WAVELET PACKETS-BASED SPEECH ENHANCEMENT TECHNIQUES FOR DIGITAL HEARING AIDS

by
Jiming Yang
B.Eng., Ryerson University, Canada, 2003
B.Sc., Shenzhen University, P.R. China, 1999

A thesis
presented to Ryerson University
in partial fulfillment of the
requirement for the degree of
Master of Applied Science
in the Program of
Electrical and Computer Engineering.

Toronto, Ontario, Canada, 2006

Author's Declaration

I hereby declare that I am the sole author of this thesis.
I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.
Signature

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.
Signature

Abstract

Wavelet Packets-Based Speech Enhancement Techniques for Digital Hearing Aids

Jiming Yang
Master of Applied Science
Electrical and Computer Engineering Department, Ryerson University, 2006

Hearing-impaired listeners often have great difficulty understanding speech in a noisy background. The problem has motived the development of a new speech enhancement scheme with the goal of improving speech in noise perception for the hearing impaired listeners. In this thesis, a novel wavelet packet based noise reduction algorithm and hearing loss compensation are presented for a single microphone hearing aids application.

The noise reduction scheme utilizes noise masking threshold based suppression rule to remove additive noise. The perceptual noise suppression rule is optimized to achieve a balance between noise removal and speech distortion. Both objective and subjective evaluations have shown superior performance of the proposed technique in a good combination of low residual noise and low signal distortion.

The hearing loss compensation is realized by the wavelet-based loudness compression in each critical band. The compensated speech is guaranteed above hearing-impaired listener's threshold of hearing and with growth of loudness corrected in the dynamic range. Preference test among normal hearing person with simulated hearing loss has shown compensated speeches are favored in various conditions.

Acknowledgments

I am deeply grateful to my supervisor Dr. Sridhar Krishnan, for his able guidance and unconditional support through the years. As a professor and friend, he is always a true mentor and a role model. What I learned from him will benefit my whole life.

I would like to thank Dr. Karthikeyan Umapathy, for his valuable support in writing this thesis. He is always there when I need his help.

Contents

List of Figures

List of Tables

List of Abbreviations

AWGN - Additive White Gaussian Noise
CBW - Critical Band Width
CWT - Continuous Wavelet Transform
DMOS - Degradation Mean Opinion Score
DWT - Discrete Wavelet Transform
EM - Ephraim-Malah
ERB - Equivalent Rectangular Bandwidth
FFT - Fast Fourier Transform
FIR - Finite Impulse Response
HLF - Hearing Loss Function
IWPT - Inverse Wavelet Packet Transform
MAD - Median Absolute Deviation
MMSE - Minimum Mean Square Error
MOS - Mean Opinion Score
MRA - MultiResolution Analysis
NIR - Noise to Input Ratio
NMT - Noise Masking Threshold
PESQ - Perceptual Evaluation of Speech Quality
PTFS - Perceptual Time-Frequency Subtraction
PWPD - Perceptual Wavelet Packet Decomposition
PWS - Perceptual Wavelet Subtraction
SegSNR - Segmental Signal to Noise Ratio
SFM - Spectral Flatness Measure
SNR - Signal to Noise Ratio
SPL - Sound Pressure Level
SS - Spectral Subtraction
STFS - Short Time Fourier Transform
STSA - Short Time Spectral Amplitude
WPT - Wavelet Packet Transform
WT - Wavelet Thresholding

Chapter 1

Introduction

1.1 Overview

Speech communication has always been the most direct and interactive way to convey information between humans. Under the modern technology settings, speech signal is now communicated through radio, television, recorder, telephony, and the Internet. In all speech communication settings, the quality and intelligibility of speech signal is of high importance to exchange information efficiently and accurately.

In the real-world environment, the fidelity of most of the communication system is corrupted by interference, which often takes the form of additive background and channel noise, reverberation, or competing speech. Sufficiently high level of interference evokes communication difficulty for listeners. Particularly for listeners with hearing impairments, who generally have great difficulty understanding the speech in the presence of noise than normal hearing listeners. The difficulties are often experienced as tiresome and fatiguing with an increased effort to understand speech in noise.

Various speech enhancement algorithms have been proposed in large number of literatures. Much of the current effort in hearing aids research has been to develop new algorithms that can perform well with speech in noise issue. However, due to the random nature of the noise and the inherent complexities of the human speech, the accuracy and robustness of the speech enhancement systems still pose considerable challenges. The complexity and ease of implementation of the speech enhancement and the noise reduction algorithms are important criterion when targeting application in portable systems such as hearing aids and cellular phones. Noise reduction techniques usually have trade off between the amount of noise removed and speech distortions introduced during the speech processing, thereby limiting the performance of speech enhancement systems.

1.2 Fundamentals of Speech Production and Human Hearing

Speech production: Speech is produced by a cooperation of lungs, glottis (with vocal cords), and articulation tract (mouth and nose cavity) . Figure 1.1 shows a schematic view of the human speech production mechanism. As air is expelled from the lungs through the trachea, the tensed vocal cords within the larynx are caused to vibrate by the air flow. The air flow is chopped into quasi-periodic pulses which are then modulated in frequency in passing through the throat, the oral cavity, and possibly the nasal cavity. Depending on the positions of the various articulators (i.e., jaw, tongue, velum, lips, and mouth), different sounds are produced [1].

To put it in more simplified terms, 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 either must pass through a constriction in the vocal tract and thereby become turbulent, and producing unvoiced sounds. Or it can build up pressure behind a point of the total closure within the vocal tract, and when the closure is opened, the pressure is suddenly and abruptly release, causing a brief transient sound.

Human hearing: The human ear has three main subdivisions: the outer, middle, and inner ear. Figure 1.2 shows a simplified view of the human ear. The outer

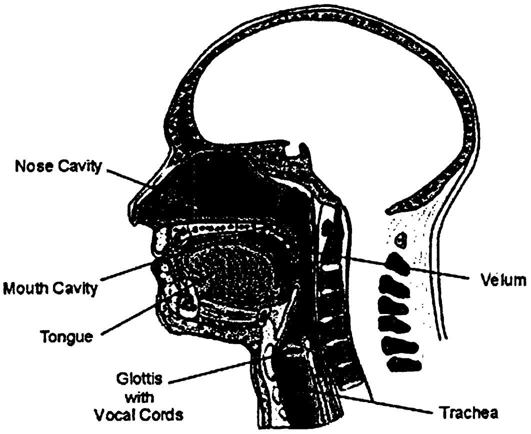


Figure 1.1: Human speech production mechanism [2]

ear consists of the pinna (visible part of the ear) and the meatus (auditory canal or ear canal). The pinna is mainly responsible for sound collecting, and aids in sound localization. The ear canal is a tube which directs the sound to the tympanic membrane (eardrum). It acts as a half closed tube resonator enhancing sounds in the range of 2-5 kHz. The middle ear consists of the eardrum and ossicles. The eardrum receives vibrations traveling from the auditory canal and transfers them through the ossicles to the oval window, which is the port to the inner ear. The ossicles act as lever, amplifying when soft sounds are received and attenuating to protect against very loud sound. The external and middle ears, which together form the conductive component of the auditory apparatus, transmit sound waves from the external environment to the sensory organ of hearing, the inner ear. As they transmit sound, they also amplify and modify the frequency spectrum. The inner ear consists of the semicircular canals that serve as the balance organ of the body and the cochlea that contains the basilar membrane and organ of Corti, which together form the complicated mechanisms that transduce vibrations into neural signal codes. During the process, the Corti performs a frequency-to-place mapping of the mechanical oscillations into electrical impulses

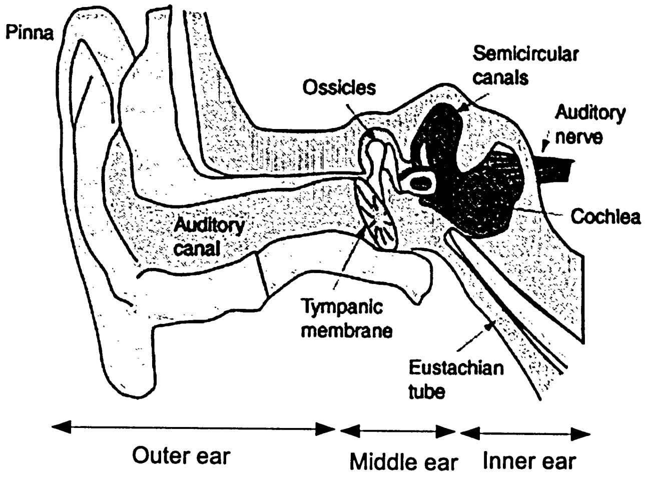


Figure 1.2: Simplified view of human ear [3]

that can be picked up by auditory nerve.

Psychoacoustics: Knowledge about the acoustic sound signals and the auditory system physiology alone is not sufficient to understand the human hearing. Ultimately only how human responded to sound that matters the most. Psychoacoustics studies the relationship between the subjective perception of audio and the scientific measurement of sound. The most important and fundamental principles of psychoacoustics are critical bands, threshold of hearing and masking phenomenon.

In order to be audible to human being, sounds require a minimum Sound Pressure Level (SPL). Often the level is determined in the absence of any external sound. The auditory threshold is the average among the minimum SPL values obtained from different people. This threshold is different from person to person and tends to increase as one ages. Figure 1.3 shows an average human auditory threshold as a function of frequency. From the figure one can see all frequencies are not heard in the same way. Frequencies that make up speech are within the smallest range but are heard better than others. For signal processing purpose, the auditory threshold

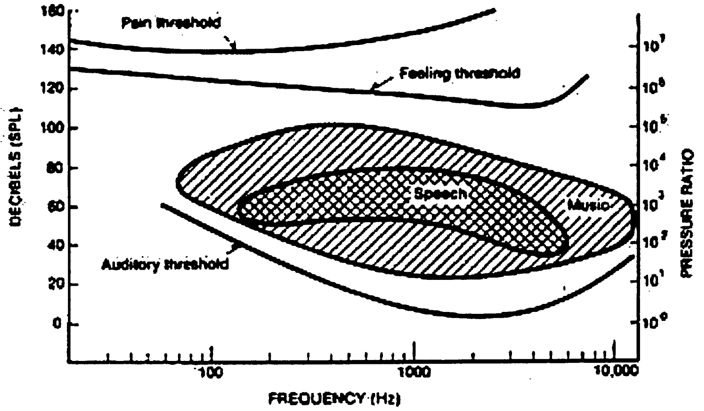


Figure 1.3: Human auditory threshold [5]

can be well approximated by a non-linear function as [4]

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 (\text{dB SPL}) \qquad (1.1)$$

where f is the frequency variable and the threshold is measured in dB SPL. Any signal component with an SPL that falls below the threshold will not be perceived by human ears. In other words, to remove or introduce components to this region will not effect signal's perceptual quality.

From physiological point of view, a critical band is the smallest band of frequencies that activate the same part of the cochlea. Human hearing system processes perceived sound in non-equal subbands. This behavior can be modeled as a bank of overlapping bandpass filters called auditory filters. The critical bands can be described as a measure of the "effective bandwidth" of the auditory filter [6]. Another view of this is that the critical bands represent an approximation of ear's ability to discriminate different frequencies. Experiments showed that the sound in each critical band can be analysed independently. While the number and the width of critical bands are still in dispute, Zwicker's 25-Bark [7] and Moore's ERB (Equivalent Rectangular Bandwidth) [6] models are widely applicable. The critical bands also can be explained in another way, adding up tones lying within a critical bands does not give any significant increase in perceived loudness over that of one strongest tone. This phenomenon promotes the auditory masking theory.

Auditory masking generally can be defined as a strong signal renders a weaker signal inaudible. The strong signal is referred as masker and the weak signal as maskee. Masking effects can be classified as simultaneous masking and temporal masking. Simultaneous masking is a frequency domain phenomenon. At certain time instance, a strong sound makes weaker neighbour frequencies sound imperceptible. Determinant factors of simultaneous masking includes masker frequency, masker amplitude, masker maskee distance (inter/intra band masking effects), and the nature of masker (tone or noise). Temporal masking is nonlinear perceptual phenomenon in the time domain, where a stronger sound occurs in time domain masks the weaker sound that is preceding it (pre-masking) or following it (post-masking). Determinant factors of temporal masking includes masker frequency, masker intensity, masker duration and temporal separation of masker and maskee. Figure 1.4 illustrates the patterns of masking. Simultaneous masking mainly happens with a critical band. Inter-band masking effect is much lower. Post-masking lasts in the order of 50 to 300 msec whereas pre-masking only has less than one-tenth of that duration. In general every masker generates simultaneous masking, pre-masking, and post-masking on the maskee signal. Simultaneous masking effect is more significant and its mechanism is better understood than temporal masking. Hence most of the audio processing research have focused on simultaneous masking.

1.3 Motivation

Hearing impairment may take many forms. The most common type of hearing loss can be characterized as sensorineural hearing loss. This type of hearing loss is mostly age-related and caused by long time exposure to excessive sound level. Sensorineural loss refers to deteriorated function in the inner ear or along the nerve pathway be-

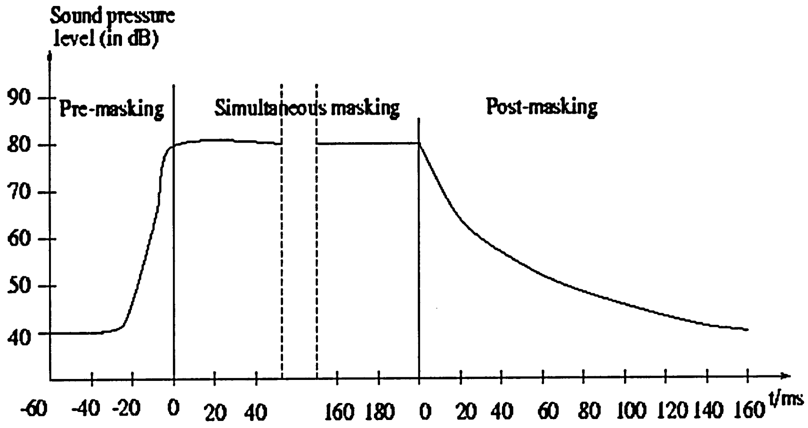


Figure 1.4: Typical patterns of masking

tween the inner ear and the brain. Typical effects of this type of hearing loss include increased absolute hearing thresholds, and abnormal perceived loudness recruitment. Quieter sounds are often inaudible, and louder sound may sound softer and distorted on the neural level. This type of hearing loss is usually permanent and not medically treatable. Wearing hearing aids is the only choice for sensorineural hearing loss individuals to achieve better hearing.

