spectral subtraction, Power Spectral Subtraction, Scalar Power Spectral Subtraction, Parametric Spectral Subtraction and Multi-Band Spectral Subtraction), wavelet based methods (Wavelet and Wavelet Packet denoising methods with soft thresholding), and adaptive filtering methods (Short time Spectral Amplitude MMSE Method, Short time Spectral Amplitude log-spectral MMSE Method, Speech Enhancement using a Noncausal A Priori SNR Estimator, Kalman Filtering Speech Enhancement for White Noise, Kalman Filter with Time-Frequency Masking, Kalman Filtering for Colored Noise, Wavelet Kalman Filtering for Colored Noise, Wavelet Packet Kalman Filtering for Colored Noise and Wavelet Kalman Filtering with Post-filter). The experimental results of Chapters 4 and 5 are generated by this toolbox. Fig. 6.9 illustrates the spectrograms of the original speech, noisy speech in speech-like noise, and speech enhanced by the Kalman filtering method. The toolbox also provides several objective measurement functions for signal-to-noise ratio (SNR) evaluation and various distortion measurements.

This toolbox has GUI with very simple viewing panes that are responsive to functional selection. Sound waveforms are stored as one-dimensional arrays which can be further processed according to the user selection. This section is not meant to be a detailed description of each available function of this toolbox. Most function descriptions including references that can be found in the manual [SPA07]. This software has been tested on Linux and Windows XP computers running with MATLAB 7.1. The codes are designed to be portable across all machines that support MATLAB programs. To demonstrate its prowess, we now present a hybrid speech recognition platform spawned from the use of this toolbox.

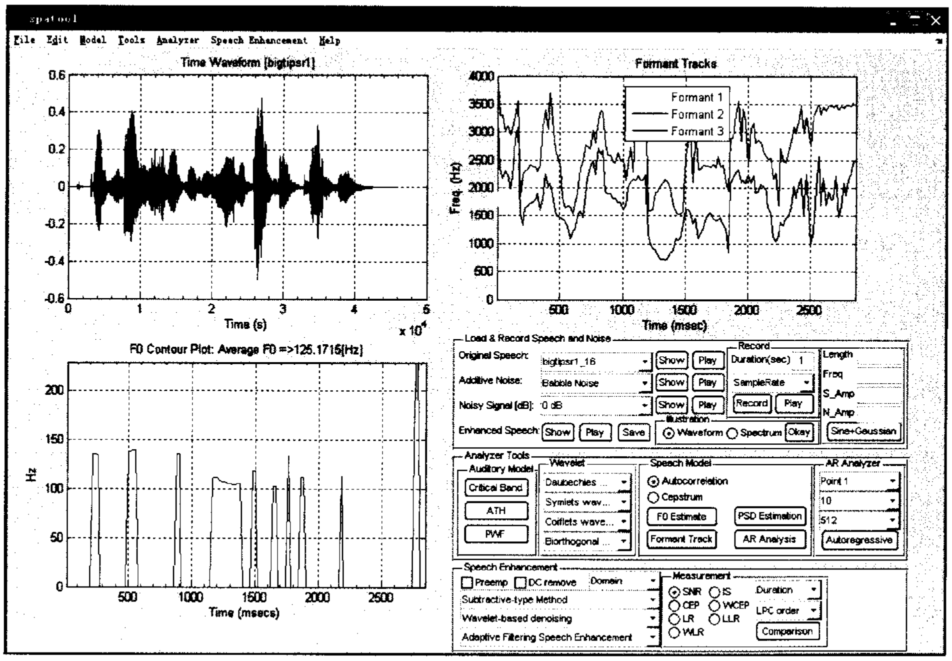

Original Speech: bigtipsr1_16
Additive Noise: Babble Noise
Noisy Signal [dB]: 0 dB
Duration(sec) 1
Illustration: ● Waveform ○ Spectrum
Speech Model: ● Autocorrelation ○ Cepstrum
AR Analyzer: Point 1, 10, 512
Measurement: ● SNR ○ IS ○ CEP ○ WCEP ○ LR ○ LLR ○ WLR

Figure 6.9 GUI display of speech spectrogram of original, noisy and enhanced speech

6.3 Hybrid Automatic Speech Recognition Framework

Sensitivity to speech variability, inadequate recognition accuracy, and susceptibility to impersonation are among the main technical hurdles that prevent a widespread adoption of speech-based recognition systems. Speech recognition systems work reasonably well in quiet conditions but poorly under noisy conditions or in distorted channels. Such mismatch in training and testing has severely limited the effectiveness of ASR in practical applications. For example, the accuracy of a speech recognition system may be acceptable if you call from the phone in your quiet office, yet its performance can be unacceptable if you try to use your cellular phone while driving. There are many types of mismatch sources, e.g., ambient background noise, microphone mismatch, and variation in speech styles. Therefore, we choose to study the amalgamation of speech processing algorithms to improve the robustness of speech recognition system in noisy conditions not originally present in the training data used for building the recognizer.

In most classifiers, tradeoff control for data fitting and classification is vital to the success of the approach. As human speech is highly variable, no single recognition approach appears to be capable of uniformly good performance. Our goal here is use the speech processing toolbox to build a hybrid speech recognizer prototype that operates well in different noise conditions and achieves high classification rate and low word error rate. A hybrid platform can combine the advantages of different design paradigms within a single system to augment the deficiencies of any approach operating in isolation. Various state-of-the-art methodologies have been carefully selected for this evaluation. Two hybrid classifiers are constructed and evaluated: a hybrid Multilayer Perceptron (MLP) [DUD01], [HAG96] and Hidden Markov Model

(HMM) [RAB89], [DUD01], [MOO97] network, and a hybrid Support Vector Machine (SVM) [BUR98], [CRI00], [DUD01], [VAP98] and HMM recognizer. These recognizers fully tap the discriminative training power of MLP and SVM onto the strong modeling ability of HMM for time-sequence structure [TRE03]. The model parameters of these original speech recognizers are typically calibrated to compensate for the mismatch during the recognition stage by incorporating appropriate optimization criteria, such as minimum classification error [JUA92]. However, perfecting the classification performance to reduce the mismatch between the training and testing conditions is impractical. A more phonetic sensitive approach is proposed to augment the hybrid classifier performances by robust speech features with speech enhancement. Instead of working with noisy features, the denoised features are attended.

To reduce the mismatch between the training and testing conditions of the classifier, the generalized perceptual time-frequency subtraction method (GPTFS) [SHA07a], and the perceptual wavelet-Kalman filtering speech enhancement (PWKF) [SHA06b], presented in Chapters 4 and 5, respectively are employed to provide an optimal estimate the clean speech from noisy observation. These methods have not been applied to the speech classification problem before. They are developed in the wavelet domain and can be made to blend well with the selected feature space. This section initiates further investigation into the potential improvement they can render to the proposed hybrid classifier on the word recognition rate under the noisy test data set.

6.3.1 The Evaluation Framework

A platform for robust ASR algorithms is constructed around three distinctive operations of preprocessing, feature extraction, and recognition. The robustness and efficiency of an ASR system are intricately related to these operations by the effect of space transformation, as shown in Fig. 6.10.

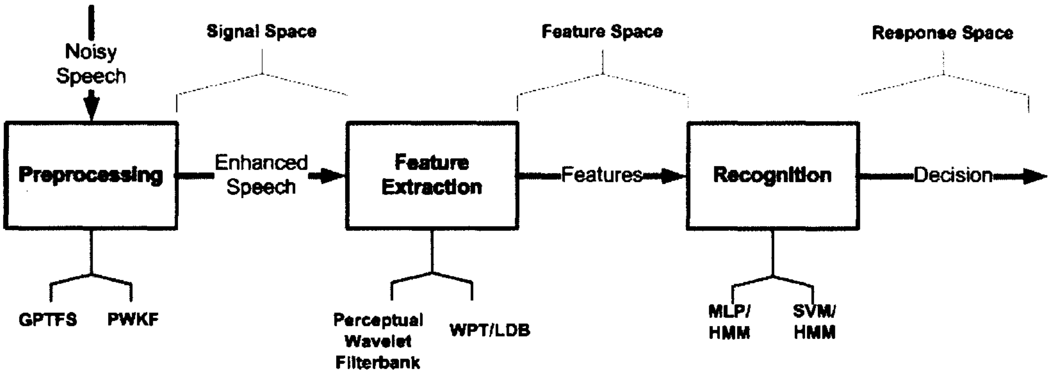


Figure 6.10 GUI display of speech spectrogram of original, noisy and enhanced speech

Speech recognition solutions can be established by selecting and mapping the appropriate subspaces of input speech signals, extracted features and responses. A speech signal space $\mathbf{X} \subset \mathbb{R}^n$ is a subset of the standard n-dimensional vector space, which contains all speech signals used in the learning and recognition tasks. The dimensionality of the speech signal space is very high compared to the response space. The additive noise mixed in the speech signals makes classification even more difficult in the noisy case. To improve the classifier's performance, from both the accuracy and efficiency perspective, it is imperative to extract only the relevant features. The resulting feature space after discarding the redundant information is denoted by $\mathbf{F} \subset \mathbb{R}^k$, where $k \leq n$. The associated feature extractor is defined as the mapping $f : \mathbf{X} \to \mathbf{F}$. The classifier is defined by a mapping $g:\mathbf{F}\rightarrow\mathbf{Y}$. The response space, $\mathbf{Y}=\{1,2,\ldots,C\}$, defines a set of class names to which the input speech signals belong.

The evaluation of the above framework is demonstrated on, and in part motivated by the task of feature extraction. The purpose is to integrate the extractor of robust features seamlessly with its front-end processor to minimize the misclassification rate at the backend due to the presence of noise. The function of feature extractor can be expressed in a general form [SAI95] as:

$$f=\Theta^{(k)}\circ\Psi \tag{6.1}$$

where Ψ is an n-dimensional orthogonal matrix representing an orthonormal base in the basis library. $\Theta^{(k)}:\mathbf{X}\rightarrow\mathbf{F}$ represents the selection rule. It selects k principal coordinates from n coordinates. To enable the feature extractor to operate harmoniously with the proposed speech enhancement method, the feature space is spanned by the basis vectors of wavelet packet transform. Wavelet packet transform provides a good unification between speech denoising, feature extraction and classification as it partitions the frequency axis smoothly with sufficiently good resolution for the analysis of transients. In what follows, we present the various components for the proposed integrated framework of ASR.

6.3.2 Feature Extraction

The goal of feature extractor is to seek distinguishing features that are invariant to irrelevant transformations of the input signal. The transformation from the input signal space to the feature space is domain specific. In speech recognition, rate variation is a serious problem as not only different people talk at different rates, but an individual may also vary in his rate of speech. Good features possess the attributes of being simple to extract, invariant against time translations and overall amplitude variation, and capable of discriminating different phonetic categories. Wavelet packet (WP) [STR96] makes good feature extractor in this sense. It recursively segments the speech signal in dyadic frequency bands and has uniform translation in time. Two methods to extract the wavelet coefficients are evaluated. One uses the local discriminant bases (LDB) to winnow the orthonormal bases of wavelet packet transform (WPT) [SAI95]. Relative entropy is used to select the best basis for maximal discrimination of different features [COI92]. The other uses the perceptual wavelet filter bank [SHA07a]. The input speech is transformed into the wavelet coefficients by imitating the frequency-spatial excitation modus of the basilar membrane of human cochlea.

a) Feature Extraction by WPT/LDB

The merits of Wavelet Packet Transform (WPT) coupled with the Local Discriminant Bases (LDB) [SAI95] are exploited to maximally separate the distances among classes. The WPT/LDB feature extractor selects a best basis from a dictionary to capture the most discriminative information for a given set of classes in the time-frequency plane. The localized features obtained from these basis functions are then fed to the recognizers adopted in our proposed ASR platform. The WPT/LDB based feature extraction is described as follows.