It is well known that hearing-impaired listeners experience more difficulty in understanding speech with background noise than normal hearing listeners. Unfortunately, the specific cochlear, conductive and linguistic mechanism responsible for diminished speech understanding still remains unknown. One of the possible explanation would be the speech is a highly redundant signal. Even if part of the speech is masked by noise; other parts of the speech signal will convey sufficient information to make the speech understandable. Such redundancy is much less for hearing-impaired person since parts of the speech are either distorted or completely inaudible [8]. Previous studies [8, 9] suggested hearing-impaired listeners generally need signal with much higher SNR, compared to normal hearing listeners, to achieve similar speech recognition rate. Hearing aids with noise reduction mechanism can greatly reduce the effect of additive noise and hence increase the speech intelligibility.

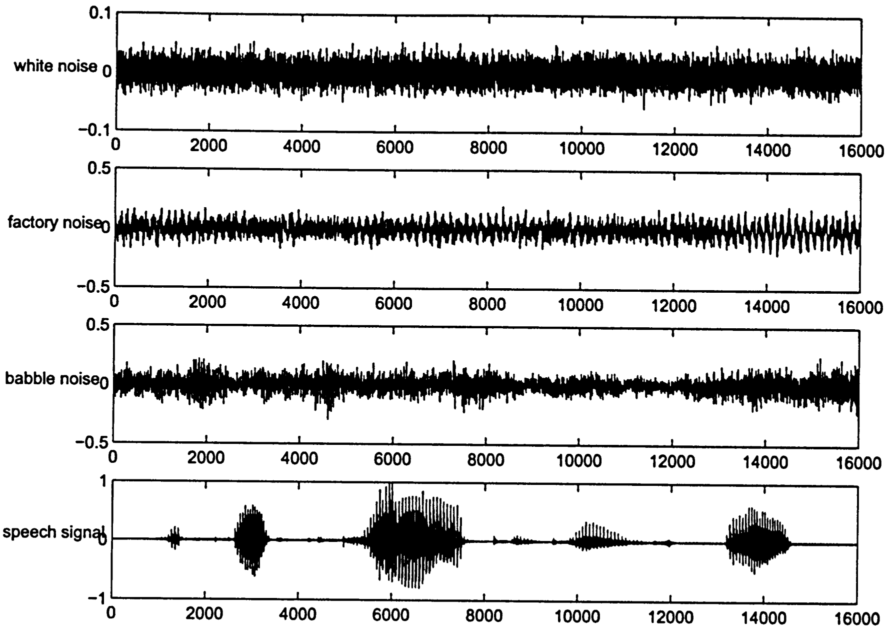


Figure 1.5: Time waveform of noise and speech

Speech noise reduction is fundamentally very difficult due to the nature of the most noise types. Another unsettled issue is how much noise removal is appropriate while minimizing speech distortion. Noise is any unwanted signal present in the desired signal. Here noise can be from the air conditioner, car, traffic or crowd of people and could be environmental. Figure 1.5 shows sound waveform plots of speech signal, white noise, factory noise , and babble noise. Some noise characteristics may be very similar to speech itself, such as babble noise. How to differentiate desired speech to babble speech? Some noise is actually meaningful such as siren and alarm sound. How can this type of noise be retained? Signal processing such as noise removal inevitably introduces speech distortion. How to determine the best trade-off between noise reduction and speech distortion? Furthermore, special consideration is needed to deal with the joint effects of hearing loss and background noise on speech perception.

The thesis presents a new approach for speech signal noise suppression with hearing loss compensation for hearing aids application, in a goal of improving speech in noise perception for hearing-impaired individuals.

1.4 Thesis Organization

This thesis is organized as follows. Chapter 2 gives a general filter representation of speech noise suppression followed by a review of various speech enhancement methods. Chapter 3 gives an introduction on wavelet analysis and its application to speech enhancement. The main contribution of this thesis begins in Chapter 4 , which presents the details of the proposed modified perceptual time frequency subtraction (PTFS) noise reduction algorithm and loudness compression hearing loss compensation schemes. Chapter 5 presents detailed comparative subjective and objective performance evaluations of the proposed method. Finally, Chapter 6 concludes the thesis with summary and suggests directions for future research work.

Chapter 2

Methodology

2.1 General Filter Representation of Speech Enhancement Techniques

In many situations, speech signals are degraded in ways of additional noise that limit their effectiveness for communication purpose. Speech enhancement aims at improving the performance of speech communication systems in noisy environments. Many multi-channel and single channel noise reduction algorithms have been proposed. However, most multi-channel scheme may not be suitable for hearing aids application due to cosmetic reasons. Hence, discussion in this thesis is limited to single channel speech signal enhancement. Discussion of the general filter representation of speech enhancement in transfer domain is mainly based on the work of Fan et al. [10].

Filtering in transform domains has been the main approach in modern speech enhancement techniques. One of the reason is because the noise spectrum could be estimated more accurately in transform domain than in time domain [10]. Figure 2.1 presents the block diagram of general speech enhancement in transform domain. The procedure is described as follows :

Noise corrupted time domain signal $x(n)$ first goes through analysis stage where signal is windowed and transformed into specific transform domain. Transformation

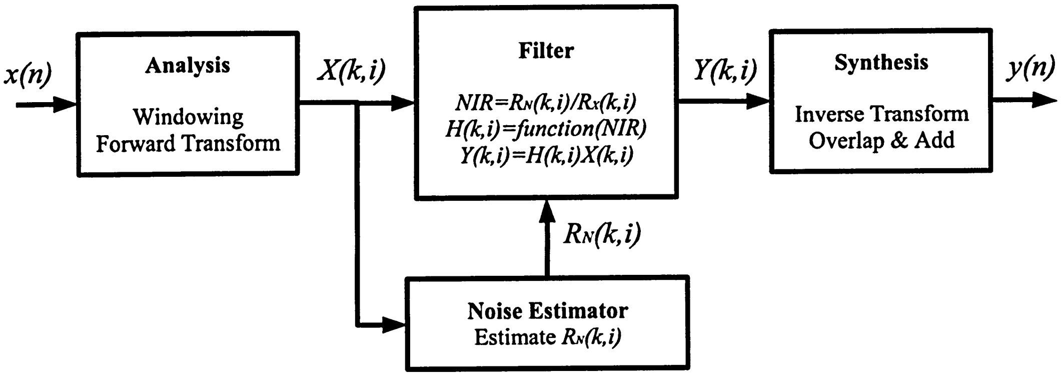


Figure 2.1: Speech enhancement: A General Filter representation

can be fast Fourier transform (FFT), short time Fourier transform (STFT), or wavelet transform. Windowing is especially important for transforms which only localize in time, such as the FFT. Transform coefficient $X(k,i)$ is filtered to remove the signal components that correspond to noise, where i is time index and k is transformed spectral component index. The filter transfer function (or gain function), $0 \leq H(k,i) \leq 1$ is a function of noise-to-input ratio (NIR). For a high NIR, it means signal contains more noise, $H(k,i)$ is made small ($\rightarrow$ 0) to suppress the noise components; while small NIR corresponds to cleaner signal, $H(k,i)$ is made large ($\rightarrow$ 1) to maintain the signal components. The filter's successfulness depends on the accuracy of NIR. Practically, prior knowledge about noise is unavailable. Therefore, a noise estimator is necessary to provide an accurate estimation of $R_N(k,i)$. To get the filtered output $Y(k,i)$ in transform domain,

$$Y(k,i) = H(k,i)X(k,i) \tag{2.1}$$

Finally, filtered coefficients are inversely transformed to the time domain to get the estimated clean signal $y(n)$.

The noise reduction scheme presented above can be commonly applied to widely used speech enhancement algorithms such as the spectral subtraction, the Wiener filter, wavelet shrinkage, and Ephraim-Malah filtering. Once the transform domain is selected, the only difference is the transfer function $H(k,i)$, which derived from either optimizing certain cost function (Wiener filter) or applying some constrain to differentiate noise and signal (spectral subtraction).

A noisy signal $x(n)$ consists of a clean signal $s(n)$ plus a noise component $d(n)$, assuming the signal and noise are independent to each other. $y(n)$ is an estimated clean version of $s(n)$. In the transform domain can be expressed as

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

where i is time index and k is transformed spectral component index. For Wiener filter, the goal is to minimize the mean square error (MSE) between the target and the estimated signal, or the cost function shown in Equation 2.3.

$$J(k,i) = \mathrm{E}\{(Y-S)\overline{(Y-S)}\} = \mathrm{E}\{(HX-(X-D))\overline{(HX-(X-D))}\} \tag{2.3}$$

Filter transfer function in Equation 2.1 is obtained by taking partial derivative with respect to H, and let the resulting equation to zero.

$$\frac{\partial J}{\partial H} = 2(H\mathrm{E}\{X\overline{X}\} - \mathrm{E}\{X\overline{X}\} + \mathrm{E}\{D\overline{X}\} = 0 \tag{2.4}$$

This leads to the Wiener filter gain function and is given by

$$H_{Wiener} = 1 - \frac{\mathrm{E}\{D\overline{S}\} + \mathrm{E}\{D\overline{D}\}}{\mathrm{E}\{X\overline{X}\}} = 1 - \frac{\mathrm{E}\{D\overline{D}\}}{\mathrm{E}\{X\overline{X}\}} = 1 - NIR \tag{2.5}$$

The most widely adopted spectral subtraction is an intuitive idea to remove noise. Estimated clean signal magnitude is obtained by subtracting estimated noise magnitude from that of noisy signal, as shown in Equation 2.6.

$$\sqrt{\mathrm{E}\{Y\overline{Y}\}} = \sqrt{\mathrm{E}\{X\overline{X}\}} - \sqrt{\mathrm{E}\{D\overline{D}\}} = H\sqrt{\mathrm{E}\{X\overline{X}\}} \tag{2.6}$$

Algorithm	Transfer Function
Spectral Subtraction	$H_{SS} = 1 - \sqrt{\frac{\lvert D\rvert^2}{\lvert X\rvert^2}}$
Wiener Filter	$H_{Wiener} = 1 - \frac{\lvert D\rvert^2}{\lvert X\rvert^2}$
Maximum Likelihood	$H_{ML} = \frac{1}{2}\left[1 + \sqrt{1 - \frac{\lvert D\rvert^2}{\lvert X\rvert^2}}\right]$
Non-linear Estimation	$H_{NLE} = f(\lvert D\rvert, \lvert X\rvert)$
Ephraim-Malah Rule	$H_{E.M.} = f(SNR_{post}, SNR_{prior})$

Table 2.1: Transfer functions for different algorithms

Therefore, the spectral subtraction filter function is

$$H_{SS} = 1 - \frac{\sqrt{\mathrm{E}\{D\overline{D}\}}}{\sqrt{\mathrm{E}\{X\overline{X}\}}} = 1 - \sqrt{NIR} \tag{2.7}$$

Building upon the original form of spectral subtraction algorithm, many improvement are proposed to resolve residue musical noise and related artifacts (missing consonants) problems. Table 2.1 shows different algorithms and their corresponding transfer functions.

After noise estimation is subtracted from the noisy signal, some portions of the noisy signal remain and the resulting signal has narrow-band peaks across the spectrum. Those isolated peaks vary in frequency from frame to frame producing short warbling sound, usually referred to as musical noise. Musical noise artifacts exist mainly because of the discrepancy between the estimated noise and the actual ones.

2.2 Summary of Previous Techniques

2.2.1 Spectral subtraction

Spectral subtraction [11] has been one of the most influential speech enhancement algorithm for its simplicity and effectiveness in dealing with wideband stationary noise. The method utilizes the fact that human sound perception is far more sensitive to the short-time speech spectrum than the phase [12]. This method essentially only requires to estimate the mean noise power. Since the phase is less important, the noisy signal phase is used to reconstruct the output signal. Implementation of spectral subtraction is extremely efficient using the FFT architecture. The major limitation of the technique is residual musical noise that degrades the intelligibility of the enhanced speech.

Since only average noise spectrum is estimated during the speech pauses, fluctuation in the actual noise spectrum of each frame leads to spectral spur in the resulting signal. To resolve the problem, good noise statistics must be known or estimated from non-speech portion of the noisy signal. This may not be trivial since no information about the noise or the speech is known in real life. Basic spectral subtract delivers fair result in terms of background noise reduction when SNR is not very low.

Goh et al. [13] suggested a simple way to reduce the musical noise by over-subtracting the noise estimation from the signal spectrum. But this method also eliminates more lower energy speech components. Low energy unvoiced speech is particularly important to hearing-impaired individuals. It is necessary to keep this kind of distortion to minimal [14]. An intuitive way will be performing‘ the spectral subtraction iteratively [15]. Their results showed better segmental SNR (SegSNR) gain as more iterations were performed. As an inevitable result, more distortion of the speech is also introduced.

Modified spectral subtraction using psychoacoustic properties reported improvement in the past decade. Singh et al. [16] found improvements in speech quality mean opinion score (MOS) test by including critical subband analysis instead of linear frequency scale in spectral subtraction. However, no significant improvement in objective segmental SNR test. Nishimura et al. [17] implemented similar method in wavelet domain. They showed SNR gain but no improvement in intelligibility test was achieved. Based on masking property of human auditory system, Virag [18] proposed another spectral subtraction algorithm. Her further publication [19] results showed the background noise was reduced and the residual noise was less structured, while the distortion of speech remained acceptable. However, objective tests used such as Itakura-Saito distortion and Articulation Index have poor correlation with subjective perception.

2.2.2 Wiener filtering

Tsoukalas et al. [20] adapted Wiener filtering rule for audio enhancement application. Wiener filter is the optimal filter for estimating speech in the minimum mean squared error (MMSE) sense. Similar to spectral subtraction, analysis performed on each frame is the power spectrum of the noise signal and an estimated noise power spectrum. As Wiener filter is non-causal and zero phase, estimated signal will use the phase information of the noisy one. Although the Wiener filter has a simple mathematical representation in the frequency domain, it requires prior knowledge of both speech and noise statistics, which are usually not known and needs to be estimated. Without accurate knowledge of the spectra, the Wiener filter approximation is not effective. It produces similar musical artifact as in spectral subtraction methods.