A dictionary of orthonormal bases from the wavelet packet transform is first generated by an iterative decomposition process. The input signal, $x \in \mathbf{X}$ of length n is divided into two frequency bands, Hx and Gx, each of length $n/2$ by the convolution-subsampling operations. The H and G operators can be specified by a large number of orthonormal wavelet bases to satisfy the perfect reconstruction criteria. Thus, repeated application of the convolution-subsampling essentially decomposes a vector space into mutually orthogonal subspaces in a binary tree. These subspaces, $\Omega_{j,k}$ with different frequency localization characteristics are spanned by the wavelet basis vectors, $\{w_{j,k,l}\}$ where the triplet (j, k, l), $j = 0, 1, \ldots, J$, $k = 0, 1, \ldots, 2^{j-1}$ and $l = 0, \ldots, 2^{n_0-j}-1$, indicate the scale, location and frequency, respectively. For a signal frame of length n, $n_0 = \lfloor \log_2 n \rfloor$, and the depth of decomposition, $J \leq n_0$.

For the classification problem of interest, a training dataset, $T = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathbf{X} \times \mathbf{Y}$ with N pairs of speech signals $x_i \in \mathbf{X}$ and class names, $y_i \in \mathbf{Y}$ is required to construct the WPT/LDB based feature extractor. The signals in the dataset are divided into C different classes so that $N = N_1 + N_2 + \ldots N_C$, where N_c denotes the number of signals belonging to class c. Let $\{x_i^{(c)}\}$ denote the set of signal in class c. The time-frequency energy (TFE) maps Γ_c for $c = 1, 2, \ldots, C$ are computed by [SAI95]:

$$\Gamma_c(j,k,l) = \sum_{i=1}^{N_c} \left(w_{j,k,l}^T x_i^{(c)} \right)^2 \Big/ \sum_{i=1}^{N_c} \left\| x_i^{(c)} \right\|^2 \tag{6.2}$$

where $\|.\|$ is the norm operator and w is an orthogonal base. The inner product, $w^T x$ is the set of wavelet expansion coefficients.

Being a good differentiator among sounds of weak intensity, for example, fricatives and stop consonants, TFE is used to evaluate the power of discrimination of each subspace in the binary decomposition tree. To select a complete basis with the maximum discriminant information from a library of orthonormal bases, an additive discrimination information function (DIF) is adopted in the best-basis search algorithm [COI92]. Since DIF is additive, it can be expressed using the TFE as [SAI95]:

$$D\left(\left\{\Gamma_c(j,k)\right\}_{c=1}^{C}\right) = \sum_{l=0}^{2^{n_0-j}-1} D\left(\Gamma_1(j,k,l),\ldots,\Gamma_C(j,k,l)\right) \tag{6.3}$$

Let $\{B_{j,k}\}$represent a set of basis vectors in the subspace $\Omega_{j,k}$. Given a decomposition depth, J, the LDB selection algorithm [SAI95] shown in Fig. 6.11 is used to obtain the best subspace $\{A_{j,k}\}$ that maximizes the TFE distributions of classes in the span of $\{B_{j,k}\}$. A training dataset of C classes of signals is assumed for the feature extractor of our ASR platform.

```
LDB($J$, $\{B_{j,k}\}$, $\{\Gamma_c\}$, $D$) {
    for ($k = 0$ to $2^J-1$) $A_{J,k} = B_{J,k}$;
    for ($j = J-1$ to $0$)
        for ($k = 0$ to $2^J-1$)
            if ($D\left(\{\Gamma_c(j,k)\}_{c=1}^{C}\right) \geq D\left(\{\Gamma_c(j+1,2k)\}_{c=1}^{C}\right) + D\left(\{\Gamma_c(j+1,2k+1)\}_{c=1}^{C}\right)$)
                $A_{j,k} = B_{j,k}$;
        else {
            $A_{j,k} = A_{j+1,2k} \oplus A_{j+1,2k+1}$;
            $D\left(\{\Gamma_c(j,k)\}_{c=1}^{C}\right) = D\left(\{\Gamma_c(j+1,2k)\}_{c=1}^{C}\right) + D\left(\{\Gamma_c(j+1,2k+1)\}_{c=1}^{C}\right)$;
            }
return $\{A_{j,k}\}$;
}
```

Figure 6.11 The LDB selection algorithm

The above procedure generates a complete orthonormal basis LDB. To obtain the best basis, the basis functions are ordered by their power of discrimination and the most discriminant basis functions are selected. The wavelet coefficients of the signal in the selected basis are used as features for the classifier. By winnowing out the irrelevant subspaces, WPT/LDB improves the recognition rate. The drawbacks of WPT/LDB are the amount of computations, and the heavy dependency of the discriminant features on the training set. This is because WPT/LDB features are selected based on the statistical energy distributions of the signals in the training dataset without considering the human perceptual sensitivity.

b) Feature Extraction by Perceptual Wavelet Filterbank

In order to overcome the above problems, a fixed set of basis is proposed. A fixed partitioning of the frequency axis is made to closely match the critical band scale of human auditory system. The result is a filterbank with an admissible binary wavelet packet tree structure.

The wavelet packet decomposition is accomplished with the Daubechies basis. A six level tree structure of cascaded filter banks is used to approximate the critical band scale of human auditory system. The details have been given in Chapters 3 and 5 and will not be iterated here. After a speech of 32 ms duration has been decomposed by the analysis filterbank, the logarithm of the energy in each of these 21 frequency bands is calculated. A discrete cosine transform (DCT) is applied on these 21 logarithmic energy coefficients and the first 13 DCT coefficients are taken as the features for classification. This is because the dynamic range effects of human auditory system cause its response to be more closely approximated by the logarithmically positioned perceptual frequency scale than the linearly-spaced frequency bands obtained directly from the transform. The calculation of the spectrum of a spectrum is inspired by the Mel Frequency Cepstral Coefficients (MFCCs), which are based on Fourier transform. The ceptrum has shown to be more robust than the direct energy spectrum in speech recognition [DEL00], [FAR04]. What we proposed here takes advantage of the fundamental merit of the critical-band scale, which has not been considered in MFCC, while keeping the spectral analysis for the extraction of the relevant parameters from the speech signals to discriminate between the phonemes.

6.3.3 Hybrid Speech Recognizers

After feature extraction, the wavelet coefficients will be fed to the speech recognizer block. The role of the recognizer is to use the feature vector provided by the feature extractor to assign it to a category. The degree of difficulty of the classification problem depends on the variability in the feature values for speech signal in the same category relative to the difference between feature values for speech signal in different categories.

The stochastic properties of the elementary sounds in a language are usually modeled with mixtures of diagonal covariance Gaussians. Thus, the inherent variability of speech is generally modeled by hidden Markov models (HMMs) [DUD01] with Gaussian emission densities in the state-of-the-art speech recognition systems. HMM consists of nodes representing hidden states, interconnected by links describing the conditional probabilities of a transition between states. Each hidden state also has an associated set of probabilities of emitting particular visible states. HMM is particularly useful in modeling context dependent sequences, such as the phonemes in speech. All the transition probabilities can be learned iteratively from sample sequences by means of the forward-backward or Baum-Welch algorithm [DUD01]. However, the classification performance of HMMs is limited intrinsically by their arbitrary parametric assumption. This can be improved by learning the decision regions discriminatively. Two discriminative techniques, MLP and SVM, respectively are adopted to improve the estimation of HMM parameters.

a) Hybrid MLP/HMM Recognizer

A series of theoretical and experimental results have suggested that multilayer perceptron (MLP) is effective for smooth estimate of highly-dimensioned probability density functions in speech recognition. The advantages of MLPs are the invariance of the structure with respect to linear transformations, the discriminative training and the availability of a posteriori probabilities that allow sound confidence measures. However, the drawback is the basic structure itself is insufficient to provide a quality solution to the problem of interest. Therefore, we introduce a hybrid MLP/HMM system, in which the MLP is trained to estimate the emission probabilities of HMM states. The approach is built on a unified framework of the hybrid systems proposed by Bourlard and Morgan [BOU94], by extending their benefits and overcoming their limitations.

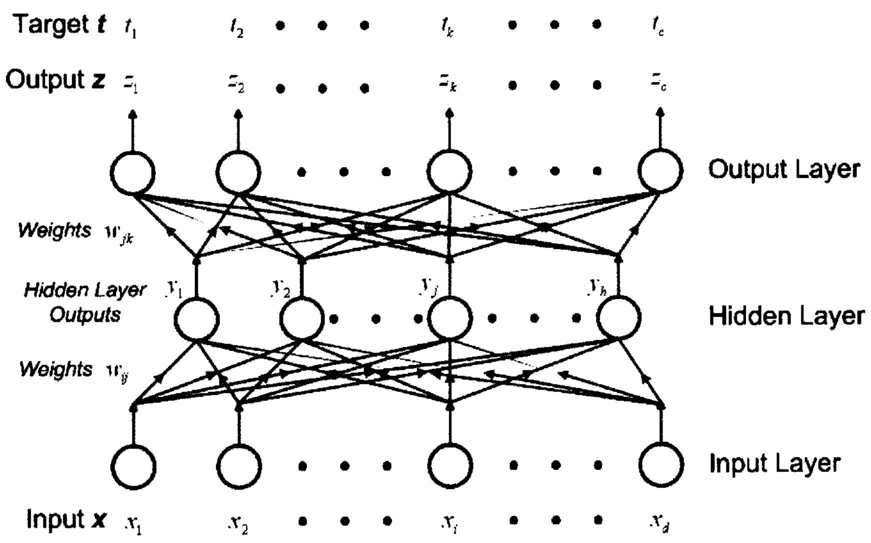

Target t t_1 t_2 t_k t_c
Output z z_1 z_2 z_k z_c
Output Layer
Weights w_{jk}
Hidden Layer Outputs y_1 y_2 y_j y_h
Hidden Layer
Weights w_{ij}
Input Layer
Input x x_1 x_2 x_i x_d

Figure 6.12 A *d-h-C* fully connected three-layer neutral network

Fig. 6.12 shows a fully connected three-layer neural network. The neurons in the input, hidden and output layers are interconnected by links between layers with modifiable weights. The weights are initialized to some nonzero values. Then, a feedforward procedure is used to establish the *C* discriminant functions in the output layer from the feature vector, **x** [HAG96], [DUD01].

1. The feature vector, $\mathbf{x} = \{x_1, x_2, \ldots, x_d\}$ is present to the input layer, where *d* is the number of neurons in the input layer.
2. The net activation, $net_j = \sum_{i=1}^{d} x_i w_{ij}$ of the *j*-th neuron in the hidden layer is computed for $i = 1, 2, \ldots, d$ and $j = 1, 2, \ldots, h$, where w_{ij} is the connection weights from the *i*-th neuron in the input layer to the *j*-th neuron in the hidden layer.
3. The output, $y_j = f(net_j)$ of the *j*-th neuron in the hidden layer is computed for $j = 1, 2, \ldots, h$. For speech recognition, a sigmoid function is used as the activation function, *f*.
4. The net activation, $net_k = \sum_{j=1}^{h} y_j w_{kj}$ and the output, $z_k = f(net_k)$ of the *k*-th neuron in the output layer are similarly computed for $k = 1, 2, \ldots, C$. i.e.,

$$z_k = g_k(\mathbf{x}) = f\left(\sum_{j=1}^{h} w_{jk} f\left(\sum_{i=1}^{d} w_{ij} x_i\right)\right) \tag{6.4}$$

After feedforwarding, the weights are to be set by supervised learning based on a training dataset and its target outputs. This is achieved by a stochastic backpropagation algorithm based on the least-mean-square (LMS) rule [DUD01]. The weights are updated iteratively by a learning function as follows:

$$\begin{aligned} w_{ij}(m) &= w_{ij}(m-1) + \eta\delta_j(m)x_i(m) \\ w_{jk}(m) &= w_{jk}(m-1) + \eta\delta_k(m)y_j(m) \end{aligned} \tag{6.5}$$

where $w_{ij}(m)$ and $w_{jk}(m)$ are the connection weights between two neurons at the m-th iteration. $\mathbf{x}(m)$ is a randomly chosen reference vector from the training set. η is the learning rate that determines the fraction of change, and $\delta_j(m)$, $\delta_k(m)$ are the sensitivity of the overall error changes to the net activation of the j-th neuron and k-th neuron, respectively. The learning converges when $\sum_{k=1}^{C} \frac{1}{2} \|\mathbf{t}-\mathbf{z}\|^2 < \varepsilon$, i.e., this learning process terminates when the distance between the desired output, t_k and the network output, z_k summed overall output neurons is less than an infinitesimal error, ε.

As the outputs of the network approximate the true a posteriori probabilities in a least-square sense, MLP is incorporated into HMM by feeding the output to the standard Viterbi recognizer [DUD01].

b) Hybrid SVM/HMM Recognizer

Support Vector Machines (SVM) [BUR98], [CRI00] are universal learning systems that use a hypothetical space of linear functions in a high dimensional feature space to increase the generalization capability of the classifier. Support vector (SV) kernel is used to map the data from the original feature space to a higher dimension feature space such that the problem becomes linearly separable. The decision boundary is obtained from the training data by finding a separating hyperplane that maximizes the margins between classes. The training algorithm [VAP98] is, in essence, solving a constrained quadratic optimization problem. However, SVMs alone can not model the temporal structure of speech effectively. Therefore, we propose a more flexible hybrid design by using SVM to estimate the a posteriori probabilities and the HMM structure to model the temporal evolution.

In SVM, each input feature, x_i, $i = 1, 2, \ldots, N$, is transformed to $y_i = \phi(x_i)$ by a nonlinear mapping function ϕ, so that a linear discriminant function can be defined [VAP98].

$$f(x) = \langle w \cdot \phi(\mathbf{x}) \rangle = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) = \sum_{i=1}^{N} w_i \phi(x_i) \tag{6.6}$$

where $\mathbf{w}$ represents the weight vector characterizing the classifier. The nonlinear mapping function, ϕ can be polynomial, Gaussian, or other basis functions depending on the problem. And $\langle a \cdot b \rangle$ denotes the inner product of vectors a and b. The dimensionality of the mapped space can be arbitrarily high. Since the data are linearly separable, there exists many hyperplanes that satisfy [VAP98]

$$y_i f(x_i) \geq 1, \quad i = 1, \ldots, N \tag{6.7}$$

The support vectors are the training data for which $y_i f(x_i) = 1$, $i = 1, 2, \ldots, N$. It defines an optimal hyperplane that separates the independent uniformly distributed training data with the largest margin. The margin of a training sample (x_i, y_i) is defined as the distance from any hyperplane to a feature vector [VAP98]:

$$\frac{y_i f(x_i)}{\|\mathbf{w}\|} \geq \gamma, \quad i = 1, \ldots, N \tag{6.8}$$

where γ denotes a positive margin.

The goal of training a SVM is to find the weight vector, **w** that maximizes γ of the training data. Under the constraint that $\gamma\|\mathbf{w}\| = 1$, we demand the solution to minimize $\|\mathbf{w}\|^2$. A constrained quadratic programming problem can be constructed by minimizing the following regularized risk function proposed by Vapnik *et al.* [VAP98]:

$$L(\mathbf{w}, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} w_i \left[y_i f(x_i) - 1 \right] \tag{6.9}$$

with respect to the weight vector, **w**, and maximizing it with respect to the Lagrange undetermined multiplier $\alpha_k \geq 0$. The last term in (6.9) expresses the goal of classifying the data correctly. Using the Kuhn-Tucker construction [VAP98], this optimization problem can be reformulated as:

$$\begin{gathered} \text{maximizing } L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{subject to the constraints } 0 \leq \alpha_i \leq Q \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0 \text{ for } i = 1, 2, \cdots, N \end{gathered} \tag{6.10}$$

where the SV kernel, $K(x_i, x_j) = <\phi(x_i) \cdot \phi(x_j)>$ must satisfy the Mercer's condition. Mathematically, a real-valued function $K(x, y)$ is said to fulfill Mercer's condition if for all square integrable functions $g(x)$, $\int\int K(x,y)g(x)g(y)dxdy \geq 0$ [ABE05]. The learning cost, Q is a positive integer constant. It is the parameter used to penalize training errors during the SVM estimation.

Three commonly used kernels to be explored in this study are [CRI00]:

$$\begin{aligned} K(x_i,x_j) &= \langle x_i \cdot x_j \rangle && \text{linear} \\ K(x_i,x_j) &= \left(\kappa\langle x_i \cdot x_j\rangle + \kappa\right)^d && \text{polynomial} \\ K(x_i,x_j) &= \exp\left\{-\kappa\left(\left|x_i - x_j\right|^2\right)\right\} && \text{radial basis} \end{aligned} \tag{6.11}$$

where κ is the kernel parameter. It is also the variance of the radial basis function (RBF). Although RBF kernel typically converges slower than the polynomial kernel [VAP98], it often delivers better performance. As the number of dot products correspond to the number of support vectors, the classification task scales linearly with N.

A less popular wavelet kernel [ZHA04] is also studied. The wavelet kernel is based on the wavelet analysis and its dot-product is defined as [ZHA04]:

$$K(x_i,x_j) = \prod_{i,j=1}^{N} h\left(\frac{x_i - \tau_i}{a}\right)h\left(\frac{x_j - \tau_j}{a}\right) \tag{6.12}$$

where $h(x)$ is a mother wavelet. a and τ denote the dilation and translation, respectively. The translation-invariant wavelet kernels can be further deduced to:

$$K\left(x_i, x_j\right) = \prod_{i=1}^{N} h\left(\frac{x_i - x_j}{a}\right) \tag{6.13}$$

The optimal wavelet coefficients in the space spanned by the multidimensional wavelet basis are found by the wavelet SVM. In general, non-separable data is addressed by the use of soft margin classifiers. Slack variables, ε_i are introduced to relax the separation constraints [CRI00]. Given a fixed target margin, $\gamma > 0$, the slack variable, ε_i of a training sample (x_i, y_i) with respect to a hyperplane is given by:

$$\varepsilon_i = \max\left(0, \gamma - y_i f\left(x_i\right)\right) \tag{6.14}$$

In order to combine SVM with HMM, a relationship between the distance from the margin and the a posteriori class probability is needed. A sigmoid distribution is used to map the output distance of the SVM to the a posteriori class probability.

$$P\left(y = 1 \mid f\right) = \frac{1}{1 + \exp\left(Af + B\right)} \tag{6.15}$$

where the parameters A and B are estimated using a model-trust minimization algorithm [PLA99].

In general, classifiers can be divided into two categories. The one-versus-one classifiers learn to discriminate one class from another class while the one-versus-all classifiers learn to discriminate one class from all other classes. According to some standard classification experiments, the one-versus-one classifier has higher classification rate and is computationally more efficient than the one-versus-all classifier. Thus, the one-versus-one classifier is used for the proposed hybrid system. The system is trained to recognize connected digit word, which can be transcribed as five phonemes: vowel, stop, fricative, sonorant consonant and silence, of variable duration. Segmentation is one of the most intricate problems as segment durations are correlated with the word choice and speaking rate. They are difficult to exploit in an SVM-type framework. In the proposed system, we have adopted a simple and yet effective approach by composing the segments in a fixed number of sections. This approach is motivated by the alignments for a baseline three-state HMM system.

The proposed hybrid system relies on a nonparametric estimate of the emission probabilities of an HMM. The SVM has an output unit for each state of an HMM in the recognition system. Each output units provides an estimate of the corresponding emission probability given the current observation in the feature space. Training of the other probabilistic quantities in the HMM, for instance, the initial and transition probabilities, still relies on the Baum-Welch algorithm [DUD01]. The Viterbi decoder [DUD01] is eventually applied to the final recognition step. It will be shown later in the experimental results that this hybrid SVM/HMM recognizer has better classification performance than the hybrid MLP/HMM recognizer.

6.3.4 Speech Enhancement Methods for ASR Front-End

The feature extractors and recognizers presented earlier are the cornerstones of an ASR system with promising performance for clean speech. In the presence of background noise, speech quality and intelligibility in the test dataset will be deteriorated and ASR systems that are designed or trained to act on clean speech signals might become fallible. To improve the performance of ASR in noisy environments, speech enhancement methods are incorporated into the front-end of the system. From the study of the previous chapters, Model-based methods normally outperform non-parametric methods, since they utilize more information in the form of a model of the signal process. However, they can be sensitive to deviations from the class of signals characterized by the model and are more complex than the former methods. Both methods are considered in the proposed ASR platform. For the non-parametric method, the generalized perceptual time-frequency subtraction (GPTFS) proposed in Chapter 4 is used and for the model-based method, the perceptual wavelet-Kalman filtering (PWKF) speech enhancement method proposed in Chapter 5 is employed.

6.4 Evaluation of ASR Systems with Denoised Wavelet Features

The techniques presented in the previous section are amalgamated into the proposed ASR framework of Fig. 6.10 to evaluate their augmented speech recognition performance. Experiments are set up to evaluate the speed recognition performance of the hybrid MLP/HMM system [BOU94], [MAR98] and hybrid SVM/HMM system against the baseline HMM system [COU00]. Different kinds of features such as Linear Predicative Cepstral (LPCC) [DEL90], Mel-frequency Cepstral Coefficients (MFCC) [VER99], Relative Spectral Technique and Perceptual Linear Prediction (RASTA-PLP) [HER94] and wavelet coefficients are applied to the prototype ASR platform. The feature vectors are extracted by either perceptual wavelet filterbank (PWF) in critical band scale [SHA07a] or wavelet packet transform of local discriminant bases (WPT/LDB) [SAI95]. For WPT/LDB extractor, the library of orthonormal bases includes Daubechies (1-10 order), Coiflets (1-5 order) and Symlets (2-8 order). Clean speech is used in the training and testing phases of the recognizer for stand alone evaluation. For complete system evaluation, in the testing phase, the noisy speech are preprocessed by either the generalized perceptual time-frequency subtraction (GPTFS) method or the perceptual wavelet Kalman filtering (PWKF) method before it is fed to the feature extractor. The feature vectors are classified into their corresponding category by the recognizer. Table 6.1 shows eight different configurations of the front-end preprocessing subsystem, feature extractor and recognizer. These combinations are evaluated in our proposed robust automatic speech recognition framework. The proposed techniques, e.g. PWF, GPTFS, and PWKF have been presented in Chapter 3, 4, 5, respectively.