In [20] experiment results claimed little effect of musical noise. One main reason may be only signals of 20 dB SNR are tested for audio restoring purpose. For speech enhancement, this algorithm is further modified according to psychoacoustics rules in [21]. This method only filtered frequency components containing sufficient large audible noise spectrum in order to reduce speech distortion. A 40% intelligibility increase at -5 dB SNR is reported. Psychoacoustically quantifying the audible noise increases the complexity of this short-time spectral amplitude (STSA) based algorithm.

2.2.3 Ephraim-Malah filter

Ephraim-Malah filter [22] is a modification of the MMSE filter by adding an estimator for the priori SNR by modeling speech and noise spectral as statistically independent Gaussian random variables. Its gain function is decided by two parameters: a priori

SNR ratio and a posterior SNR ratio. The smoothness of the priori SNR results in a much lower residual noise without over filtering the speech itself. The implementation of Ephraim-Malah filter is straightforward despite its complicated mathematical expression comparing to spectral subtraction and the Wiener filter. Since the priori SNR needs to be estimated from the observation of noisy signal, again noise estimation determines the effectiveness of the Ephraim-Malah filter.

Ephraim-Malah algorithm has been widely modified to achieve perceptually better processed signal. The algorithm is adapted to discrete cosine domain [23] and wavelet domain [24]. Both reported to achieve better noise reduction and more pleasant sound than the original scheme. Recently, more sophisticated modification incorporating psychoacoustics rules [25, 26] have been developed aiming at reducing only audible part of noise spectral or finding best trade off between speech signal distortion and amount of unnatural residual noise. Such schemes reported to have better overall speech quality, even though some musical tones and speech distortion were still audible.

From above discussion in this chapter, one can conclude that successfulness of spectral based speech enhancement techniques largely depends on the accuracy of noise estimator or speech pause detection. The better the estimation is, the less residual noise and speech distortion result. Over the last decade, there have been great number of works on theory and practical methods to address the hearing-in-noise problem. Unfortunately, the problem is fundamentally very difficult for the most common types of noise. Furthermore, there are severe limits as to how much noise reduction is practically possible [8]. Especially for hearing-impaired subjects, who experience major difficulties when hearing in noisy environment. On the other side, considering joint effects of hearing impairment and background noise on auditory perception will allow the development of signal processing scheme for hearing aids application. By using a better suppression rule and incorporating psychoacoustics model, the determinant effect of noise estimation can be lessen on the performance

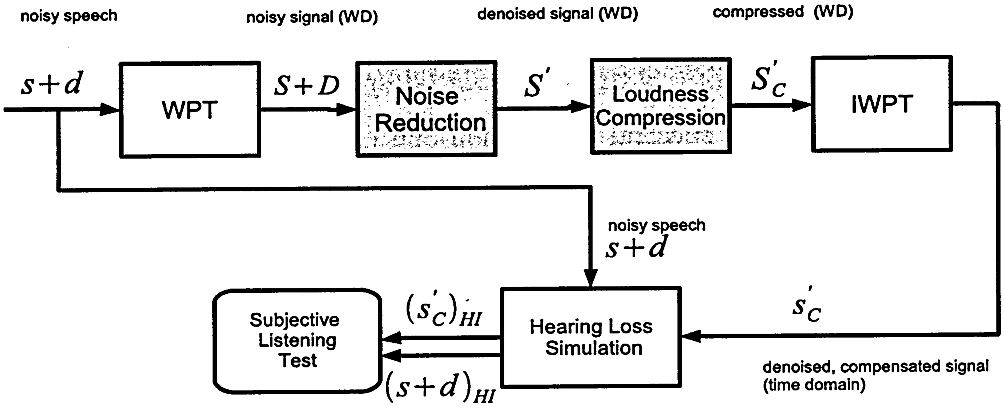


Figure 2.2: Block diagram of proposed method

of a speech enhancement algorithm. Instead of targeting to remove as much noise possible, the strategies used in this thesis is to reduce the effects of background noise on overall sound quality and speech intelligibility if possible.

2.3 Proposed Method

In this thesis, we proposed a robust speech enhancement system for hearing aids which will simultaneously conduct noise removal and hearing compensation by using critical band wavelet packet transform. With a reasonable complexity, the new speech enhancement scheme can improve perception in noise for hearing-impaired listeners. The proposed speech enhancement scheme can be used to improve next generation hearing aids device performance.

Figure 2.2 shows the block diagram of the proposed speech enhancement scheme. First the noisy speech time series is decomposed according to psychoacoustic critical bands by using wavelet packet transform (WPT). Wavelet coefficients of the noisy speech are perceptually weighted through a weighting function incorporating masking properties. The denoised wavelet coefficients are then compressed to compensate recruitment of loudness problem of hearing-impaired on the Compression stage. Inverse wavelet packet transform (IWPT) transform the processed signal back to time domain. To evaluate the processed speech on normal hearing subjects, Hearing loss simulation introduces effect of hearing loss to both original noisy signal and processed signal to simulate certain types of hearing impairment.

Detail of the proposed methodology will be discussed in Chapter 4.

Chapter 3

Wavelet Packet Analysis

This chapter provides an overview of the fundamentals of the wavelet transform and wavelet packet decomposition.

First, the underlying mathematics of the wavelet analysis is explained and then it is extended to wavelet packet transform. Since this work is based on the wavelet packet transform (WPT), fast filter bank implementation of the wavelet transform and wavelet packet transform are presented with corresponding computational complexity analysis.

3.1 Wavelet Transform Analysis

In order to present the idea, the Fourier transform is compared with the wavelet transform.

The Fourier transform transforms the time domain signal to frequency domain by using sinusoidal basis function to approximate the original signal. There are many advantages for this kind of approximation, as signal can be analyzed for its frequency content. However, the Fourier transform representation has a major drawback due to using sinusoidal basis function. Fourier sine and cosine functions are localized in frequency but not in time. In other words, they stretch infinitely in time, in transforming to frequency domain, and the time information of the signal is lost. With

Fourier analysis, it is impossible to tell which frequencies appear at what time. As a result, Fourier transform can not be used to approximate a signal whose properties change over time, i.e. non-stationary signal.

To address this problem, it requires a joint time-frequency representation. The STFT is an intuitive modification of the Fourier transform to analyze non-stationary signal. The basic idea behind the STFT is segmenting the signal by time-localized windowing and performing Fourier transform to each segment at a time. The STFT maps a signal into a two dimensional time-frequency representation. STFT achieves some degree of compromise between time and frequency representation of signals. It provides information about both when and what frequencies occur in a signal with limited precision. The imprecision drawback comes from the fixed length time window used to analyze the entire signal regardless of its frequency content of each segment. STFT is able to do wideband frequency analysis using narrow window, or narrowband frequency analysis using wide window, but not both simultaneously once the window size is selected.

To analyze the signal flexibly, a size-variable window is needed to access accurate view either in time or frequency. The wavelet transform gave exactly what to achieve this. Continuous wavelet transform (CWT) is defined as

$$W(a,b) = \frac{1}{\sqrt{a}} \int x(t)\psi^*(\frac{t-b}{a})dt \tag{3.1}$$

where $a > 0$ and b are scale and translation parameters respectively, ψ denotes the basis function or mother wavelet, and $W(a,b)$ is the continuous wavelet transform of $x(t)$. Equation 3.1 can be interpreted as inner product of $x(t)$ and the complex conjugate of the scaled and translated version of the basis function (wavelet) ψ

$$W(a,b) = \int x(t)\psi^*_{(a,b)}(t)dt =< x(t), \psi^*_{(a,b)}(t) > \tag{3.2}$$

where $\frac{1}{\sqrt{a}}\psi^*(\frac{t-b}{a})$ and $a > 0, b \in \Re$ are real continuous variables. CWT can also be expressed in convolution form as

$$W(a,b) =< x(t), \psi^{*}_{(a,b)}(t) >= x(t) \star \psi^{*}_{(a,b)}(-t) \tag{3.3}$$

CWT in convolution form can be interpreted as the output of an infinite bank of linear filters described by the impulse response $\psi^{*}_{(a,b)}(t)$ over the continuous range of scale a [27].

While the FFT maps a one dimensional series into an one dimensional sequence of coefficients, wavelet analysis maps it into a two dimensional array of coefficients. The extra dimension of information allows localizing signal in both time and frequency. But the two dimensional time-scale representation is highly redundant because of the non-orthogonal basis functions used in CWT. In addition, infinite number of the wavelets and the lack of analytical solutions of CWT make wavelet transform not practical for signal analysis.

To solve the problem, the discrete wavelet transform (DWT) has been introduced. To obtain DWT, the parameters a and b from CWT is discretized, much the same way as DFT does, DWT can be scaled and translated in discrete steps. There are many possible ways to discretize CWT, and most of the applications are primarily interested in dyadically spaced wavelets, which is a natural choice for computational effectiveness. Figure 3.1 shows a resulting time-frequency (time-scale) tiling from dyadic sampling. For dyadic case, CWT parameters is discretized by $a = 2^{j}$and $b = 2^{j}k$, $(j, k \in \mathbb{Z})$. Wavelet function and DWT are define as

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j/2}t - k) \tag{3.4}$$

$$W_{j,k} = \sum x(t)\psi_{j,k}(t) \tag{3.5}$$

Note, indexing rather than bracket is used for a more compact form, and also the conjugation is dropped since only the real-valued wavelet is used.

Dyadic sampling and restricting orthonormal basis wavelets functions will enable the use of Mallat's multiresolution analysis (MRA) to obtain DWT of the discrete signal.

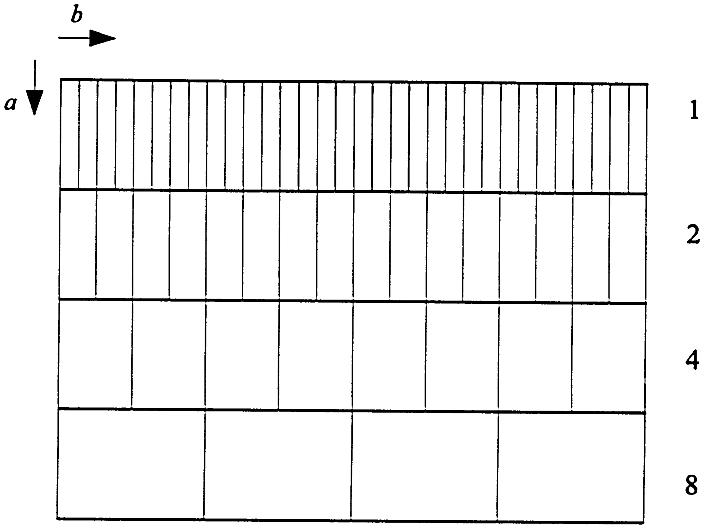


Figure 3.1: Time-frequency tiling of dyadic DWT

To form a complete basis in $L^2(R)$, it requires an infinite numbers of wavelets to cover the spectrum. In practice, when the lower frequency part of the signal is small enough, a low pass spectrum called scaling function φ, introduced by Mallat, can be used to represent the lower frequency portion. The scale function takes care of the spectrum covered by wavelets up to a scale of j. The rest is represented by wavelet function [28]. Scaling function at a given decomposition depth is a scaled and translated version of the mother scaling function.

$$\varphi_{j,k}(t) = 2^{-j/2}\varphi(2^{-j/2}t - k) \tag{3.6}$$

A signal sequence $x(t)$ can then be represented by its projection on $\psi_{j,k}$ for $j = 1, 2, 3, \cdots, j_{max}$, and its projection on to $\varphi_{j_{max},k}$, where $j_{max} = \log_z N$ and N is the block size of the sequence. Equation 3.7 shows the wavelet representation of any function $x(t) \in L^2(R)$.

$$x(t) = \sum_{j,k} c_{j,k}\psi_{j,k}(t) + d_{j_{max},k}\varphi_{j_{max},k}(t), \tag{3.7}$$

$j = 1, 2, 3, \cdots, j_{max}$, where j_{max} is the coarsest scale. The coefficients c_{jk} and d_{jk} are the discrete scaling coefficients and the discrete wavelet coefficients of $x(t)$ respectively.

3.2 Multiresolution Analysis and Filter Banks

The wavelet transform can be thought of as a filter bank. At different filter stages, output of the filters are wavelet and scaling coefficients. Filter bank is an analysis scheme originally applied in subband coding. The filter bank can be implemented in several ways. One way is to use many bandpass filters to split the spectrum into a number of frequency bands. Another way is to split the signal spectrum into two equal parts by lowpass and highpass filters. Since the highpass band has enough detail information, only the lower half band is further iteratively separated in the same way until desired resolution is reached. The advantage of the second scheme is only needed to design a pair of highpass and lowpass pass filters, while every bandpass filter has to be designed separately in the first scheme. As a result, the disadvantage of the second scheme is that the band coverage is fixed, while subband width of the first scheme could be freely chosen [28].

The wavelet function gives the bandpass bands with doubled bandwidth and the scaling function provides the lowpass band. In other words, the wavelet transform can perform the same filter band subband analysis by feeding signal to a bank of iterative highpass and lowpass filters. Mallat [29] derived a discrete orthogonal constant-Q filter bank implementation of the wavelet analysis, generally referred to as multiresolution analysis.

The MRA implementation of DWT analyzes the signal at different frequency bands with different resolution by decomposing the signal in to a coarse and detail

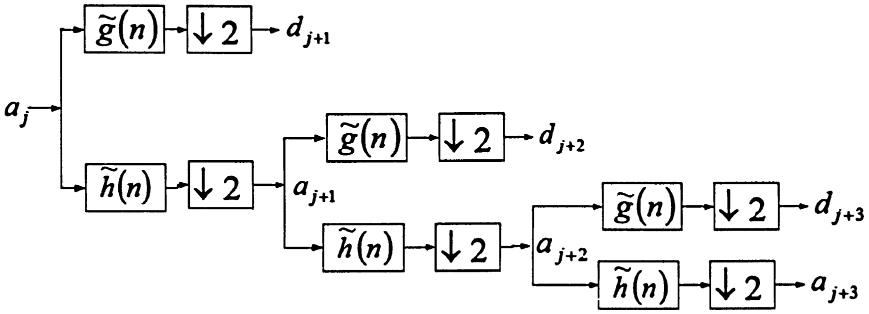


Figure 3.2: DWT decomposition tree [30]

information. The scaling functions and the wavelet functions are associated with the lowpass and highpass filters respectively. Equations 3.8 and 3.9 show at level $j+1$, scaling (coarse) and wavelet (detail) coefficients can be obtained from level j scaling coefficients by filtering with finite impulse response (FIR) lowpass filter $\tilde{h}(n)$ and highpass filter $\tilde{g}(n)$[30].

$$c_{j+1}(k) = \sum_m \tilde{h}(m-2k)c_j(m) \tag{3.8}$$

$$d_{j+1}(k) = \sum_m \tilde{g}(m-2k)c_j(m) \tag{3.9}$$

The filtering operation is achieved by convolving the sequence with the filter coefficients, or impulse response. The two equations also show a down-sampling, or decimating, by 2 of the filtering outputs. Down-sampling of the filter outputs is an important part of multirate filter banks. Without down-sampling, output of each filter stage will be doubled of previous stage by having two filters. The decomposition of the complete sequence into different frequency bands is obtained by successive highpass and lowpass filtering of the coarse coefficients followed by decimation by 2. A tree-like structure in Figure 3.2 illustrates the recursive nature of the approach.