Table 6.1 ASR Configurations

S/N	Enhancement	Feature Extraction	Recognition
1	GPTFS	PWF	MLP/HMM
2	PWKF	PWF	MLP/HMM
3	GPTFS	PWF	SVM/HMM
4	PWKF	PWF	SVM/HMM
5	GPTFS	WPT/LDB	MLP/HMM
6	PWKF	WPT/LDB	MLP/HMM
7	GPTFS	WPT/LDB	SVM/HMM
8	PWKF	WPT/LDB	SVM/HMM

Two English databases are used for the speaker-independent connected digit recognition. The first database is the Studio Quality Speaker-Independent Connected-Digits Corpus (TIDIGITS) from the National Institute of Standard and Technology in the USA [TID]. It extends the Isolated Digits test to recognize more than one word at a time (i.e., continuous speech). The vocabulary is merely the spoken digits from 0 through 9, with a single utterance containing a sequence of digits. The training dataset is taken from 112 speakers (57 males and 55 females) and the test data is taken from 113 different speakers (57 males and 56 females). The second database is CDIGITS corpus we collected ourselves. The speech was recorded in a laboratory with an average signal-to-noise ratio (SNR) of 35 dB. There were 5 speakers and each speaker was asked to speak each digit ten times, making up a total of 500 speech files. These data was sampled at 16 kHz and digitized to 14 bit resolution. 300 speech files were used as reference data and the remaining speech files as test data. The speech signal was pre-emphasized by a first order filter with coefficient 0.95 and multiplied by a 20 ms Hamming windows with an overlap of 10 ms.

The above settings are used to evaluate the classification performance of employing different hybrid recognizers with different extracted features and speech enhancement preprocessing on the ASR platform. The results are reported in terms of the recognition rate on the specific test database. The word recognition ratio (WRR) and word error ratio (WER), which are the two most popular speech recognition evaluation criteria, are used.

6.4.1 Hybrid Recognizer Evaluation

In this section, the performance of the SVM/HMM hybrid recognizer is evaluated by adopting WPT/LDB and our proposed PWF as the feature extractor, respectively. The purposes of our work are to evaluate which kernel is better adopted in SVM classifier and which feature extractor has better performance when it is combined with SVM classifier. Table 6.2 compares the results of four types of SVM kernels, i.e., linear, 4th degree polynomial, RBF and wavelet, in the training and testing phases using the CDIGITS database. The respective kernel parameters are set as follows: the amplification of the polynomial kernel is $\kappa = 4$. The RBF radius is tuned to $\kappa = 0.5$ to strike a good balance between the generalization ability for unknown data and the strong discrimination on the training set. As learning cost increases, the radius of the RBF kernel decreases. The learning cost Q defined in (6.10) is empirically set to 10. The parameter of wavelet kernel is $a = 4$.

It is found that the WRR of the wavelet kernel is higher than the linear and other nonlinear kernels. Based on this result, the wavelet kernel is adopted in the hybrid SVM/HMM for all the other experiments. From Table 6.2, it is found that PWF has almost the same performance as WPT/LDB. Although WPT/LDB can better discriminate the difference of each class based on the relative entropy, it requires a lot more computation than PWF, and the resultant tree of WPT/LDB depends heavily on the training set used. Besides, too many independent subbands lead to the negligence of much of the spectral dependency, which could result in poor phonetic discrimination in the presence of noise and interference. This is not the case for PWF. As pointed out in Chapter 3, the division of the speech frequency-band into subbands by PWF can effectively isolate any local frequency corruption from the other usable bands. In view of this, PWF is used to extract the proposed wavelet feature vectors for the following experiments.

Table 6.2 Word Recognition Ratio (WRR) of SVM Classifiers using Different Kernel Functions

Feature Extractor	Mode	WRR %			
		Linear	Polynomial	RBF	Wavelet
WPT/LDB	Training	82.81	93.85	96.55	96.67
	Testing	81.15	90.74	93.00	93.02
PWF	Training	81.53	92.24	94.39	94.54
	Testing	80.09	89.37	91.71	91.83

Table 6.3 shows the comparison results of speech recognition using different recognizers. All these recognizers are existing methods. The purpose of this evaluation

is to find which recognizer has better performance when it is combined with our proposed PWF. Each digit is represented by a HMM containing eight emitting states and one Gaussian distribution per state. Training is conducted with the training set using the Baum-Welch algorithm, and testing is accomplished by applying the test set to a Viterbi decoder. For the hybrid MLP/HMM recognizer, the architecture of the MLP is defined as a four-layer feed-forward net, with 500, 200 sigmoid units in two hidden layers and 80 sigmoid units in the output layer.

The learning rate, η and the termination criterion, ε are set to 0.1 and 0.01, respectively. From some preliminary experiments conducted on a reduced-scale task, on-line training in the linear domain yields slightly better results than batch training [TRE03]. For this reason, training was accomplished via on-line Bayesian and maximum a posteriori (MAP) algorithms in the linear domain. The Bayesian framework adopted here allows an objective setting of the regularization parameters, according to the training data [BOU94], [DUD01]. The MAP estimation employs an optimization objective that includes the a priori distribution of the quantity to be estimated [DUD01]. It can therefore be seen as a regularization of ML estimation.

Table 6.3 Word recognition ratio (WRR) of different recognizers

Speech Recognizer	WRR%	
	CDIGIT	TIDIGIT
HMM with 8-Gaussian mixtures	92.50	89.25
MLP/HMM hybrid trained via Bayes	94.54	92.45
MLP/HMM hybrid trained via MAP	95.45	93.74
SVM/HMM hybrid system	96.53	94.63

From Table 6.3, the proposed hybrid SVM/HMM with wavelet kernel has the best performance with the WRR of 96.53% and 94.63% for CDIGIT and TIDIGIT databases, respectively. HMM is good at dealing with sequential inputs and SVM possesses good generalization properties. They can be combined to capitalize on these advantages to yield an effective hybrid classifier.

6.4.2 Feature Vectors Evaluation

The word error recognition ratio (WER) due to the use of different feature vectors is investigated in this section. In this case, a lower WER implies a better performance. Fig. 6.13 shows the WER comparison of speaker independent discrete HMM, MLP/HMM and SVM/HMM recognizers using different kinds of parameters to represent the speech signal. The speech database used in this experiment is the CDIGITS corpus. We used the LPCC, MFCC and their delta derivative (MFCC+delta) and RASTA-PLP as the competitive features against the proposed wavelet coefficient features. Comparing with other competitive feature representations, the proposed wavelet coefficients produce the lowest WER on any system. In particular, a very low word error ratio of 3.5% was observed in Fig. 6.13 when they are applied to the hybrid SVM/HMM recognizer. This is because the wavelet coefficients are able to capture the localized information of the transient signal such as fricative, nasal and unvoiced speech, which is difficult to be represented by other traditional features.

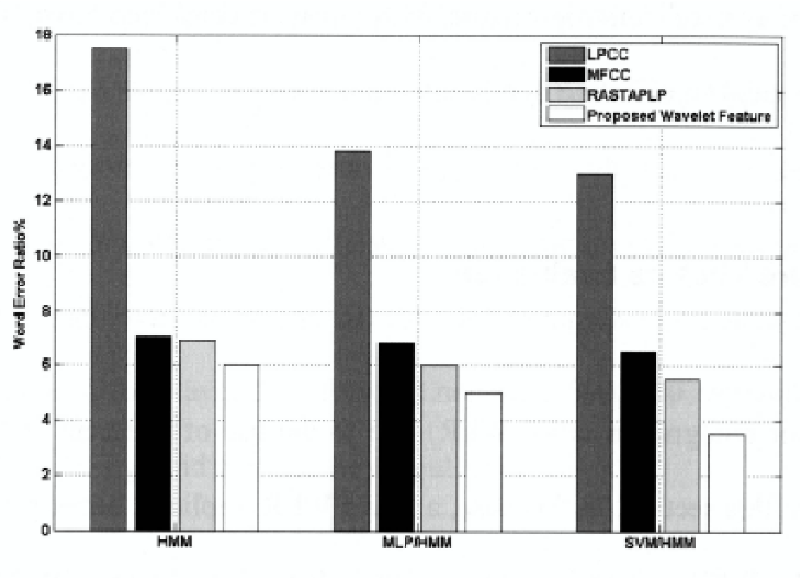


Figure 6.13 WER performances of HMM, MLP/HMM and SVM/HMM recognizers using different kinds of features

6.4.3 Robustness Evaluation

In this section, the effect of speech enhancement, which is incorporated as a front-end preprocessor of the proposed ASR platform, is tested for speaker independent digits recognition. It was evaluated by using a database created by adding different types of background noise to the CDIGITS corpus. In the testing phase, the utterances were corrupted by various types of additive noise, i.e., white Gaussian noise, pink noise, car engine noise, and speech-like noise from the Noisex-92 database [NOI92]. The noise can be stationary or time-varying, and no knowledge about the noise characteristics is assumed.

Table 6.4 Comparison of speaker independent recognition rate improvements based on hybrid SVM/HMM recognizer with different speech enhancement methods

Noise	Method	0 dB	5 dB	10 dB
WGN	GPTFS	20%	16%	13%
	PWKF	25%	22%	19%
Pink noise	GPTFS	17%	15%	12%
	PWKF	24%	20%	16%
Car Engine Noise	GPTFS	21%	19%	17%
	PWKF	26%	23%	20%
Speech-like Noise	GPTFS	15%	2%	0%
	PWKF	18%	5%	0%

The ASR word recognition rate improvements with the use of GPTFS and PWKF speech enhancement methods over the same ASR without speech enhancement preprocessing are summarized in Table 6.4 for various types of noise and SNR conditions. The proposed ASR platform based on hybrid SVM/HMM recognizer leads to a significant recognition rate improvement in all cases, except for speech-like noise at SNR above 5 dB. This is because speech and noise are in the same frequency range, causing an increased distortion in the speech enhancement process. Both PWKF and GPTFS can reduce the mismatch between training and testing conditions. The best speech recognition ratio is obtained from PWKF. From the perspective of noise reduction, PWKF outperforms GPTFS due to a more optimal estimate of the clean speech from the noisy speech. The statistical-based model harnesses a more precise time-frequency analysis of speech and noise.

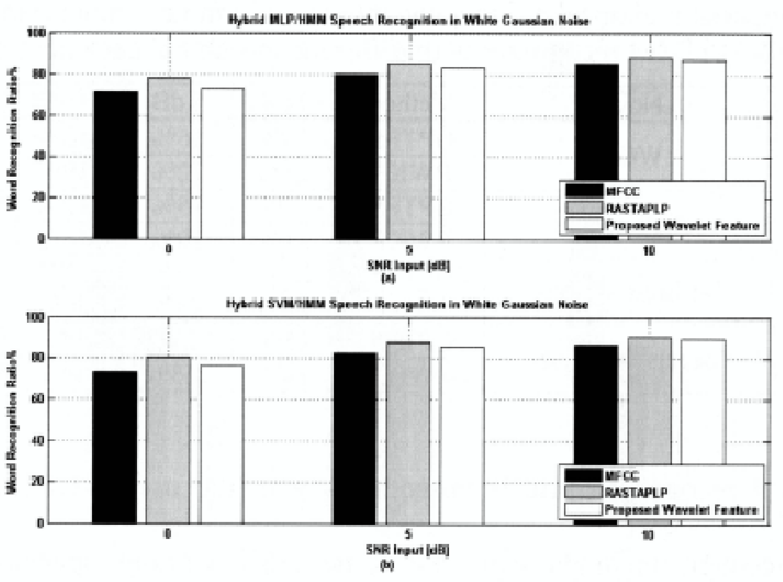


Figure 6.14 Performances of ASR using different features in additive white Gaussian noise: (a) Hybrid MLP/HMM (b) Hybrid SVM/HMM

To evaluate the robustness of ASR to denoised feature vectors, different features are used on the proposed ASR platform with additive Gaussian noise, colored noise, car engine noise and speech-like noise of varying SNRs deliberately introduced in the test speech. It was found that the recognition rate degrades more gracefully for wavelet coefficients than other coefficients. The results are charted in Figs. 6.14 to 6.17. From these figures, the proposed wavelet coefficients yield a more robust system than MFCC. RASTA+PLP has been conceived as the most complex but also the best performance ASR to date. Wavelet coefficients based feature extraction is inferior to it at low input SNR but is otherwise comparable at SNR above 10 dB. The wavelet coefficient has consistently good performance in various noise conditions for both systems, and hybrid SVM/HMM system has better recognition accuracy than MLP/HMM recognition system.

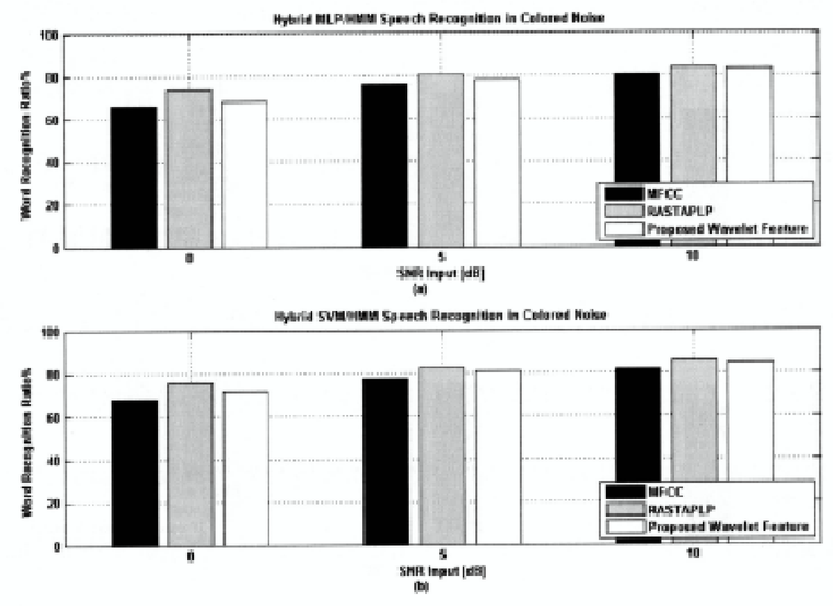


Figure 6.15 Performances of ASR using different features in additive colored noise: (a) Hybrid MLP/HMM (b) Hybrid SVM/HMM

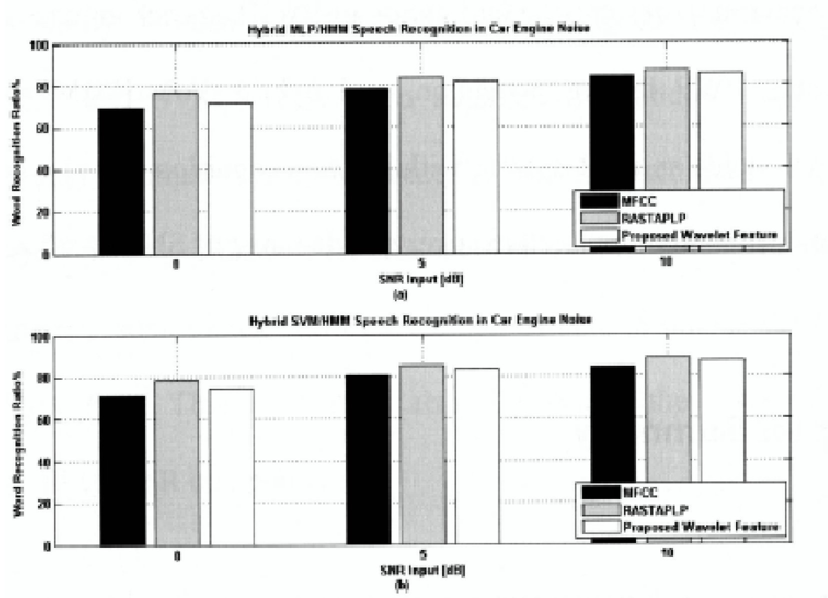


Figure 6.16 Performances of ASR using different features in additive car engine noise: (a) Hybrid MLP/HMM (b) Hybrid SVM/HMM

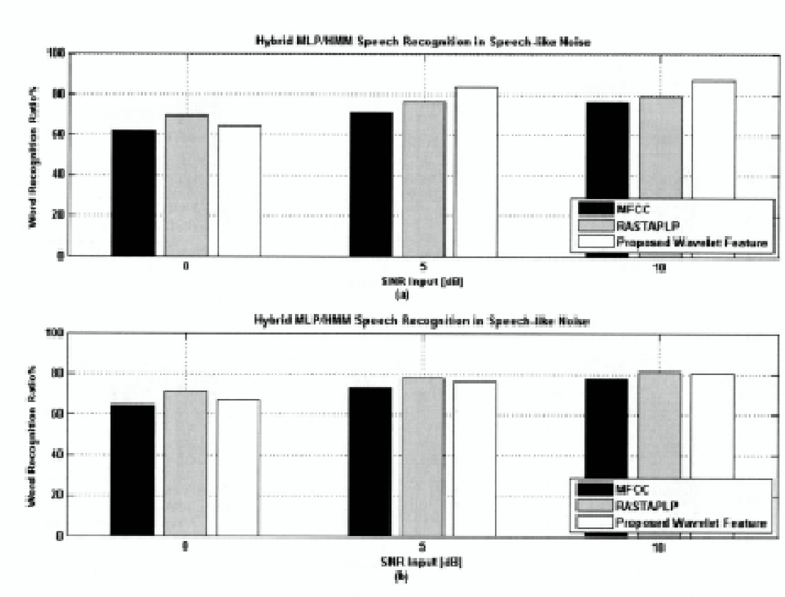


Figure 6.17 Performances of ASR using different features in speech-like noise: (a) Hybrid MLP/HMM (b) Hybrid SVM/HMM

From these experiment results, the unification of PWKF speech enhancement for front-end processing, PWF feature extraction, and hybrid SVM/HMM recognizer with wavelet kernel yields an ASR system with better recognition accuracy and robustness than conventional recognition systems over a wide range of SNR conditions.

6.5 Chapter Summary

A comprehensive speech processing and analysis toolbox has been developed to ease the analysis and evaluation of speech enhancement applications. The toolbox provides many important functions for digitized speech signal analysis and processing. It is portable to any machine that runs Matlab program and can directly interface to other Matlab codes. The convenient graphical user interface makes the analyses easier and more efficient. The prowess of this toolbox is demonstrated by using its functions to investigate the effect of front-end speech enhancement processing on recognition task, compares the robustness of wavelet feature vectors against other competitive features, and explores hybrid recognition systems based on HMM model. The proposed configurations of speech enhancement, feature extraction, and hybrid recognizer in a unified wavelet domain to improve the performance of ASR system are motivated by two factors. One is to bridge the phonetic gap between the low-level data and the high-level model by considering the perceptual properties of human auditory system. The other aim is to reduce the mismatch between the laboratory training and practical testing conditions for speech recognition. The front-end speech enhancement, GPTFS and PWKF methods incorporated into the proposed ASR platform, have shown to have positive impact on harnessing the recognition accuracy. The suggested wavelet feature extractors have also enhanced the robustness of the ASR system by lowering the classifier's sensitivity to mismatched training and testing conditions. Discriminative approaches, MLP and SVM, are used to augment the hidden Markov model by adjusting the model parameters to compensate for the mismatch during the recognition stage. With appropriate optimization criteria, such as minimum classification error, the hybrid recognition system yields performance comparable to one of the best previous systems, RASTA+PLP. This evaluation also illuminates the potential of optimal configurations of key ASR components.

CHAPTER 7.

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

With the popularity of wireless communication and hand-free car kit, speech enhancement has been gaining considerable attention in recent years. Over the last three decades, a flurry of interdisciplinary works that bridged the applied mathematics, psychoacoustic and electronic engineering communities have been proposed to address the fundamental issues of speech denoising problem. The emphasis of this research is on the reduction of the effect of residual noise and speech distortion in the denoising process, and the enhancement of the denoised speech to improve its intelligibility. This thesis has presented several new insights into versatile speech enhancement methods and its associated applications by improving existing noise reduction and speech enhancement methods, such as subtractive-type method and adaptive filtering method.

By exploiting the subtractive-type method, the physiology of human auditory system, and wavelet technique, a new generalized perceptual time-frequency subtraction speech enhancement algorithm has been proposed to recover high quality speech from noise contaminated speech. The system consists of two functional stages working cooperatively to perform perceptual time-frequency subtraction by adapting the weights of the perceptual wavelet coefficients. The noisy speech is first decomposed into critical bands by perceptual wavelet transform. A temporal and spectral model of human psychoacoustics is developed to calculate the masking threshold to be applied to the GPTFS for noise reduction. Different spectral resolutions of the wavelet representation preserve the energy of the critical transient components so that the background noises, distortion, and residual noise can be attenuated adaptively according to the masking threshold and the signal to noise ratio. The unvoiced speech is also enhanced by a soft-thresholding scheme. The effectiveness of the proposed system to extract a clear and intelligible speech from various noisy environments has been demonstrated through both objective and subjective measurements. The performance robustness under varying signal-to-noise conditions of the proposed algorithm is attributable to the use of both the temporal and simultaneous maskings for the tuning of subtraction parameters. Together with the unvoiced speech enhancement, the proposed system makes an average SNR improvement of 5.5% over [HU04] by objective measurements, and an average intelligibility improvement of 8% by subjective evaluation over other well-known methods under different noisy conditions and various SNR levels.

This thesis also dwells on another effective speech enhancement technique that bridges the multi-resolution discrete wavelet packet transform and the model based Kalman filter. Wavelet packet transform provides a flexible time-frequency excision of speech features whereas Kalman filter makes an optimal estimator of signal in noise. A perceptual wavelet filter bank is realized by a six-level dyadic subband tree decomposition to approximate the critical band of human auditory system. The compaction of the energy of the critical transient components into different spectral resolutions in wavelet representation allows the background noise and distortion to be efficiently excised by the Kalman filter. The state-space equations of the proposed wavelet Kalman filter is formulated based on a subband voiced-unvoiced speech model. The residue noise of the clean speech estimated by the wavelet Kalman filter is further attenuated by a perceptually weighted filter. Comparing with other speech enhancement methods, the effectiveness of the proposed wavelet Kalman filter to extract a clear and intelligible speech from adverse environments is largely due to the incorporation of both the statistical and perceptual properties into the wavelet transform excision process for optimal estimation of clean speech.

In this research work, a speech processing and analysis toolbox for Matlab program has been developed. This software tool provides a simple and useful graphical user interface for prototyping speech processing algorithms. Besides the routines required to perform the basic operations in digital speech analysis and modification, it also provides many versatile speech enhancement functions including the subtractive-type methods, wavelet denoising methods and adaptive filtering methods.