Upon resynthesis, start from coarsest resolution, first up-sampling by two then highpass lowpass bank filtering. Perfect reconstruction can be achieved by appropriate

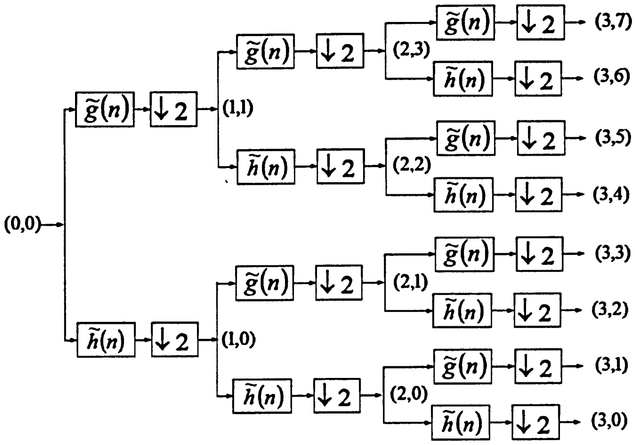


Figure 3.3: The full binary tree for three-stage wavelet packet transform [30]

filter design technique.

3.3 Wavelet Packet Analysis

As a natural extension to the DWT, the wavelet packet transform (WPT) was introduced by Coifman, Meyer and Wickerhauser [31]. Wavelet packet provide a powerful and versatile multiresolution analysis of a signal. A logarithmic frequency resolution of the DWT is not appropriate for some signal analysis. WPT as a generalization of the wavelet transform, can achieve flexible time-frequency resolution according to the signal characteristics. WPT allows high frequency band for further decomposition just as in lower band. Splitting both highpass and lowpass band results in a full binary tree filter bank structure, as shown in Figure 3.3.

Any node (j, k) of the binary tree is labeled by its depth j and location at the same depth k from left to right. Node (j, k) corresponding to a space $\mathbf{W}_{j,k}$. The recursive splitting defines a binary tree of wavelet packet spaces where each parent node is divided into two orthogonal subspaces, as shown in Equation 3.10 [32].

$$\mathbf{W}_{j,k} = \mathbf{W}_{j+1,2k} \oplus \mathbf{W}_{j+1,2k+1} \tag{3.10}$$

Wavelet packet coefficients are computed with a filter bank algorithm that generalizes the fast DWT as discussed in previous section.

3.4 Computational Complexity of Wavelet Analysis

Wavelet analysis of a length-N sequence using Mallat's algorithm with filter banks require $O(N)$ operations [30]. The computational cost of the wavelet transform is the convolutions carried out in each of the $\log_2 N$ stages. Since the number of coefficients in the convolution is halved after down-sampling in each stage, the total operations required for a full DWT decomposition is

$$O(N + \frac{N}{2} + \frac{N}{4} + \frac{N}{8} + \cdots + 1) < O(2N) \tag{3.11}$$

Therefor the complexity is linear with the length of the signal. In comparison, complexity of the FFT implementation of the DFT is $O(N\log(N))$ and that of DFT direct implementation is $O(N^2)$.

In the wavelet packet case, uniform division is used, that means number of coefficients in each stage are the same. Length of N sequence has a complete decomposition of $\log(N)$ stages. The complexity of the WPT becomes $O(N\log(N))$, similar to that of the FFT.

Fair computational complexity has enabled wavelet analysis to be applied for real-time application such as image processing, speech codec, and fingerprint recognition. Applications on real-time portable system such as hearing aid device could become a possibility in the foreseeable future.

3.5 Wavelet Transform for Noise Reduction

The wavelet transform introduced as an alternative way to analysis non-stationary signal provides a new way of representing signal into well behaved expression that has useful properties. A wavelet transform decorrelates signal structures, and filters uncorrelated noise. The noise is spread out evenly over all coefficients. In many signals such as human speech, the energy is concentrated in a small number of wavelet coefficients. The wavelet coefficients of a signal are considerably larger than that of noise. Hence, by thresholding the wavelet coefficient useful signal component is preserved while eliminating the noise. One popular technique is the wavelet shrinkage algorithm by Donoho and Johnstone [33, 34]. It can be summarized into following steps:

1. Decompose noise corrupted signal into wavelet coefficients
2. Apply a soft or hard thresholding to the noisy wavelet coefficients
3. Synthesize thresholded wavelet coefficients to obtain enhanced signal

For removing additive Gaussian noise, a global threshold

$$\lambda = \sigma\sqrt{2\ln(N)} \tag{3.12}$$

is used for thresholding. Noise standard deviation σ can be estimated by $\sigma = \mathrm{MAD}/0.675$, where the median absolute deviation (MAD) is obtained from the first decomposition stage of the wavelet transform [34]. N is the length of a frame. For the wavelet-packet transform case, the threshold becomes

$$\lambda = \sigma\sqrt{2\ln[N\log_2(N)]} \tag{3.13}$$

One may notice that above wavelet denoising scheme does not require to estimate the noise spectra. The threshold solely depends on the input level. Wavelet thresholding provides a simple yet fair noise reduction solution. However, some problems arise when the basic wavelet thresholding is applied to speech enhancement. Due to the time variability and highly non-stationary nature of real-life additive noise to the speech signal, a time-constant global threshold tends to over threshold the speech signal, especially in higher frequency bands. Since unvoiced sounds are high frequency and noise like, removing these results in poor perceptual quality. Basic wavelet with global thresholding is not sufficient for speech corrupted by real-life noises.

To improve wavelet based speech denoising method, effort has be focused on the modification of threshold selection and thresholding function. Level-dependent threshold [34] and node dependent threshold [35] have been proposed to better cope with colored noise. Semi-soft thresholding [36] and μ-law [35] thresholding were used to lessen the effect of time-frequency discontinuities that lead to speech artifact. Some [37, 36] suggested using a different threshold for voiced and unvoiced speech segment. All modifications of basic wavelet thresholding improve speech enhancement results to some degree. Among them algorithms that incorporating psychoacoustical model [38, 39, 40, 41] reported best improvement in term of speech quality. Therefore, the proposed wavelet based noise reduction algorithm utilizing psychoacoustical model is introduced in next Chapter.

Chapter 4

Wavelet-based Noise reduction and Compression

In this Chapter a new single-microphone speech enhancement scheme is proposed. The approach reduces noise by perceptually filtering the wavelet coefficients corresponding to noisy speech. Denoised speech is furthered enhanced by non-linear gain derived from hearing-impaired person's profile. Finally hearing loss simulation is used to efficiently evaluate the proposed approach.

4.1 Noise Reduction: Modified PTFS

The flowchart of the proposed noise reduction algorithm is presented in Figure 4.1. The detail of the algorithm will be explained in the following subsections. This noise reduction algorithm by itself can be used for speech enhancement application. Or as proposed in the thesis, cascaded with loudness recruitment compensation algorithm for hearing aids signal processing application.

4.1.1 Perceptual Time-Frequency Subtraction

Li et al. [38] proposed a perceptual time-frequency subtraction (PTFS) algorithm for speech noise reduction. Their approach provides the basis for the noise reduction

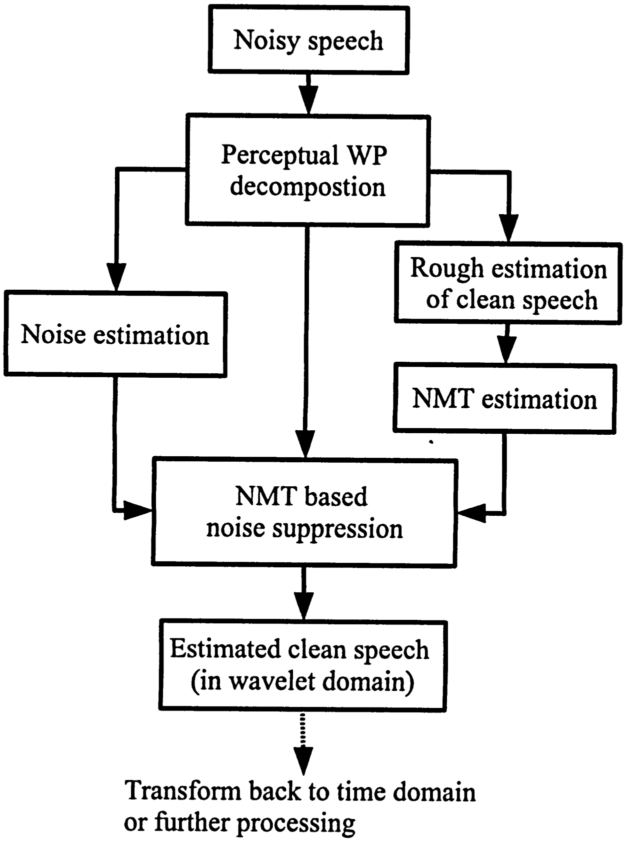

Noisy speech
Perceptual WP decompostion
Noise estimation
Rough estimation of clean speech
NMT estimation
NMT based noise suppression
Estimated clean speech (in wavelet domain)
Transform back to time domain or further processing

Figure 4.1: Flowchart of the proposed noise reduction algorithm

algorithm proposed in this thesis. It is a subtraction scheme in wavelet domain where auditory masking properties are incorporated to form a non-linear weighting function. Its basic assumption is the auditory system perceives sounds based on the SNR (or signal-to-masking ratio, SMR) in each critical band. Their proposed weighting function is given in Equation 4.1

$$H_{j,k} = 1 - \frac{\beta[\mathrm{E}(w_{j,k}^2(d))]^{1/2}}{|w_{j,k}(x)|} \tag{4.1}$$

where $w_{j,k}(x)$ are wavelet coefficients of the noisy signal, $[\mathrm{E}(w_{j,k}^2(d))]^{1/2}$ is the standard deviation of noise and β is an adjustable constant. The weighting function can be viewed as a filter that produces a non-linear transfer function as shown in Figure 4.2. The conventional one refers to the filter function that generated from spectral subtraction algorithm.

The weighting function is divided into three regions according to (Signal+Noise)-to-Noise ratio: noise-masking, signal-noise, and signal masking regions by two thresholds $\mathrm{T}_{\mathrm{low}}$ and $\mathrm{T}_{\mathrm{high}}$. In noise masking region where noise power is strong enough to make speech inaudible, weighting function is set to be close to zero. In signal masking region where signal is sufficiently strong to make noise inaudible, a unity gain is applied to retain all speech components. In signal-noise region where both speech and noise are audible, a non-linear gain is applied to suppress noise from the signal. In wavelet domain, the noise power is estimated during non-speech segment. Experimental results show that the PTFS algorithm yielded improvement in speech quality in terms of SNR gain, especially in unvoiced portions. A significant noise reduction result was achieved during speech pause segments.

The new noise reduction scheme proposed in the thesis includes two primary modification to the above original PTFS algorithm: (1) the signal processing is done in the wavelet domain via adaptive wavelet packet transform instead of fixed wavelet transform. Furthermore, considering auditory masking is observed over a Bark scale, the decomposition tree structure of the conventional wavelet packet transform is altered

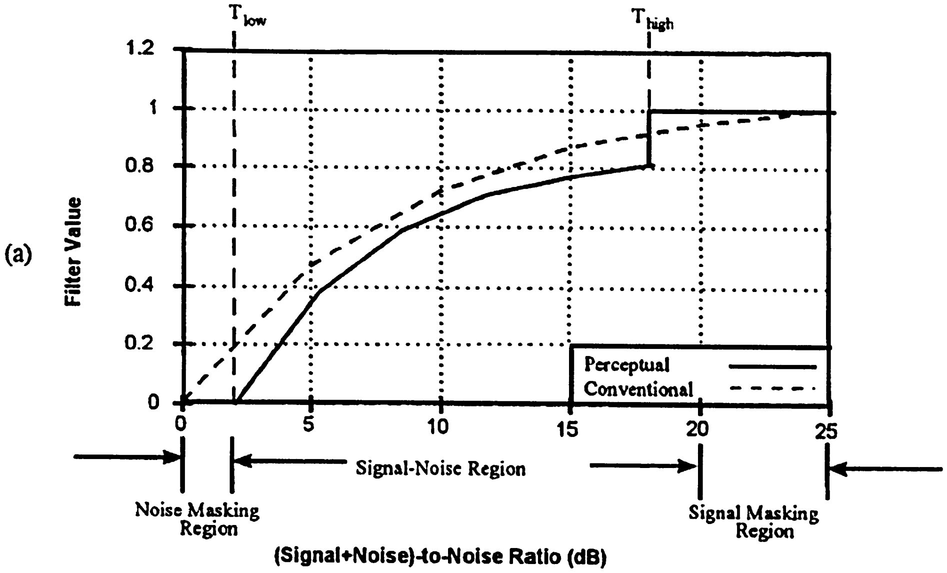


Figure 4.2: Weighting function of PTFS algorithm [38]

to approximate critical bands. (2) The second modification is the noise suppression structure. The proposed weighting function is adapted based on noise masking threshold instead of SNR. Algorithms that utilized masking threshold to adjust noisy signal showed superior perceptual speech quality to SNR-based rules [42, 25, 21].

4.1.2 Perceptual Wavelet Packet Decomposition

In this research, we utilize a wavelet packet decomposition algorithm mentioned in the previous works [40, 41, 25, 43]. This algorithm, usually referred as perceptual wavelet packet decomposition (PWPD), is designed to adjust the decomposition tree structure of the conventional wavelet packet transform to approximate the critical bands of the psychoacoustic model as close as possible. The primary advantage of PWPD is integrating a psychoacoustic critical bands model. As mentioned in Chapter 1, critical bands are of great importance since subjective responses to each bands are significantly different. In fact, a music noise reduction can already be obtained by applying noise reduction to subband derived from frequency group of the human auditory system instead of applying it to each frequency components [44, 45].