With the help of the speech analysis and processing toolbox, we can easily investigate the effect of front-end speech enhancement processing on recognition task, compares the robustness of wavelet feature vectors against other competitive features, and explores hybrid recognition systems based on hidden Markov model. Different configurations of speech enhancement, feature extraction and hybrid recognizer in a unified wavelet domain have been evaluated and have shown to improve the performance of automatic speech recognition system. The mismatch between the laboratory training and practical testing conditions for speech recognition has been significantly reduced by the incorporation of the proposed generalized perceptual time-frequency subtraction method and perceptual wavelet Kalman filtering speech enhancement method speech enhancement methods into the proposed ASR platform. The suggested wavelet feature extractors have also helped to lower the classifier's sensitivity to mismatched training and testing conditions. The benefit of a hybrid recognizer has also been corroborated by the experimental results. It has been demonstrated that the drawback of hidden Markov model can be mitigated by the discriminative power of multilayer perceptron and support vector machine. The best performance is obtained by the proposed hybrid support vector machines and hidden Markov model with wavelet kernel with a word recognition ratio of 96.53%. With appropriate optimization criteria, such as minimum classification error, the hybrid recognition system yields performance comparable to one of the best previous systems, RASTA+PLP. This evaluation indicates the promise of an optimal configuration of ASR components.

In summary, the objectives set forth in this thesis on the design of versatile speech enhancement methods capable of producing high intelligibility in various noisy conditions have been met. These humble contributions shall pave the way to the advancement of speech enhancement and hence, broaden the applications of digital speech processing.

7.2 Recommendations for Further Work

As usual, no research will be completed, since a new discovery naturally triggers the pursuit of the new frontier and dimension it projected. Through the research conducted in this thesis, several relevant topics and directions worthy of further exploration have been identified.

The analysis–synthesis systems based on maximally decimated filter banks have emerged as one of the important techniques for wideband speech and audio coding. In Chapter 3, the analysis–synthesis wavelet filter bank has been deduced to model the human auditory system, where the critical band model of aural perception is reflected in the design of the filter banks in both the time and frequency domains. The orthogonal wavelet packet transforms have been designed by a hierarchical association of perfect reconstruction paraunitary filter banks, which led naturally to the tree structures. One shortcoming of this design is the long filter length. To make the implementation of this wavelet packet transform fast and efficient, a lattice structured paraunitary filter bank can be adopted [DRY93], [RIO94], [DRY96]. In general, the existing methods generalize the fast STFT approach by offering multiframe and multiwindow processing algorithms. The main characteristic of a lattice structure implementation is the paraunitary filter bank is composed of cascading orthogonal operators, delays, downsamplers (for analysis filters) and upsamplers (for synthesis filters). This lattice structure has a hierarchical property, i.e., higher order paraunitary filter banks can be obtained from those of the lower order by adding more lattice sections. The lattice structure implements two-channel maximally downsampled paraunitary FIR filter banks of even length. The filter banks have perfect reconstruction property and are based on in-place butterfly operations. Butterfly operations can be characterized by lattice rotation and delay parameters. By changing the rotation parameters, all paraunitary filter banks can be generated. Further research is necessary to work out the optimal parameters. These parameters are dependent on the desired characteristics of the prototype lowpass filter. Optimization procedures can be explored to obtain a variety of orthogonal prototype filters with balanced regularity, frequency selectivity, number of orthogonal operators and phase. To investigate an elegant and efficient way to implement multiresolution auditory model, the existing lattice-based method could be enhanced by deriving the wavelet packet (WP) tree structure according to perceptual frequency criterion. We can select a WP tree that best matches the signal's time-spectral characteristics. Different types of filters, such as Daubechies filters, Malvar MLT and ELT filters, as well as those developed for orthogonal wavelet transforms and other paraunitary filter sets, could be implemented in this way to evaluate which one could be used to better approximate the human auditory system.

An EM-based iterative algorithm has been used for the proposed Kalman filtering approach in this thesis. The speech and colored noise are modelled as white Gaussian noise excited autoregressive process. The experiment results show that the proposed Kalman filter outperforms the Wiener filter and white noise Kalman filter. However, Niedzwiecki indicates that EM-based algorithms do not work well for the case of impulsive disturbances [NIE96]. To solve this problem, an extended Kalman filter (EKF) can be considered. In an EKF, there are two Kalman filters. One Kalman filter is used for tracking the parameter changes in the speech model rather than estimating the parameters iteratively. The other Kalman filter is designed to estimate the clean speech from the noisy observation. The computation load of the system can be reduced considerably if a fast Kalman tracking algorithm can be developed. This proposal of future work on EKF can also be considered for non-stationary noise, which is a very important issue for single channel speech enhancement problem.

The third potential research direction is to upgrade the proposed hybrid ASR platform for large vocabulary continuous speech recognition application (LVCSR), and to evaluate and demonstrate its improved error rate performance on large vocabulary conversational speech task. However, a related problem of LVCSR is the variable segment length or duration problem. Segment durations are correlated with the word choice and speaking rate. The motivating factor for most segment-based approaches is that the acoustic model needs to capture both the temporal and spectral structure of speech that is clearly missing in frame-level classification schemes. Segmental approaches overcome the assumption of conditional independence between frames of data in traditional HMM systems. In this study, Segmental minimum risk Bayesian decoding for automatic speech recognition can be considered [GOE04]. Segmental data takes better advantage of the correlation in adjacent frames of speech data. Currently, this framework allows us to decompose an utterance level minimum Bayes-risk recognizer into a sequence of smaller sub-utterance recognizers. Therefore, a large search problem is decomposed into a sequence of simpler, independent search problems. Any progress in this advanced segmentation will enable the hybrid ASR system to efficiently and significantly reduce the word recognition error.

AUTHOR'S PUBLICATIONS

Journal Papers:

[1] **Yu Shao** and Chip-Hong Chang, "A generalized time-frequency subtraction method for robust speech enhancement based on wavelet filter bank modeling of human auditory system," *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*. vol. 37, no. 4, pp. 877- 889, Aug. 2007.

[2] **Yu Shao** and Chip-Hong Chang, "Evaluation of Hybrid Speech Recognition Systems with De-noised Wavelet Features," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Applications.* (under review).

Conference Papers:

[1] **Yu Shao** and Chip-Hong Chang, "A novel hybrid neuro-wavelet system for robust speech recognition," *Proc. 2006 IEEE International Symposium on Circuits and Systems* (ISCAS' 06), Kos, Greece, pp. 1852-1855, May 21-24, 2006.

[2] **Yu Shao** and Chip-Hong Chang, "A Kalman filter based on wavelet filter-bank and psychoacoustic modeling for speech enhancement," *Proc. 2006 IEEE International Symposium on Circuits and Systems (ISCAS' 06)*, Kos, Greece, pp. 121-124, May 21-24, 2006.

[3] **Yu Shao** and Chip-Hong Chang, "A generalized perceptual time-frequency subtraction method for speech enhancement," *Proc. 2006 IEEE International Symposium on Circuits and Systems (ISCAS' 06)*, Kos, Greece, pp. 2537-2540, May 21-24, 2006.

[4] **Yu Shao** and Chip-Hong Chang, "A versatile speech enhancement system based on perceptual wavelet denoising," *Proc. 2005 IEEE International Symposium on Circuits and Systems (ISCAS' 05),* Kobe, Japan, vol: 2, pp: 864 – 867, May 23-26, 2005.

[5] **Yu Shao** and Chip-Hong Chang, "Wavelet transform to hybrid support vector machine and hidden Markov model for speech recognition," *Proc. 2005 IEEE International Symposium on Circuits and Systems (ISCAS' 05)*, Kobe, Japan, vol: 4, pp: 3833 – 3836, May 23-26, 2005.

[6] **Shao Yu**, Tong Yit Chow and Wang Chao, "Wavelet statistical model of speech for feature extraction and denoising," *IEEE Asia-Pacific Conference on Circuits and Systems (APCCAS' 04)*, Tainan, Taiwan, vol: 1, pp: 189 – 192, Dec., 6-9, 2004.

[7] Chao Wang, Yit-Chow Tong and **Yu Shao**, "VLSI design and analysis of a critical-band transform processor for speech recognition," *IEEE International System-on- Chip Conference (SOCC' 04)*, Hilton Santa Clara, California, pp:365 – 368, Sep., 12-15, 2004.

BIBLIOGRAPHY

[ABE05] S. Abe, *Support Vector Machines for Pattern Classification.* New York: Springer, 2005.

[AAR04] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Syst., Man, Cybern. B*, vol. 34, no. 4, pp. 1763 – 1773, Aug. 2004.

[AHM89] M. S. Ahmed, "Comparison of noisy speech enhancement algorithms in terms of LPC perturbation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 1, pp. 121 – 125, Jan. 1989.

[AYA04] S. Ayat, M. T. Manzuri and R. Dianat, "Wavelet based speech enhancement using a new thresholding algorithm," *In Proc. 2004 Int. Symp. on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, pp. 238 – 241, 20-22 Oct. 2004.

[BAH01] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *IEEE Signal Processing Lett.*, vol. 8, no. 1, pp. 10 – 12, Jan. 2001.

[BEN93] J. J. Benedetto and A. Teolist, "A wavelet auditory model and data compression," *Appl Comput Harmon Analysis 1*, in press, 1993.

[BER79] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, (ICASSP '79),* vol. 4, pp. 208 – 211, Apr. 1979.

[BOL79] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113 – 120, Apr. 1979.

[BOU94] H. A. Bourland and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer, Boston, 1994

[BRE07] C. Breithaupt, T. Gerkmann and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Processing Lett.*, vol. 14, no. 12, pp. 1036 – 1039, Dec. 2007.

[BRO] M. Brookes, VOICEBOX: Speech Processing Toolbox for MATLAB, [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[BRO02] J. Brocker, U. Parlitz and M. Ogorzalek, "Nonlinear noise reduction," *Proc. IEEE*, vol. 90, no. 5, pp. 898 – 918, May 2002.

[BUR98] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Knowledge Discovery Data Mining*, vol. 2, no. 2, pp. 121-167, 1998.

[CHA95] D. M. Chabries, D. V. Anderson, T. G. Stockham and R. W. Christiansen, "Application of a human auditory model to loudness perception and hearing compensation," in *IEEE Int. Conf. Acoust. Speech and Signal Proc.*, pp. 3527–3530, 1995.

[CHE91] Y. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Processing*, vol.39, no. 9, pp. 1943-1954, Sept. 1991.

[CHE05] K. Chen, "On the use of different speech representations for speaker modeling," *IEEE Trans. Syst., Man, Cybern. C*, vol. 35, no. 3, pp. 301 – 314, Aug. 2005.

[CHO07] G.F. Choueiter and J.R. Glass, "An Implementation of Rational Wavelets and Filter Design for Phonetic Classification," *IEEE Trans. Speech Audio Processing*, vol. 15, no. 3, pp. 939 – 948, Mar. 2007.

[COI92] R. R. Coifman and M.V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp: 713 -718, Mar. 1992.

[COU00] L. Couvreur and C. Couvreur, "Wavelet-based method for nonparametric estimation of HMMs," *IEEE Signal Processing Lett*, vol. 7, no. 2, pp. 25 – 27, Feb. 2000.

[CRI00] N. Cristianini, and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[DAV80] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 4, pp. 357 – 366, Aug 1980.

[DEM77] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Roy. Stat. Soc. Ser.*, vol. 39, pp. 1-38, 1977.

[DEL99] J. Deller, J. Proakis and J. Hansen, *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[DON95] D.L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 613 – 627, May 1995.

[DRY93] A. Drygajlo, "Fast orthogonal transform algorithms for multiresolution time-sequency signal decomposition and processing," in *Proc. SPIE Math. Imaging: Wavelet Applicat. Signal Process. Image Process.*, San

Diego, CA, vol. 2034, pp. 349–358, July 1993.

[DRY96] A. Drygajlo, "New fast wavelet packet transform algorithms for frame synchronized speech processing ," *In Proceedings Fourth International Conference Spoken Language, 1996, (ICSLP 96)*, Philadelphia, PA, USA, 3-6 Oct 1996.

[DUD01] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification (2nd ed.)*. John Wiley and Sons, New York, 2001.