Critical band analysis scales is based on the widely used Zwicker's Bark scale [7]. Bark scale z can be approximately expressed as a function of linear frequency in Equation 4.2:

$$z(f) = 13\arctan(7.6 \times 10^{-4} f) + 3.5\arctan(1.33 \times 10^{-4} f)^2 \text{ [Bark]} \tag{4.2}$$

The corresponding critical band width (CBW) of the center frequency in each Barks can be expressed as

$$\mathrm{CBW}(f_c) = 25 + 75(1 + 1.4 \times 10^{-6} f_c^2)^{0.69} \text{ [Hz]} \tag{4.3}$$

where f_c is the center frequency in Hz. From physiological point of view, the human auditory frequency ranges from 20 Hz to 20 kHz, i.e. about 25 Barks. Since this research mainly focuses on the speech signal, its bandwidth mainly falls in the 200 Hz - 2500 Hz frequency range. Speech signal is sampled at 8 kHz per second yielding bandwidth of 4 kHz. For 4 kHz bandwidth, approximately 18-Bark scale is sufficient instead of the original 25 Barks. Specific to our implementation, a 5 level non-symmetry decomposition is employed to create 18 approximate critical bands (or subbands). Table 4.1 lists the center frequencies, CBW, and the lower/upper cutoff frequencies of critical bands up to 4000 Hz defined in [7].

According to the critical bands model above, the tree structure of PWPD is constructed as shown in Figure 4.3. Considering the binary tree nature of wavelet packet decomposition of the signal, PWPD results in 3 different bandwidths: 125 Hz at level 5, 250 Hz at level 4 and 500 Hz at level 3. Resulting subband and its bandwidth is described in Table 4.2. A discrete time domain speech signal $x(n)$ of length N after PWPD will be corresponding to a set of wavelet coefficients $w_{j,m}$, where $j = 3, 4, 5$ is the decomposition level and $m = 1, 2, ..., 18$ is subband index. Theoretically, the number of coefficients in each $w_{j,m}$ would be $N/2^j$.

Critical Band (Bark)	Center Frequency (Hz)	Critical Bandwidth (Hz)	Lower/Upper Cutoff Frequency (Hz)
1	62.5	125	—/125
2	187.5	125	125/250
3	312.5	125	250/375
4	437.5	125	375/500
5	562.5	125	500/625
6	687.5	125	625/750
7	812.5	125	750/875
8	937.5	125	875/1000
9	1062.5	125	1000/1125
10	1187.5	125	1125/1250
11	1375	250	1250/1500
12	1625	250	1500/1750
13	1875	250	1750/2000
14	2125	250	2000/2250
15	2375	250	2250/2500
16	2750	500	2500/3000
17	3250	500	3000/3500
18	3750	500	3500/4000

Table 4.1: Critical bands under 4 kHz

Level (j)	Subband index (m)	Bandwidth (Hz)
5	1,2,...,8	125
4	9,...,15	250
3	16, ...,18	500

Table 4.2: Subbands of PWPD

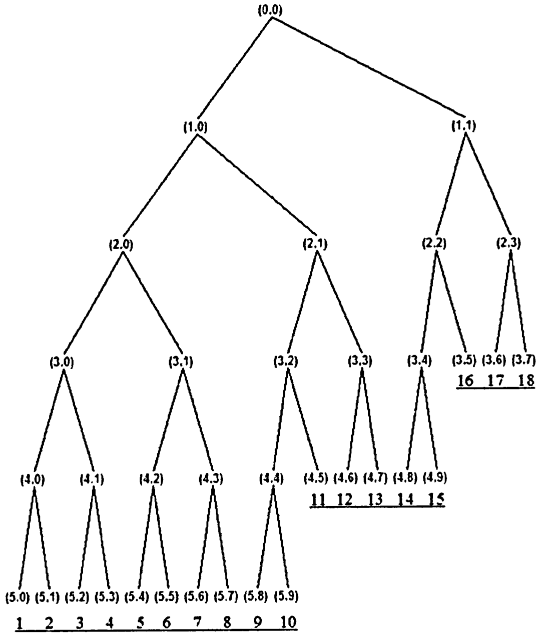


Figure 4.3: Tree structure of PWPD: numbers in brackets indicate the depth position; underlined numbers are subband

4.1.3 Modified PTFS

Equation 2.2 could be rewritten as

$$X_{j,k}(m) = S_{j,k}(m) + D_{j,k}(m) \tag{4.4}$$

where $X_{j,k}$,$S_{j,k}$, and $D_{j,k}$ are the wavelet coefficients of noisy, clean and noise signal of frame m respectively. The estimated clean speech signal $\widehat{S}_{j,k}$ can be expressed as noisy signal multiplied by the weighting function or gain H_j.

$$\widehat{S}_{j,k}(m) = H_j(m) \cdot X_{j,k}(m) \tag{4.5}$$

The error or deviation between clean and enhanced signal is defined as:

$$\varepsilon_{j,k}(m) = \widehat{S}_{j,k}(m) - S_{j,k}(m) \tag{4.6}$$

Substituting Equations 4.4 and 4.5 into Equation 4.6

$$\begin{aligned}\varepsilon_{j,k}(m) &= H_j(m)\cdot(S_{j,k}(m)+D_{j,k}(m))-S_{j,k}(m)\\ &= [H_j(m)-1]\cdot S_{j,k}(m)+H_j(m)\cdot D_{j,k}(m)\\ &= \varepsilon_S(m)+\varepsilon_D(m)\end{aligned}\tag{4.7}$$

where $\varepsilon_S(m) = [H_j(m) - 1]\cdot S_{j,k}(m)$ represents the speech distortion and $\varepsilon_D(m) = H_j(m)\cdot D_{j,k}(m)$ represents the residual noise in wavelet domain.

As a relatively simple single-input speech enhancement algorithm, PTFS showed superior noise suppression compared to classical techniques such as spectral subtraction and hard/soft wavelet thresholding. A weighting function solely depended on the segmental SNR achieved similar result as in spectral flooring and over-subtracting noise floor [46], and it eliminated the musical noise. However, it was done at the expense of introducing speech distortion. Experiment results showed that PTFS processed speech sounds unnatural and muffled, especially under low SNR condition. One main reason is the thresholds $\mathrm{T}_{\mathrm{low}}$ and $\mathrm{T}_{\mathrm{high}}$ are based on global wavelet threshold. As a result, it will significantly reduce the speech intelligibility for the listeners. Unless perfect information of noise is available, it is not practical to minimize the speech distortion and the residual noise simultaneously. Instead, the proposed method in the thesis is to find a balance between the two under the consideration of perceptual properties. Therefore, it is necessary to modify the weighting function so that it is adapted based on the noise masking threshold. If the noise is smaller than NMT, human ear cannot perceive the noise. In this case, unity gain is applied to minimize speech distortion. In other words, noise suppression is only applied to noise that is audible to human ears.

The weighting function in Equation 4.8 is derived from [42]

$$H_j = \frac{1}{1 + max\left\{\sqrt{\frac{S^2}{Th_m}} - 1, 0\right\}} \tag{4.8}$$

where s^2 is estimated noise power and Th_m is the noise masking threshold. Accurate NMT estimation facilitates speech enhancement scheme to achieve balance between noise reduction and potential processing distortion. Most of the existing masking threshold estimation algorithms are based on a similar model of the human auditory system. Among the algorithms that of Johnston's [47] and ISO MPEG-1 are widely used. The masking threshold estimation implemented in this thesis is mainly based on Johnston's. The following is a brief description of the algorithm.

1. Compute the Bark spectrum energy

The masking threshold is estimated on the wavelet coefficients of PWPD. The wavelet coefficients are grouped in to critical bands. Energy in each critical band is summed as shown in Equation 4.9

$$B(\xi) = \sum_{k=bl_\xi}^{bh_\xi} |X_{j,k}|^2 \tag{4.9}$$

where $\xi = 1, 2, ..., 18$ are critical band index, bl_ξ and bh_ξ denote the lower and higher boundary of the critical band respectively.

2. Spreading function applied on the critical band energy

The spreading function is used to estimate the effects of masking across critical bands. It is calculated as a matrix $S(i, j)$ given as

$$S(i, j) = 15.81 + 7.5(i - j + 0.474) - 17.5\sqrt{1 + (i - j + 0.474)^2} \tag{4.10}$$

where i is the bark frequency of the masked signal and j is the bark frequency of the masked signal, spreading is calculated for Bark frequency distance limited to $(|i-j| < 25$. Figure 4.4 shows the spreading function for Bark distance of $-5 < i - j < 12$.

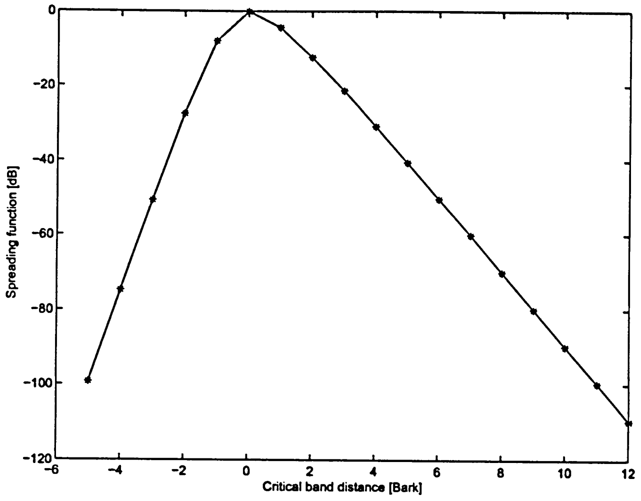


Figure 4.4: Spreading function

The convolution of the $B(\xi)$ with the spreading function is implemented as a matrix multiplication given as

$$C(\xi) = S(i,j) * B(\xi) \tag{4.11}$$

The value of $C(\xi)$ denotes the spread critical band spectrum for band ξ.

3. Calculating the noise masking threshold

There are two kinds of noise masking thresholds: tone masking noise and noise masking tone. In order to determine the noise-like or tone-like nature of the signal, the Spectral Flatness Measure (SFM) is used. The SFM is defined as the ratio of the geometric mean G_m of the power spectrum to the arithmetic mean A_m of the power spectrum.

$$SFM_{dB} = 10\log_{10}\frac{G_m}{A_m} \tag{4.12}$$

Then the coefficient of tonality is defined as follows:

$$\alpha = \min(\frac{SFM}{SFM_{max}}, 1) \tag{4.13}$$

where SFM_{max}corresponds to a signal which is assumed to be entirely tone-like and

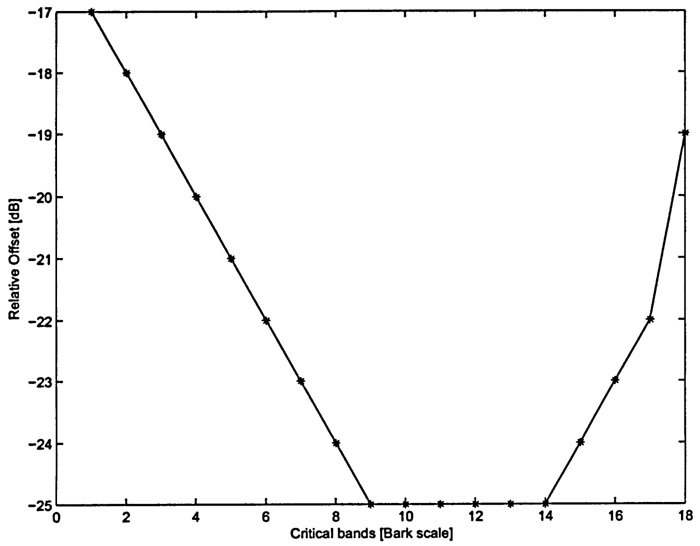


Figure 4.5: Relative offset $O(\xi)$ for noise masking threshold calculation

is set to -60 dB A zero dB SFM indicates a signal that is completely noise like. To find the masking threshold the following offset is subtracted from the spread critical band spectrum in dB.

$$O(\xi) = \alpha(14.5 + \xi) + 5.5(1 - \alpha) \tag{4.14}$$

To avoid accurate estimation of signal tonality that reduces the computation load, a simple estimation of $O(\xi)$ by Sinha and Tewfik [48] is used in the thesis. Based on the fact that the speech signal has a tone-like nature in lower critical bands and a noise-like nature in higher bands, the resulting value of $O(\xi)$ are represented in Figure 4.5

The noise masking threshold $T(\xi)$ is obtained by subtracting the offset $O(\xi)$ from the spread critical spectrum $C(\xi)$

$$T(\xi) = 10^{\log_{10}(C(\xi) - O(\xi)/10)} \tag{4.15}$$

<u>4. Renormalization and integration of the absolute threshold of hearing</u>

Since convolving subband energy with spreading function increases energy in each wavelet subband. The renormalization process takes this into account by multiplying each $T(\xi)$ by the inverse of the energy gain. Final noise masking threshold $Th_m(\xi)$ is shown in Equation 4.16.

$$Th_m(\xi) = T'(\xi) = \left(\frac{B(\xi)}{C(\xi)}\right) T(\xi) \tag{4.16}$$

Finally, the masking threshold is compared with absolute threshold of hearing. If in any critical band noise masking threshold is lower than the absolute hearing threshold, it is changed to the absolute threshold of that critical band.

An accurate estimation of the NMT is very important to the algorithm performance. It is found out in the experiment that a roughly denoised speech will greatly help finding the NMT accurately. Especially if the signal is under low SNR condition. A simple level-dependent thresholding method [49] is used for finding the rough estimation of clean speech in the proposed algorithm. In each frequency band the threshold is proportional to the standard deviation of the noise in that band. The threshold used at level j is

$$T_j = \sigma_j \sqrt{2\log(N_j)} \tag{4.17}$$

where $\sigma_j = MAD_j/0.6745$, N_j is the number of samples in scale j. MAD_j is the absolute median estimation at scale j.

After the noise reduction process, the estimated clean speech is passed to next stage to compensate for hearing loss effect.