[EPH84] Y. Ephraim and D. Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109-1121, Dec. 1984.

[EPH92a] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526 – 1555, Oct. 1992.

[EPH92b] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 40, no. 4, pp. 725 – 735, April 1992.

[EPH95] Y. Ephraim and H.L. Van Trees, "A signal subspace approach for speech enhancement", *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp.251 – 266, July 1995.

[FAR04] O. Farooq and S. Datta, "Wavelet based robust sub-band features for phoneme recognition," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 151, no. 3, pp. 187 – 193, 1 June 2004.

[FED89] M. Feder, A.V. Oppenheim and E. Weinstein, "Maximum likelihood noise cancellation using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 2, pp. 204 – 216, Feb. 1989.

[FLE40] H. Fletcher, "Auditory Patterns," *Review of Modern Physics*, vol.12, pp. 47-65, 1940.

[GAB99] M. Gabrea, E. Grivel and M. Najun, "A single microphone Kalman filter-based noise canceller," *IEEE Signal Processing Lett.*, vol. 6, no. 3, pp. 55 – 57, Mar. 1999.

[GAB01] M. Gabrea, "Adaptive Kalman filtering-based speech enhancement algorithm," *in Proc. Canadian Conf. Electrical and Computer Engineering*, Fredericton, AB, Canada, vol. 1, pp. 521-526, 2001.

[GAN98] S. Gannot, D. Burshtein and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 4, pp.373 – 385, July 1998

[GIB91] J. D. Gibson, B. Koo and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, no. 8, pp. 1732 – 1742, Aug. 1991.

[GOE04] V. Goel, S. Kumar and W. Byrne, "Segmental minimum Bayes-risk decoding for automatic speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 3, pp. 234 – 249, May 2004.

[GOH98] Z. Goh, K.Tan and T. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech Audio Processing,* vol. 6, pp. 287-292, May 1998.

[GOH99] Z. Goh, K. C. Tan and B.T.G.Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 5, pp.510 – 524, Sept. 1999.

[GOL94] M. H. Goldstein, "Auditory periphery as speech signal processor," *IEEE Eng. Med. Biol.*, pp. 186–196, Apr. 1994.

[GRA00] D. Graupe and D. Veselinovic, "Blind adaptive filtering of speech from noise of unknown spectrum using a virtual feedback configuration," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 2, pp. 146 -158, Mar. 2000.

[GRA06] V. Grancharov, J. Samuelsson and B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Trans. Audio, Speech, Language Processing*, vol.14, no. 3, pp. 764 – 773, May 2006.

[HAG96] M. T. Hagan, H. B. Demuth and M. H. Beale, *Neural Network Design*, PWS Publishing Co., Boston, MA, USA, 1996.

[HAN06] J. H. L. Hansen, V. Radhakrishnan and K. H. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 6, pp. 2049 – 2063, Nov. 2006.

[HAS04] M. K. Hasan, S. Salahuddin and M.R. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Processing Lett.*, vol. 11, no. 4, pp. 450 – 453, Apr. 2004.

[HAY01] S. Haykin, *Adaptive Filter Theory*. New Jersey: Prentice Hall, 2001.

[HE99] C. He and G. Zweig, "Adaptive two-band spectral subtraction with multi-window spectral estimation," *ICASSP*, vol.2, pp. 793-796, 1999.

[HER94] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 578 -589, Oct. 1994.

[HOY05] T. Hoya, T. Tanaka, A. Cichocki, T. Murakami, G. Hori and J.A. Chambers, "Stereophonic noise reduction using a combined sliding subspace projection and adaptive signal enhancement," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, pp. 309 – 320, May 2005.

[HU02] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," *IEEE Signal Processing Letters*, vol. 9, no. 7, pp. 204 – 206, Jul. 2002.

[HU03] Y. Hu and P. C. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 457 – 465, Sept. 2003.

[HU04] Y. Hu and P. C. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Processing Lett.*, vol. 11, no. 2, pp. 270 -273, Feb. 2004.

[ITU01] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommend*, p.862, Feb. 2001.

[JAB03] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 700 – 708, Nov. 2003.

[JEA01] W. L. B. Jeannes, P. Scalart, G. Faucon and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: a survey," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 8, pp. 808 – 820, Nov. 2001.

[JEO01] S. Jeong, and M. Hahn, "Speech quality and recognition rate improvement in car noise environments," *Electronics Letters*, vol. 37, no. 12, pp. 800 – 802, 7 Jun 2001.

[JOH88] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, no. 2, pp. 314 – 323, Feb. 1988.

[JUA92] B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.

[JU07] G.-H. Ju and L.-S. Lee, "A Perceptually Constrained GSVD-Based Approach for Enhancing Speech Corrupted by Colored Noise," *IEEE Trans. on Audio, Speech, Language Processing*, vol. 15, no. 1, pp. 119 – 134, Jan. 2007.

[KAH97] M. Kahrs, G.W. Elko, S.J Elliot, S. Makino, J.M. Kates, M. Bosi and J.O. Smith, "The past, present and future of audio signal processing," *IEEE Signal Processing Mag.*, vol. 14, no. 5, pp. 30 – 57, Sep 1997.

[KAM02] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, (ICASSP '02)*, vol. 4, pp .IV-4164, 13-17 May 2002.

[KAR01] M. Karnjanadecha and S.A. Zahorian, "Signal modeling for high-performance robust isolated word recognition," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 6, pp. 647 – 654, Sept. 2001.

[KAR06] A. Karmakar, A. Kumar and R.K. Patney, "A Multiresolution Model of Auditory Excitation Pattern and Its Application to Objective Evaluation of Perceived Speech Quality," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 6, pp. 1912 – 1923, Nov. 2006.

[KAR07] A. Karmakar, A. Kumar and R.K. Patney, "Design of Optimal Wavelet Packet Trees Based on Auditory Perception Criterion," *IEEE Signal Processing Lett.*, vol. 14, no. 4, pp. 240 – 243 Apr. 2007.

[KIM00] W. Kim, S. Kang and H. Ko, "Spectral subtraction based on phonetic dependency and masking effects," *Proc. IEE Vis. Image Signal Procs.*, vol. 147, no. 5, pp. 423 – 427, Oct. 2000.

[LAI99] P. C. Laizou, "Signal-processing techniques for cochlear implants," *IEEE Eng. Med. Biol. Mag.*, vol. 18, no. 3, pp. 34 – 46, May-June 1999.

[LEE96] K. Y. Lee and K. Shirai, "Efficient recursive estimation for speech enhancement in colored noise," *IEEE Signal Processing Lett.,* vol. 3, no. 7, pp.196 – 199, Jul. 1996.

[LEE97] K. Y. Lee, B. G. Lee and S. Ann, "Adaptive filtering for speech enhancement in colored noise," *IEEE Signal Processing Lett.*, vol. 4, no. 10, pp. 277 – 279, Oct. 1997.

[LEE00] K. Y. Lee and S. Jung, "Time-domain approach using multiple Kalman filters and EM algorithm to speech enhancement with nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 282 – 291, May 2000.

[LI01] M. Li, H.G. McAllister, N.D. Black and T.A. De Perez, "Perceptual time-frequency subtraction algorithm for noise reduction in hearing aids," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 9, pp. 979 – 988, Sept. 2001.

[LIM78] J. Lim and A.Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 26, no. 3, pp. 197 – 210, Jun. 1978.

[LIM79] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[LIN03] C. T. Lin, "Single-channel speech enhancement in variable noise-level environment," *IEEE Trans. Syst., Man, Cybern. A*, vol. 33, no. 1, pp. 137 – 143, Jan. 2003.

[LIN05] Y. L. Lin and G. Wei; "Speech emotion recognition based on HMM and SVM," *Proc. IEEE Int. Conference on Machine Learning and Cybernetics 2005*, vol. 8, pp. 4898 – 4901, 18-21 Aug. 2005.

[LIU92] W. Liu, "An analog cochlear model: signal representation and VLSI realization." *Ph.D. Dissertation*, Johns Hopkins University, 1992.

[LIU06] H. Liu, Q. Zhao, M. Wan and S. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 5, pp. 865 – 874, May 2006.

[LOC92] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2-3, pp. 215-228, 1992.

[LOI98] P. C. Loizou, "Mimicking the human ear," *IEEE Signal Processing Mag.*, vol. 15, no. 5, pp. 101 – 130, Sept. 1998.

[LOI99] P. Loizou, COLEA: A Matlab Software Tool for Speech Analysis, [Online]. Available: http://www.utdallas.edu/~loizou/speech/colea.htm, 1999.

[LOW91] J. M. Lowerre, "Time domain use of the EM algorithm in noise cancellation," *IEEE Trans. Speech Audio Processing*, vol. 39, no. 4, pp. 986 – 989, Apr. 1991.

[LU04] C.-T. Lu and H.-C. Wang, "Speech enhancement using perceptually-constrained gain factors in critical-band-wavelet-packet transform," *Electronics Letters*, vol. 40, no. 6, pp. 394 – 396, 18 Mar. 2004.

[MA06] N. Ma, M. Bouchard and R. A. Goubran, "Speech enhancement using a masking threshold constrained Kalman filter and its heuristic implementations," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 1, pp. 19 – 32, Jan. 2006.

[MAL99] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego: Academic Press, 1999.

[MAO00] J. S. Mao, S. C. Chan, W. Liu and K. L. Ho, "Design and multiplier-less implementation of a class of two-channel PR FIR filterbanks and wavelets with low system delay," *IEEE Trans. Signal Processing,* vol. 48, no. 12, pp. 3379 – 3394, Dec. 2000.

[MAR98] J. A. Martins and F. Violaro, "Comparison of parametric representations for hidden Markov models and multilayer perceptron recognizers," *Proc. SBT/IEEE Int. Telecom. Symp.*, Sao Paulo, Brazil, vol. 1, pp. 141 – 145, Aug. 1998.

[MAR02] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 2, pp. 109 – 118, Feb. 2002.

[MCA80] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 2, pp. 137 – 145, Apr. 1980.

[MER99] A. Mertins, *Signal Analysis: Wavelet, Filter Banks, Time-Frequency Transforms and Applications,* Chichester, John Wiley & Sons, 1999.

[MIT00] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Processing*, vol.8, no.2, pp.159 – 167, Mar. 2000.

[MOL73] A. R. Moller, Ed., *The frequency selectivity of the cochlea*, London, Academic, 1973.

[MOO97] S. Moon and J. N. Hwang, ”Robust speech recognition based on joint model and feature space optimization of hidden Markov models,” *IEEE Trans. Neural Networks*, vol. 8 , no. 2, pp.194 – 204, Mar. 1997.

[MOU07] A. Mouchtaris, J. Van der Spiegel, P. Mueller and P. Tsakalides, “A Spectral Conversion Approach to Single-Channel Speech Enhancement,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 15, no. 4, pp. 1180 – 1193, May 2007.

[NAN95] S. Nandkumar and J. H. L. Hansen, “Dual-channel iterative speech enhancement with constraints on an auditory-based spectrum,” *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 22 – 34, Jan. 1995.

[NIE96] M. Niedzwiecki and K. Cisowski, “Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals,” *IEEE Trans. Signal Processing*, vol. 44, no. 3, pp. 528 – 537, Mar. 1996.

[NOC85] N. Nocerino, F. Soong, L. Rabiner and D. Klatt, “Comparative study of several distortion measures for speech recognition” *In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, (ICASSP '85*), vol. 10, pp. 25 -28, Apr. 1985.

[NOI92] NOISEX-92, “Speech and noise data base,” *NATO AC243-panel 3/RSG.10*, 1992.

[ORT05] A. Ortega, E. Lleida and E. Masgrau, “Speech reinforcement system for car cabin communications,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, Part 2, pp. 917 – 929, Sept. 2005.

[P.862] ITU-T P.862 Version 1.1 software, ITU-T Rec. P.862, 02/2001. availabe online: http://eu.sabotage.org/www/ITU/P/P862%20Software/, [PAL87] K. Paliwal, and A. Basu, "A speech enhancement method based on Kalman filtering," *In Proc. IEEE Int. Confe. on Acoustics, Speech, and Signal Processing, (ICASSP '87)*, vol. 12, pp.177 – 180, Apr 1987.

[PLA99] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," *Advances in Large Margin Classiers*. MIT Press, Cambridge, 1999.

[PAR02] S. J. Park, C. G. Cho, C. Lee and D. H. Youn, "Integrated echo and noise canceler for hands-free applications," *IEEE Trans. Circuits Syst. II*, vol. 49, no. 3, pp. 188 – 195, Mar. 2002.

[PET81] T. Peterson and S. Boll, "Acoustic noise suppression in the context of a perceptual model", *Proc. IEEE Inter. Conf. Acoust. Speech Signal Procs.*, pp. 1086-1088, 1981.

[PET83] T. Petersen and S. Boll, "Critical band analysis-synthesis," *IEEE Trans. Signal Processing*, vol. 31, no. 3, pp. 656 – 663, Jun. 1983.

[PLA99] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," *Advances in Large Margin Classiers*. MIT Press, Cambridge, 1999.

[POP98] D. C. Popescu and I. Zeljkovic, "Kalman filtering of colored noise for speech enhancement," *In Proc. IEEE Int. Confe. on Acoustics, Speech, and Signal Processing, (ICASSP '98)*, vol. 2, pp. 997 - 1000, 12-15 May 1998.

[PUJ05] P. Pujol, S. Pol, C. Nadeu, A. Hagen and H. Bourlard, "Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, pp. 14-22, Jan. 2005.

[QUA02] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*, Prentice Hall PTR, NJ, 2002

[RAB89] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[RAM06] J. Ramirez, P. Yelamos, J. M. Gorriz and J.C. Segura, "SVM-based speech endpoint detection using contextual speech features," *Electronics Lett.*, vol. 42, no. 7, pp. 426 – 428, 30 March 2006.

[REZ01] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 87-95, Feb. 2001.

[RIO94] O. Rioul and P. Duhamel, "A Remez exchange algorithm for orthonormal wavelets," *IEEE Trans. Circuits Syst. II*, vol. 41, pp. 550–560, Aug. 1994.

[SAI95] N. Saito and R. R. Coifman, "Local discriminant bases and their applications," *J. Math. Imag. Vision*, vol. 5, pp. 337-358, 1995.

[SAM98] H. Sameti, H. Sheikhzadeh, L. Deng and R.L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 5, pp. 445-455, Sept. 1998.

[SCA96] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP-96)*, vol. 2, pp. 629-632, 7-10 May 1996.

[SCH75] M. Schroeder, “Models of hearing,” *Proc. IEEE,* vol. 63, No. 9, pp. 1332-1350, Sept. 1975.

[SCH79] M. R. Schroeder, B. S. Atal and J. L. Hall, “Optimizing digital speech coders by exploiting masking properties of the human ear.” *J. Acoust. Soc. Amer.*, vol. 66, no. 6, pp. 1647-1652, Dec. 1979.

[SHA04] Y. Shao, Y. C. Tong and C. Wang, “Wavelet statistical model of speech for feature extraction and denoising,” *In Proc. IEEE Asia-Pacific Symp. on Circuits and Systems (APCCAS' 04)*, Tainan, Taiwan, vol. 1, pp. 189-192, Dec., 6-9, 2004.

[SHA05a] Y. Shao and C. H. Chang, "A versatile speech enhancement system based on perceptual wavelet denoising", *In Proc. IEEE Int. Symp. on Circuits and Systems, (ISCAS’05)*, Kobe, Japan, vol. 2, pp. 864-867, May 23-26, 2005.

[SHA05b] Y. Shao and C. H. Chang, "Wavelet transform to hybrid support vector machine and hidden Markov model for speech recognition", *In Proc. IEEE Int. Symp. on Circuits and Systems, (ISCAS’05)*, Kobe, Japan, vol. 4, pp. 3833 – 3836, May 23-26, 2005.

[SHA06a] Y. Shao and C. H. Chang, “A novel hybrid neuro-wavelet system for robust speech recognition,” *In Proc. IEEE Int. Symp. on Circuits and Systems, (ISCAS’06)*, Greece, May 21-24, 2006.

[SHA06b] Y. Shao and C. H. Chang, “A Kalman filter based on wavelet filter-bank and psychoacoustic modeling for speech enhancement,” *In Proc. IEEE Int. Symp. on Circuits and Systems, (ISCAS’06)*, Greece, May 21-24, 2006.

[SHA06c] Y. Shao and C. H. Chang, "A generalized perceptual time-frequency subtraction method for speech enhancement," *In Proc. IEEE Int. Symp. on Circuits and Systems, (ISCAS'06)*, Greece, May 21-24, 2006.

[SHA07a] Y. Shao and C. H. Chang, "A generalized time-frequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system," *IEEE Trans. Syst., Man, Cybern. B*, vol. 37, no. 4, pp. 877- 889, Aug. 2007.

[SHA07b] Y. Shao and C. H. Chang, "Evaluation of hybrid speech recognition system with de-noised wavelet features," *IEEE Trans. Syst., Man, Cybern. A*, 2007. (26th Dec. 2007 Submitted)

[SHY92] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Mag.*, pp. 15–37, Jan. 1992.

[SIM98] B. L. Sim, Y. C. Tong, J. S. Chang and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 4, pp. 328-337, July 1998.

[SIN93] D. Sinha and A.H. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. Signal Processing*, vol. 1, no. 12, pp. 3463-3479, Dec. 1993.

[SLA98] M. Slaney, Auditory Toolbox, [Online]. Available: http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/, 1998.

[SPA07] Y. Shao, Speech Processing and Analysis toolbox, internal report, 2007 (available on request).

[SPR07] A. Spriet, G. Rombouts, M. Moonen and J. Wouters, "Combined feedback and noise suppression in hearing aids," *IEEE Trans. Audio,* *Speech and Language Processing*, vol. 15, no. 6, pp. 1777-1790, Aug. 2007.

[SRE96] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 4, pp. 383-389, Sept. 1996.

[SRI98] P. Srinivasan and L.H. Jamieson, "High-quality audio compression using an adaptive wavelet packet decomposition and psychoacoustic modeling," *IEEE Trans. Signal Processing* vol. 46, no. 4, pp. 1085-1093, Apr. 1998.

[SRI06] S. Srinivasan, R. Aichner, W.B. Kleijn and W. Kellermann, "Multichannel Parametric Speech Enhancement," *IEEE Trans. Signal Processing*, vol. 13, no. 5, pp. 304-307, May 2006.

[STR96] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, MA, Wellesley, 1996.

[STR99] N. Ström, L. Hetherington, T. J. Hazen, E. Sandness and J. Glass, "Acoustic modeling improvements in a segment-based speech recognizer," in *Proc. IEEE ASR Workshop*, Keystone, CO, Dec. 1999.

[TAN97] O. Tanrikulu, B. Baykal, A. G. Constantinides and J. A. Chambers, "Residual echo signal in critically sampled subband acoustic echo cancellers based on IIR and FIR filter banks," *IEEE Trans. Signal Processing*, vol. 45, pp. 901-911, Apr. 1997.

[TER79] E. Terhardt, "Calculating virtual pitch," *Hearing Res.*, vol. 1, pp. 155-182, 1979.

[TIM90] Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1, DARPA-TIMIT, 1990.

[TRE03] E. Trentin and M. Gori, "Robust combination of neural networks and hidden Markov models for speech recognition," *IEEE Trans. Neural Networks*, vol. 14, no. 6, pp. 1519-1531, Nov. 2003.

[VAI87] P. Vaidyanathan, "Quadrature mirror filter banks, M-band extensions and perfect-reconstruction techniques," *IEEE ASSP Magazine*, vol. 4, no. 3, part 1, pp. 4-20, Jul 1987.

[VAI93] P.P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[VAP98] V. N. Vapnik, *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

[VAS00] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction.* 2nd Ed., John Wiley & Sons, Chichester, 2000.

[VER99] R. Vergin, D. O'Shaughnessy and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 5, pp. 525-532, Sept. 1999.

[VES03] D. Veselinovic and D. Graupe, "A wavelet transform approach to blind adaptive filtering of speech from unknown noises," *IEEE Trans. Circuits Syst. II*, vol. 50, no. 3, pp. 150-154, Mar. 2003.

[VET95] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[VIR99] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 2, pp.126 – 137, Mar. 1999.

[WAN04] C. Wang, Y. C. Tong and Y. Shao, "VLSI design and analysis of a critical-band transform processor for speech recognition," *IEEE*

International System on Chip Conference (SOCC' 04), Hilton Santa Clara, CA, pp.365 – 368, Sep., 12-15, 2004.

[WAT93] L. Watts, "Cochlear Mechanics: Analysis and Analog VLSI," *PhD Thesis*, California Institute of Technology, 1993.

[WES97] F.A. Westall, "Review of speech technologies for telecommunications," *Journal Electronics & Communication Engineering*, vol. 9, no. 5, pp.197 – 207, Oct. 1997.

[WOO00] K. Wooil, K. Sunmee and K. Hanseok, "Spectral subtraction based on phonetic dependency and masking effects," *IEE Proceedings-Vision, Image and Signal Processing,* vol. 147, no. 5, pp.:423 – 427, Oct. 2000.

[WU98] W. R. Wu and P. C. Chen, "Subband Kalman filtering for speech enhancement," *IEEE Trans. Circuits Syst. II*, vol. 45, no. 8, pp. 1072-1083, Aug. 1998.

[WU00] G. D. Wu and C. T. Lin, "A recurrent neural fuzzy network for word boundary detection in variable noise-level environments," *IEEE Trans. Syst., Man, Cybern. B*, vol. 31, pp. 84-97, Feb. 2000.

[WU01] K. Wu and P. Chen, "Efficient speech enhancement using spectral subtraction for car hands-free application," *International Conference on Consumer Electronics*, vol. 2, pp. 220-221, 2001.

[YAO01] J. Yao and Y.-T. Zhang, "Bionic wavelet transform: a new time-frequency method based on an auditory model," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 8, pp. 856-863, Aug. 2001.

[YAO02] J. Yao and Y.-T. Zhang, "The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 11, pp. 1299-1309, Nov. 2002.

[YOS00] W. A. Yost, *Fundamentals of Hearing: An Introduction*. San Diego: Academic Press, 2000.

[ZHA04] L. Zhang, W. Zhou and L. Jiao, "Wavelet support vector machine," *IEEE Trans. Syst., Man, Cybern. B*, vol. 34, no. 1, pp. 34-39, Feb. 2004.

[ZHE99] L. Zheng, Y. T. Zhang, F. S. Yang and D. T. Ye, "Synthesis and decomposition of transient-evoked otoacoustic emissions based on an active auditory model," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 9, pp. 1098-1106, Sep. 1999.

[ZHO05] W. Zhong, S. Li and H. M. Tai, "Signal subspace approach for narrowband noise reduction in speech," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 152, no. 6, pp. 800-805, 9 Dec. 2005.

[ZWI90] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models* Berlin, Germany: Springer-Verlag, 1990.