4.2 Recruitment Compensation: Loudness Compression

As stated in the previous Chapter, elevated auditory threshold and associated loudness recruitment are the symptoms of sensorineural hearing loss. Recruitment here means a growth or increase. Threshold of hearing is raised non-uniformly with fre-

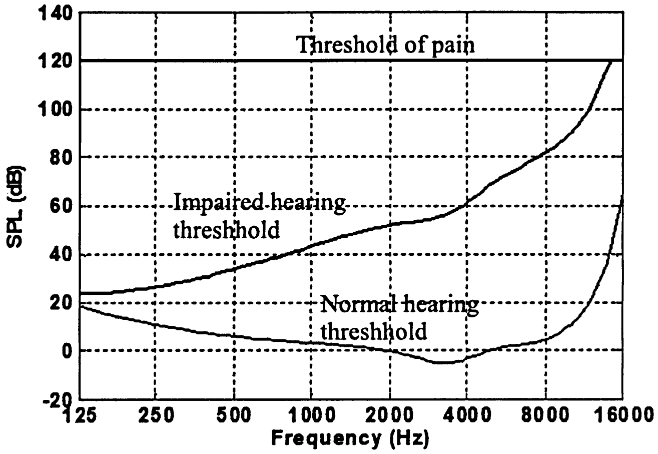


Figure 4.6: Hearing threshold and dynamic range

quencies and sometimes the threshold of pain is lowered. Thus the distance between the two thresholds so called the dynamic range is reduced. Figure 4.6 shows a hearing threshold of a typical hearing-impaired person. Individuals with this type of hearing loss experience unusual loudness relationship. Equal sound intensity increments do not produce equal increments in the loudness perception uniformly across all the frequency range, as it should be for normal hearing listeners. Reduced dynamic range causes small change in intensity to give larger changes in perceived loudness.

The signal processing approach to the problem is to find a preprocessing operator to enhance a signal that will undergo a known distortion. In the hearing aids case, the distortion comes from the effects of the hearing impairment, and the preprocessor forms the hearing aids device [50]. The hearing aids' function is to offset the effect of hearing impairment. So individual with hearing loss can perceive sound as a normal hearing will do. As an aid to hearing-impaired, simple linear amplification will not solve the problem. Since it makes the high frequency consonants audible while amplifies the low frequency vowels over the threshold of pain. So a limited maximum signal level is used to avoid this kind of discomfort. Linear amplification provides fair hearing loss compensation only when the input level is not very low and there is insignificant word to word amplitude variation [51].

To make the sound audible, the dynamic range of the incoming sound needs to be manipulated to fit in the dynamic range of impaired hearing. That leads to amplitude compression concept described by Villchur [52] . The basic idea of loudness compression is to restore the natural dynamic range of input speech to the impaired listener to provide all the speech cues in a more natural way. Signal should be in the audible range of impaired ears in order to be heard. Soft sound for normal should be soft for impaired ear and loud sound should be loud. It is called compression since the processing reduces the dynamic range of the signal in a certain way.

Conventional amplitude compression chooses parameters based on normal conversation and remains fixed regardless of input condition. It will tolerate only small input variations. An input too weak will not get a sufficient gain and an input too strong will be clipped, resulting in unnatural speech and reduced intelligibility. Thus, compression with gain varying in time and frequency according to the input level is highly desired. In order to compensate loudness, the spectral analysis should be similar to human auditory system, such as the PWPD used in this thesis.

The loudness compression proposed in the thesis is input independent, wavelet-based amplitude compression over a human auditory model. It is based on the work of Drake et al.[53] and Li et al. [51].

In this thesis, loudness compression is calculated in each critical bands, since loudness summation is strongly associated with critical bandwidth. Compression can be viewed as a gain applied to the original input. Gain is applied in such a way as to amplify the coefficient from a given equal loudness curve in the normal-hearing person's hearing profile to the corresponding equal loudness curve in the hearing-impaired person's profile. Gain is calculated for each critical band such that the log intensity above hearing threshold to dynamic range of hearing is the same for the hearing-impaired listener as the corresponding ratio for the normal-hearing listener [53]. Refer the idea to Figure 4.7, and mathematically the following equation should

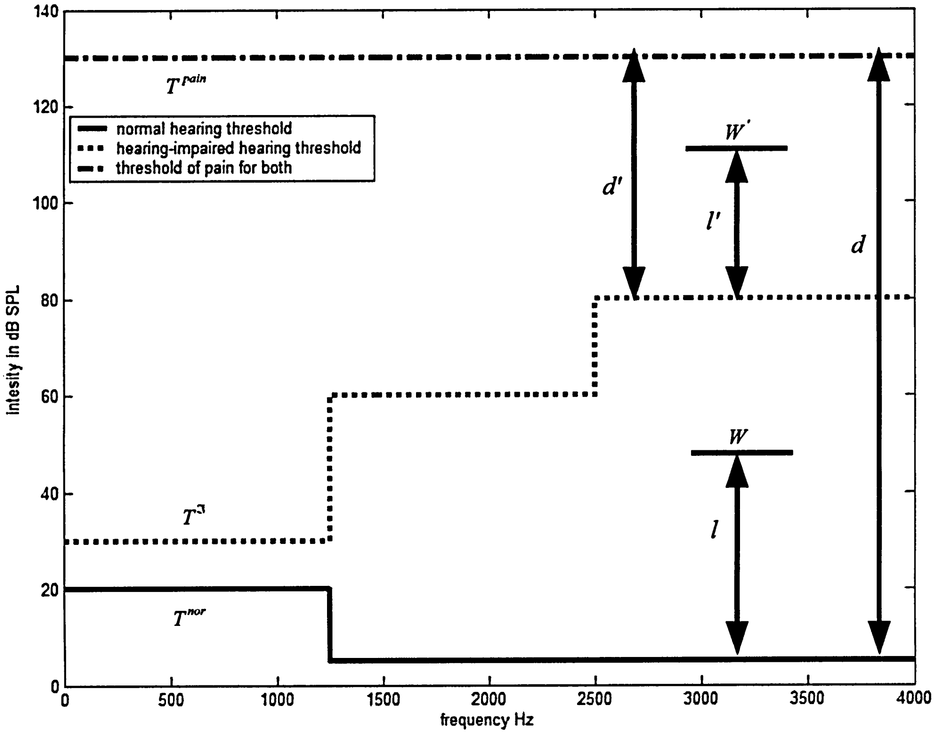


Figure 4.7: Compression gain computation

hold.

$$\frac{d'_\xi}{d_\xi} = \frac{l'_\xi}{l_\xi} \tag{4.18}$$

where d and d' are the dynamic ranges of normal hearing and impaired hearing respectively; l is the distance between uncompressed wavelet coefficients W and normal hearing threshold T^{nor}, and l' is the distance between compressed wavelet coefficients $W'$and impaired hearing threshold T^{im}; ξ is critical band index. ξ footnote is omitted for simplicity of the following equations:

$$\begin{aligned} d &= T^{pain} - T^{nor} \\ d' &= T^{pain} - T^{im} \\ l &= W - T^{nor} \\ l' &= W' - T^{im} \end{aligned} \tag{4.19}$$

Note the above variables in Equation 4.19 are in dB SPL. From Equation 4.19 and 4.18, compressed wavelet coefficient in dB in one critical band can be expressed as

$$W' = T^{im} + l' = T^{im} + \frac{d'}{d}l = T^{im} + \frac{T^{pain} - T^{im}}{T^{pain} - T^{nor}}(W - T^{nor}) \tag{4.20}$$

T^{nor} is taken from typical normal hearing person's profile. T^{im} and T^{pain} can be retrieved from hearing-impaired individual's audiogram test. Upon reconstruction from wavelet domain, a loudness compensated speech will result.

Since a critical band decomposition in wavelet domain already exists in the previous noise reduction module, the complexity of compressed coefficients computation is minimum. The compressed coefficients are guaranteed to be within the impaired hearing person's dynamic range, regardless of the input speech intensity.

4.3 Hearing Impairment Simulation

It is important to know how the signal processing helps the impaired listeners. Given the most of the research setting, it is not realistic to have impaired hearing persons to do listening tests repeatedly. Therefor, simulation of hearing loss on normal listening person is highly desirable.

Simulated hearing loss gives a realistic experience of auditory consequences of hearing impairment to the normal hearing listeners. It can be used as a hearing protection educational tool. By limiting simulation to certain aspect of hearing loss, it will greatly facilitate the development of newer generation of hearing aids devices. Real hearing-impaired subjective listening tests are often very expensive and time consuming. Subjects with different types and degree of hearing loss introduce individual bias effects in the test results. These issues make test on hearing-impaired subjects prohibitive in the signal processing algorithm development stage. A good hearing loss simulation algorithm will improve signal processing algorithm development and fitting for hearing aids. An ideal hearing loss simulation should account for the following psychoacoustics characteristics: elevated threshold in quiet, abnormal growth of loudness, and reduced temporal and spectral resolution [9].

There are two major classes of hearing loss simulations. The first one simulates elevated threshold in quiet by adding spectrally shaped masking noise. The second class based on multi-band dynamic range expansion. Unfortunately, there is very few detailed formulation of any kind of hearing loss simulation in the literature.

Simulation with shaped masking noise gives only general effect of hearing loss. It is more suitable for educational demonstration purpose. The method in the second category is used here so it is possible to control the characteristics of certain hearing loss. Implementations in [54] and [55] both belong to this category. Subjective tests have confirmed the effectiveness of the method in simulating sensorineural hearing loss. A hearing loss function (HLF) is utilized to calculate how much the hearing loss gain should be introduced. The gain is frequency and input level dependent.Figure

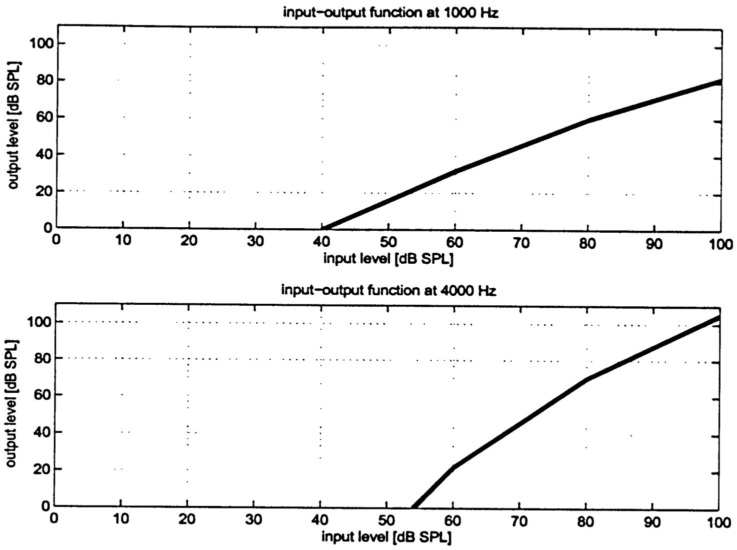


Figure 4.8: Expansive input-output function at 1000 Hz and 4000 Hz

4.8 shows a sample input-output function at 1000 Hz and 4000 Hz for simulating a specific hearing impaired listener.

The hearing loss simulation is a reverse process of loudness compression to some degree. Since the PWPD already provided the critical band decomposition, the next step is to calculate the energy in each band and apply the right hearing loss gain. After inverse transform the normal speech is mapped into the impaired one. The hearing loss simulation algorithm can simulate sensorineural hearing impaired. It is useful in the algorithm developing stage but the intention is not to replace the subjective test altogether.

Chapter 5

Performance Evaluation

Both objective and subjective measures are used to evaluate the performance and to find efficient parameter settings for speech enhancement in this thesis. Speech quality measures based on the ratings by human listeners are called subjective speech quality measures. Using human as an acoustic measuring device is the intuitive and ultimate way to evaluate the speech quality. However, it is not desirable due to potential various subject bias, poor control of listening conditions, and the ambiguity associated with the interpretation of "good quality". Subjective tests are generally simple but usually very expensive in terms of time and cost consumption. Objective quality assessment for noise reduction schemes is needed to reduce time-consuming and cost-intensive subjective listening tests. Objective quality measure should be consistently repeatable, and most importantly should incorporate psychoacoustics knowledge.

The speech samples used in the evaluation tests are taken from the TIMIT speech database. They are sampled at 8 kHz. To evaluate the noise reduction performance, both synthetic and real life noise are added to corrupt the clean speech signal. Different types of background noises from the NOISEX-92 database have been used including car noise, factory noise, and babble noise. The variance of noise is adjusted to obtain SNRs in the noisy signals according to Equation 5.1, ranging from -5 dB to

15 dB. SNR<10 dB is considered very low SNR condition.

$$x(n) = s(n) + \left(\sqrt{\frac{var(s(n))}{var(d(n))} \cdot \frac{1}{10^{SNR/10}}} \right) \cdot d(n) \tag{5.1}$$

5.1 Noise Reduction Evaluation

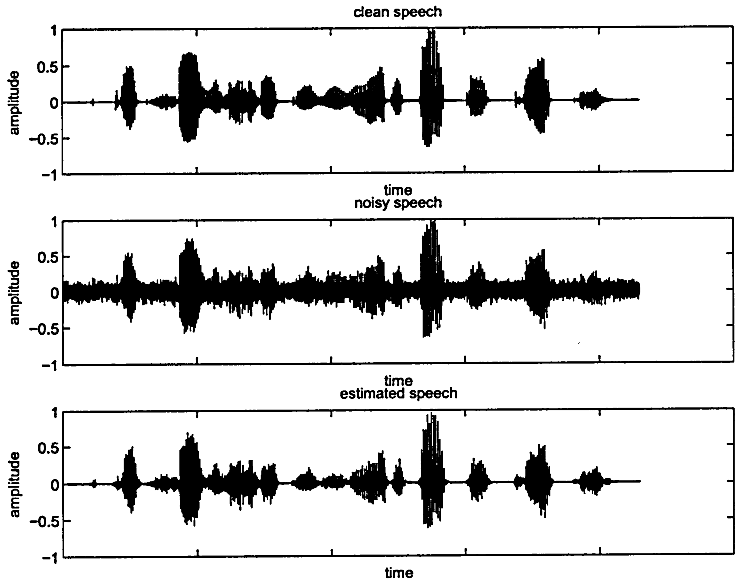
In this section, a detailed performance evaluation is presented in the forms of objective and subjective speech quality tests. To start, temporal and spectral plots of clean, noise degraded, and enhanced speech using the proposed algorithm is shown to illustrate the performance.

The processed sentence is "Good service should be rewarded by big tips", degraded with AWGN at SNR=5 dB. From temporal waveform plot in Figure 5.1(a), one can see noticeable noise reduction by the proposed PWS algorithm. In Figure 5.1(b), the spectrum of the estimated speech in a voice segment is mostly preserved with noise removed. The spectrum in segment corresponding to the speech pause is almost clean. If observed carefully, one will notice some low frequencies components of noise is left untouched, this will help preserve speech intelligibility.

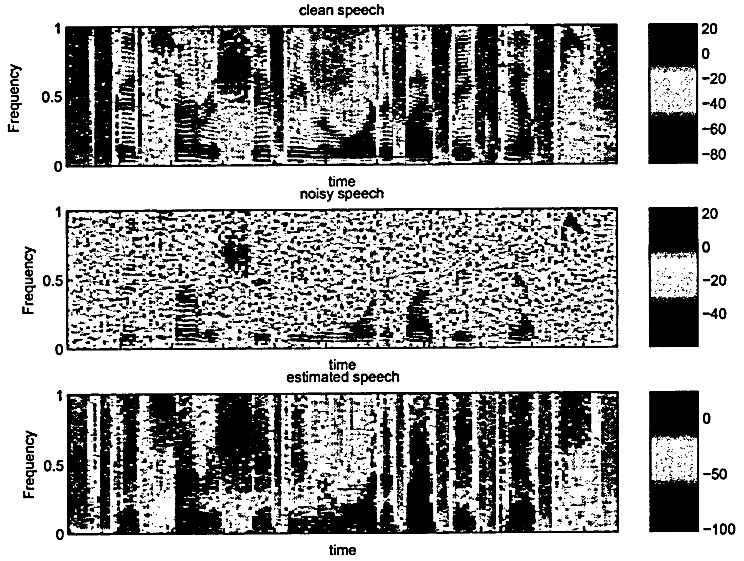
The remaining of this Section will cover the details of objective and subjective evaluations with comparison to other techniques.

5.1.1 Objective Evaluation

The PESQ (Perceptual Evaluation of Speech Quality) measure (ITU 2001) [56] and segmental SNR are utilized for the objective evaluations in this research. PESQ was recently adopted as an ITU-T recommendation (P.862) for signal distortion in audio and speech codec evaluation. Objective measures such as Itakura-Saito distortion, Articulation Index, Segmental SNR, and SNR have been correlated to subjective tests at 59%, 67%, 77%, and 24%, respectively [57], while the PESQ results have a 93.5% correlation with subjective tests [56]. Poor correlation measures tend to emphasize more on the similarity of actual waveforms of the signals instead of how



(a) Waveforms of the clean, noisy, and estimated speech



(b) Spectrograms of the clean, noisy, and the estimated speech

Figure 5.1: Waveforms and spectrograms of the clean, noisy, and estimated speech (SNR=7 dB, AWGN)

they actually sound. As a result, PESQ and Segmental SNR tests are chosen for objective evaluation.

For comparison purpose, the sample signals are also processed by the following speech enhancement algorithms.

- SS: Spectral subtraction
- WT: Wavelet thresholding
- EM: Ephraim-Malah algorithm
- PTFS: Perceptual time-frequency subtraction
- PWS: Proposed perceptual wavelet subtraction

5.1.1.1 Segmental SNR Test

Segmental SNR (SegSNR) is the most widely used time domain measure of speech quality. SegSNR is defined as an average of the SNR values of all short segments (frames) over the entire signal.. It can be formulated as in Equation 5.2.

$$\mathrm{SegSNR} = \frac{1}{M}\sum_{m=0}^{M-1} 10 \cdot \log\left[\frac{\sum_{i=0}^{N-1}[(s(i+mN)]^2}{\sum_{i=0}^{N-1}[(s(i+mN)-\hat{s}(i+mN)]^2}\right] \tag{5.2}$$

where N is the number of signal samples in one segment (frame) and M is the number of frames.

From the formulation, it is easy to figure out that the technical measure such as the SegSNR incorporates little speech distortion information. In terms of objective measures, a SegSNR gain is define as the difference between the SegSNR value of the processed signal and the unprocessed signal. SegSNR poses a problem in which silence intervals in the speech appear. In silence frame, the original speech is close to zero, any amount of noise can give rise to a large negative SNR for that segment, which could significantly bias the overall measure of SegSNR. This problem can be resolved by including the SegSNR of the frame only if the frame energy is above a specified threshold [57]. Since SegSNR has been so extensively used in most speech processing research as objective quality evaluation, it is included in this research as a direct comparison to other techniques proposed in the literatures, even though SegSNR is not an ideal perceptual measure.

Tables 5.1, 5.2, 5.3, 5.4 show a comparison of SegSNR gain of different algorithms with four types of noises. Figure 5.2 shows comparison of SNR gain performance of different noise reduction algorithms. Blue line is segmental SNR of the noisy signal.

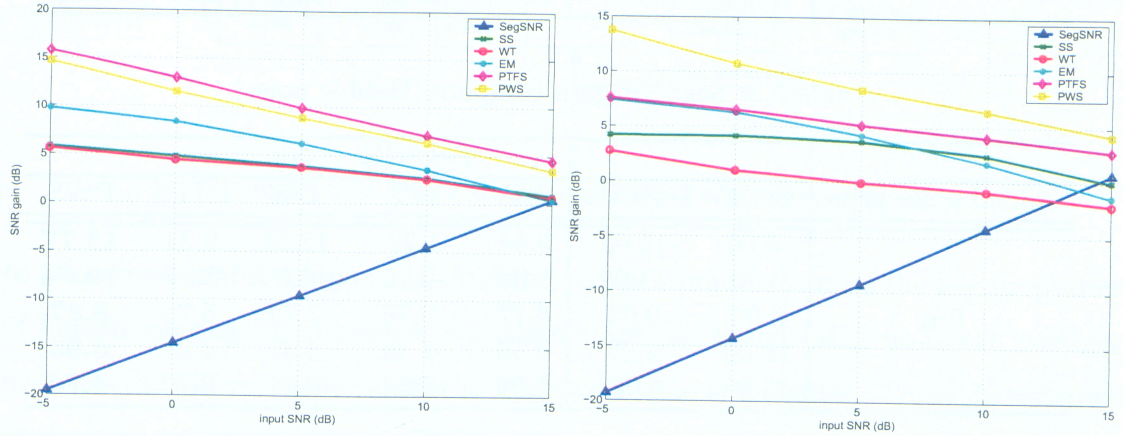
			SegSNR Gain				
Added noise type	SNR	SegSNR	SS	WT	EM	PTFS	PWS
AWGN	-5.00	-19.55	5.78	5.59	9.72	**15.75**	14.62
	0.00	-14.55	4.81	4.45	8.38	**12.98**	11.54
	5.00	-9.55	3.84	3.71	6.18	**9.91**	8.89
	10.00	-4.55	2.60	2.55	3.50	**7.09**	6.33
	15.00	0.45	0.83	0.64	0.41	**4.49**	3.50

Table 5.1: SegSNR gain compare (AWGN noise)

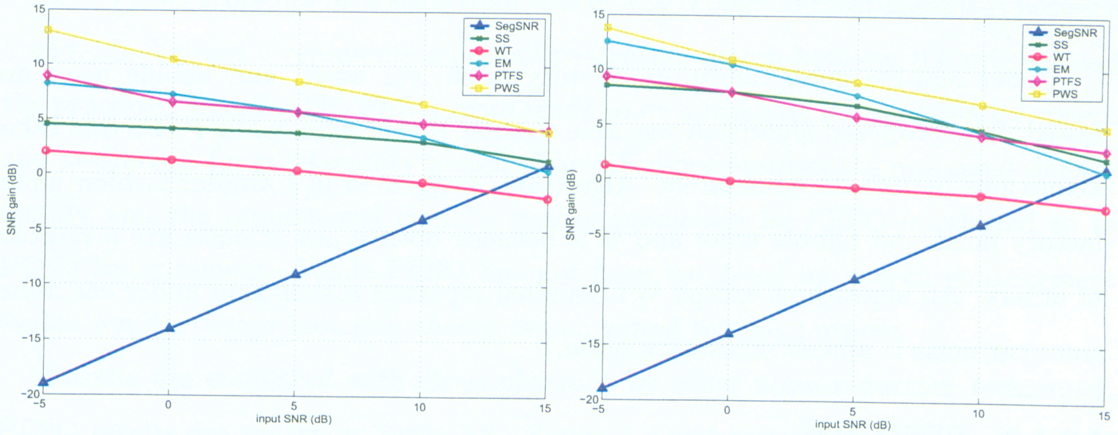
Table 5.1 shows under additive white Gaussian noise, proposed algorithm and original PTFS performed very well with PTFS outperforming by small margin. E-M algorithm achieved better results than WT and SS did. At high SNR algorithms SegSNR gain values converged, which showed most algorithms performed well under AWGN noise at high SNR (SNR>15 dB).

			SegSNR Gain				
Added noise type	SNR	SegSNR	SS	WT	EM	PTFS	PWS
Factory	-5.00	-19.30	4.16	2.72	7.38	7.54	**13.74**
	0.00	-14.30	4.11	0.95	6.23	6.49	**10.71**
	5.00	-9.30	3.68	-0.04	4.25	5.19	**8.41**
	10.00	-4.30	2.38	-0.89	1.66	4.06	**6.44**
	15.00	0.70	0.04	-2.09	-1.36	2.80	**4.25**

Table 5.2: SegSNR gain compare (Factory noise)



(a) SegSNR gain (AWGN noise) (b) SegSNR gain (Factory noise)



(c) SegSNR gain (Babble noise) (d) SegSNR gain (Car noise)

Figure 5.2: SegSNR gain comparison under different types of noise

			SegSNR Gain				
Added noise type	SNR	SegSNR	SS	WT	EM	PTFS	PWS
Babble	-5.00	-19.01	4.51	1.99	8.23	8.92	**13.07**
	0.00	-14.01	4.20	1.30	7.33	6.65	**10.50**
	5.00	-9.01	3.88	0.39	5.80	5.77	**8.55**
	10.00	-4.01	3.10	-0.63	3.53	4.79	**6.50**
	15.00	0.99	1.36	-2.02	0.44	4.16	**4.02**

Table 5.3: SegSNR gain compare (Babble noise)

			SegSNR Gain				
Added noise type	SNR	SegSNR	SS	WT	EM	PTFS	PWS
Car	-5.00	-19.02	8.54	1.22	12.58	9.34	**13.77**
	0.00	-14.02	8.00	-0.16	10.50	7.97	**10.94**
	5.00	-9.02	6.77	-0.76	7.72	5.75	**8.87**
	10.00	-4.02	4.58	-1.40	4.37	4.07	**6.95**
	15.00	0.98	1.91	-2.54	0.69	2.70	**4.70**

Table 5.4: SegSNR gain compare (Car noise)

Under real-life noise condition: factory noise, car noise, and babble noise are tested. The proposed algorithm was able to show higher gain, while PTFS and other algorithms performed moderately. Algorithms performed in a similar fashion under factory noise and babble noise and it is because both types of noise are a mixture of others. All algorithms except WT handled reported higher gain under car noise, since car noise is almost stationary noise.

5.1.1.2 PESQ Test

PESQ represents both the original and the degraded speech using psychophysical parameters. It is done by transforming the signals from the physical domain to the psychophysical domain through frequency warping and level compression. A distance measure is used to calculate the PESQ score [56]. Figure 5.3 presents the block diagram of PESQ algorithm. Signals under analysis are time aligned and level aligned

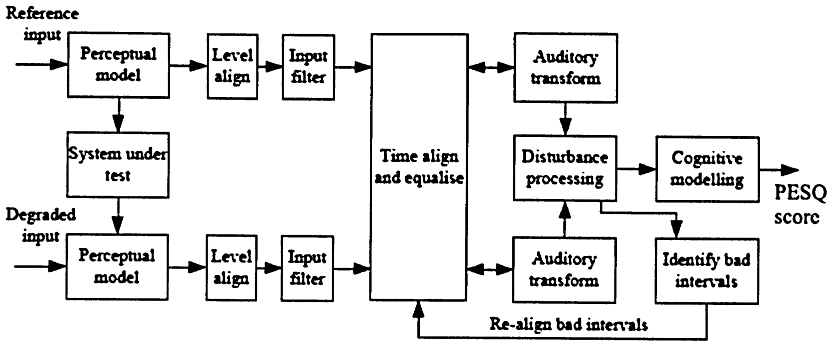


Figure 5.3: Block digram of PESQ algorithm [56]

to allow meaningful comparison. Auditory transformation mimics the key properties of human hearing. Disturbance parameters are calculated using no-linear averages over specific differences between the loudness of the two signals. These disturbance parameters are converted to a PESQ score.

The PESQ score lies on a scale from -0.5 to 4.5, though in most cases it is between 1 and 4.5. Value 4.5 represents a perfect perceptual match between the original and the processed signals. The PESQ score tends to be optimistic for poor quality speech and pessimistic for good quality speech. The difference between the processed signal's quality and the unprocessed signal's quality prediction by PESQ is referred to as PESQ Improvement. While PESQ has not been validated for speech enhancement evaluation, its design indicates that it is well suited for this purpose.

Results are compared with those obtained by other noise reduction techniques. PESQ results are shown in Table 5.8. Figure 5.4 elaborates the same result in bar chart.

With AWGN noise condition in Table 5.5, the proposed method consistently yielded better results than other techniques. Particularly at very low SNR condition (SNR<0 dB) where basic SS and WT methods failed to provide good improvement. At higher SNR, all techniques' performance improved as expected.

		PESQ Score					
Added noise type	SNR	Noisy	SS	WT	EM	PTFS	PWS
AWGN	-5	1.61	1.79	1.72	2.09	2.07	**2.30**
	0	1.87	2.07	2.01	2.26	2.33	**2.56**
	5	2.15	2.34	2.24	2.46	2.47	**2.81**
	10	2.42	2.55	2.51	2.60	2.76	**2.98**
	15	2.65	2.69	2.71	2.77	2.96	**3.13**

Table 5.5: PESQ results comparison (AWGN noise)

		PESQ Score					
Added noise type	SNR	Noisy	SS	WT	EM	PTFS	PWS
Factory	-5	1.92	1.95	2.13	**2.61**	2.31	2.35
	0	2.21	2.32	2.32	2.57	2.45	**2.61**
	5	2.41	2.54	2.50	2.71	2.48	**2.87**
	10	2.59	2.78	2.69	2.85	2.52	**3.11**
	15	2.81	3.00	2.87	2.97	2.78	**3.27**

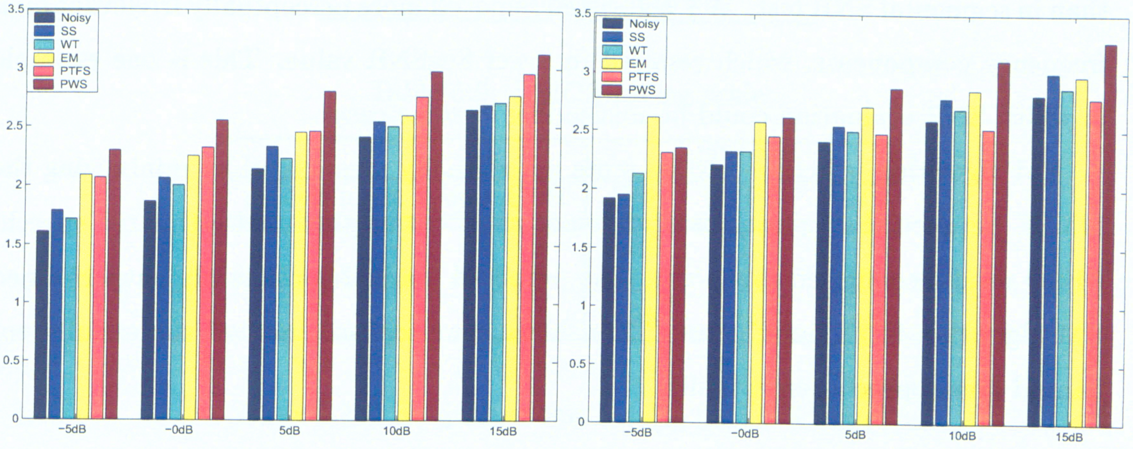
Table 5.6: PESQ results comparison (Factory noise)

		PESQ Score					
Added noise type	SNR	Noisy	SS	WT	EM	PTFS	PWS
Babble	-5	1.85	1.81	1.79	1.92	1.80	**2.12**
	0	1.70	1.91	1.86	2.05	1.93	**2.41**
	5	2.38	1.82	2.45	2.06	2.25	**2.69**
	10	2.59	2.62	2.64	2.66	2.65	**2.95**
	15	2.78	2.83	2.85	2.85	2.82	**3.15**

Table 5.7: PESQ results comparison (Babble noise)

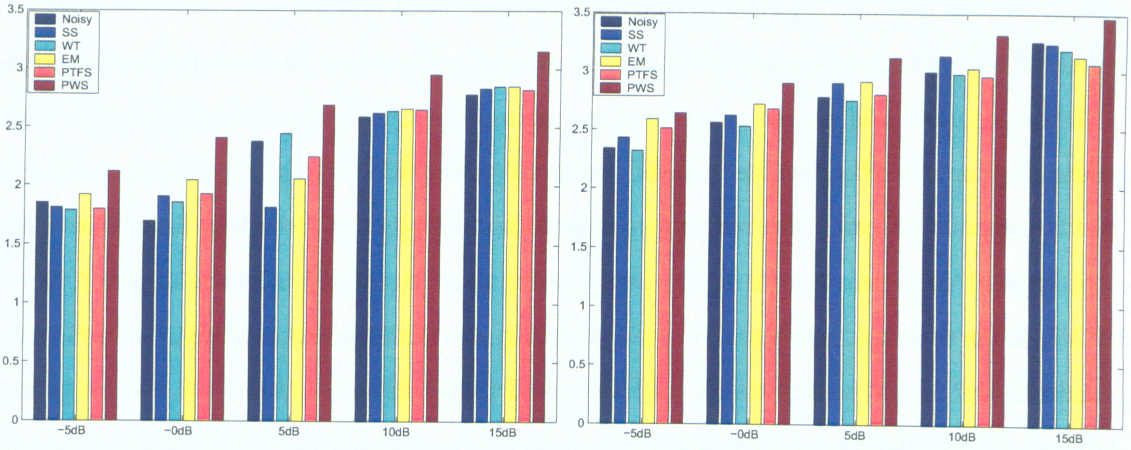
		PESQ Score					
Added noise type	SNR	Noisy	SS	WT	EM	PTFS	PWS
Car	-5	2.34	2.43	2.32	2.59	2.51	**2.64**
	0	2.56	2.62	2.53	2.72	2.68	**2.90**
	5	2.78	2.90	2.75	2.91	2.80	**3.12**
	10	3.00	3.14	2.98	3.03	2.96	**3.32**
	15	3.26	3.24	3.19	3.13	3.07	**3.47**

Table 5.8: PESQ results comparison (Car noise)



(a) PESQ score (AWGN noise)

(b) PESQ score (Factory noise)



(c) PESQ score (Babble noise)

(d) PESQ score (Car noise)

Figure 5.4: PESQ score comparison under different types of noise

In the real-life noise cases, PESQ results are shown in Tables 5.6, 5.7, and 5.8. The performance is the worst in cases of babble noise, since this noise has similar statistics as speech signal. The proposed method managed to retain gain in this situation. It is interesting to note that PESQ evaluated WT algorithm much better than in segmental SNR test. WT algorithm removed more perceptually irrelevant high frequency components, which resulted in lower SegSNR value. This is one example in which SNR evaluation could be misleading in some cases.

EM algorithm is believed to be more perceptually meaningful by minimizing the MMSE between the logarithms of the spectra of the original and estimated speech. PESQ results confirmed this point. The proposed method consistently outperformed EM algorithm in all noise cases. This shows the proposed method successfully enhanced noisy speech perceptually.

5.1.2 Subjective Quality Evaluation

Since SNR gain cannot faithfully reflect perceptual quality of the enhanced speech. Although more and more researchers are using PESQ for speech enhancement evaluation purpose, PESQ has not yet been validated for evaluating effects and artifacts generated from noise reduction algorithms. Subjective listening tests were done to confirm evaluation results. In this thesis, the Degradation Mean Opinion Score (DMOS) is used to do subjective evaluation. In the DMOS, listeners hear the reference and the test signal sequentially, and are asked to rate degradation level by comparing them. The DMOS provides greater sensitivity than the MOS, in evaluating speech quality, because the reference speech is provided.

Listeners are asked to give rating base on overall effect in terms of noise reduction, signal distortion, and sentence intelligibility. Grades of DMOS are provided in Table 5.9.

Eleven normal hearing subjects (5 male and 6 female) in the age group of 21 to 34 years participated in the experiments. All of them had no difficulty in clearly hearing

Rating	Degradation Level
5	Inaudible
4	Audible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Table 5.9: DMOS rating scale

		DMOS		
Added noise type	SNR	EM	PTFS	PWS
AWGN	0	**2.86**	2.14	2.71
	5	3.08	2.51	**3.21**
	10	3.35	3.00	**3.43**

Table 5.10: Subjective test results (AWGN noise)

the test stimuli. Test signals were produced by two males and two females, including 3 speech sentences and one singing sentence. The DMOS results are showed in Table 5.10 and Table 5.11.

By comparing subjective DMOS score with PESQ prediction, results in both tables are consistent with the perceived speech quality improvement trend. Under AWGN noise case, proposed PWS algorithm performed slightly better than the EM algorithm. At low SNR both performances are comparable. PTFS algorithm reported poorer performance than what was predicted by PESQ . Main reason could be listeners' dis-likeness of "muffled" speech. Even though PWS processed speech signals that

		DMOS		
Added noise type	SNR	EM	PTFS	PWS
Babble	0	2.24	2.06	**2.67**
	5	2.76	2.45	**3.16**
	10	3.02	2.56	**3.32**

Table 5.11: Subjective test results (babble noise)

contained some residual noise, listeners preferred these to EM processed ones.

The overall performances in babble noise case are poorer than in AWGN case for all algorithms as expected. Listeners consistently preferred PWS results in babble noise condition. Especially at lower SNR, PWS results are noticeably better.

From above SegSNR, PESQ, and DMOS evaluations, the proposed method has been shown to perceptually remove background noise and minimize speech distortion. It is able to improve quality of speech corrupted with strong color noise, under which conventional techniques normally fail.

5.2 Hearing Loss Compensation Evaluation

In order to compensate hearing loss, frequency spectrum shaping and dynamic range compression of the input signal need to be performed. Transparency of signals is not necessarily maintained in processed signal. The compressed speech does not sound quite "normal" to normal hearing person. PESQ is not directly applicable to measure speech quality, since it computes distance between input and output signal with fixed perceptual criteria. Only subjective listening test was done to evaluate hearing loss compensation performance.

The listening tests were conducted by normal hearing subjectives with masking noise to simulate hearing impairment. A subjective "A-B" preference listening test was done. Each listener is presented with signals with and without hearing loss compensation, and has to indicate which one is preferred. A no-preference choice is also allowed.

Table 5.12 and Table 5.13 show the "A-B" preference test results. WITH: signal processed with hearing loss compensation algorithm; W/O: signal without hearing loss compensation process.

In both additive noise conditions, more listeners favored hearing loss compensation. Especially at higher SNR, majority preferred with compensation. By observing results of different SNR condition, it is found out that signal distortion significantly

		A-B test		
Added noise type	SNR	preferred WITH	preferred W/O	No preference
AWGN	0	**47.2%**	38.9%	13.9%
	5	**61.1%**	27.8%	11.1%
	10	**83.3%**	13.9%	2.8%

Table 5.12: Subjective test results on compression (AWGN noise)

		A-B test		
Added noise type	SNR	preferred WITH	preferred W/O	No preference
Babble	0	**44.4%**	38.9%	16.7%
	5	**52.8%**	33.3%	13.9%
	10	**75.0%**	19.4%	5.6%

Table 5.13: Subjective test results on compression (Babble noise)

effects preference. One explanation is that the loudness compression algorithm is designed to take clean signal as input. Any distortion in the signal will be applied with some gain in the same way as speech components.

Overall, hearing loss compensation is effective provided the input is relatively free of distortion.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The main goal of this thesis is to improve the perceptual quality of degraded speech for hearing-impaired listeners when only one microphone is available. The study provided a working framework for speech enhancement for hearing-impaired person by incorporating psychoacoustics model and perceptual auditory features of hearing loss. The speech enhancement scheme developed can facilitate the design of hearing aids.

The noise reduction techniques proposed is based on the speech enhancement system by Li et al. [38] . The proposed algorithm employs a noise masking threshold (NMT) to balance the degree of noise suppression and speech distortion. By evaluating the proposed method subjectively and objectively, results showed improvement in many aspects. In SegSNR tests, the proposed method reported over 10 dB gain in low SNR condition (SNR$<$ 0 dB). PESQ results showed the proposed method received 10% higher score (average of all SNR conditions) than the next best performed technique in all noise cases. In subjective DMOS tests, the proposed method received 14% and 6% higher score (average of all SNR conditions) than EM algorithm in AWGN and babble noise respectively. For both white Gaussian noise and real-life noise conditions, the proposed method consistently outperformed other conventional techniques. Even under low SNR condition, the proposed method managed to give fair result while others failed.

The proposed hearing loss compensation scheme utilized critical-band loudness compression in wavelet domain. The processed speech is guaranteed to be within the residual dynamic range of the impaired listener. As a result, it could improve speech quality and intelligibility. Subjective tests showed strong preference for the compensated speech. When input signal is clean, benefit of loudness compression is apparent. About 83% of the subjects preferred hearing loss compensation in AWGN noise, and 75% in babble noise. As for the noisy signal case, input can be preprocessed by the proposed noise reduction algorithm.

The major contributions of this thesis include: Proposed a speech enhancement framework that would simultaneously perform noise reduction and hearing loss compensation. Cascading the two processing stages in the same domain is computationally efficient. The cleaner speech after noise reduction stage could further enhance the performance of hearing loss compensation.

- Developed a wavelet based critical band noise reduction strategy utilizing psychoacoustics model. Speech processed by the new algorithm showed reduced level of residual noise and better perceptual quality.

- Developed a wavelet based loudness compression algorithm. Compression gain is calculated in each critical band rather than in linear subbands. Algorithm parameters can be easily modified according to specific individual's profile.

- Detailed performance evaluation with comparison with other techniques. Both subjective and subjective correlated objective tests were performed. Tests from both categories showed consistent results, therefor the evaluations are reliable.

6.2 Future Work

There are a number of potentially promising further developments of the proposed method in this thesis. These include:

- To better represent impaired cochlear frequency resolution, equivalent rectangular bandwidths (ERB) instead of critical-band width is worth investigating.
- The possibility of estimating the NMT considering elevated hearing threshold of impaired person, i.e. customizing the estimation of the NMT to each hearing-impaired individual.
- Formulate an algorithm to estimate the NMT without relying on the pre-estimated speech.
- The techniques could be extended to the wider-band audio signal.

Bibliography

[1] J. D. Jr, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals.* Prentice Hall, New Jersey, 1987.

[2] K. Fellbaum, "Human speech production," www.kt.tu-cottbus.de/speech-analysis/tech.html, Tech. Rep., 2006.

[3] E. Zwicker and H. Fastl, *Psychoacoustics: facts and models.* Springer Verlag, 1999.

[4] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Am.*, vol. 71, pp. 679–688, 1982.

[5] S. Handel, *Listening: An Introduction to the Perception of Auditory Events.* Cambridge, MA: MIT Press, 1989.

[6] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. Academic Press, 1997.

[7] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.*, vol. 68, pp. 1523–1525, 1980.

[8] H. Levitt, "Noise reduction in hearing aids: A review," *Journal of Rehabilitation Research and Development*, vol. 38, pp. 111–121, 2001.

[9] B. C. J. Moore, "Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids," *Ear & Hearing.*, vol. 17(2), pp. 133–161, 1996.

[10] N. Fan, R. Balan, and J. Rosca, "Comparison of wavelet- and FFT-based single-channel speech signal noise reduction techniques," *Intelligent Systems in Design and Manufacturing Proceedings of the SPIE*, vol. 5607, pp. 127–138, Nov. 2004.

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

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

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

[14] N. Whitmal III, "Wavelet-baed noise reduction for speech enhancement," Ph.D. dissertation, Northwestern university, USA, 1997.

[15] S. Ogata and T. Shimamura, "Reinforced spectral subtraction method to enhance speech signal," *Electrical and Electronic Technology, 2001. TENCON*, vol. 1, pp. 242–245, 2001.

[16] L. Singh and S. Sridharan, "Speech enhancement using critical band spectral subtraction," in *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, Dec. 1998.

[17] R. Nishimura, F. Asano, Y. Suzuki, and T. Sone, "Speech enhancement using spectral subtraction with wavelet transform," *Electronics communication Japan, Part III fundam electron sci.*, vol. 81, pp. 24–31, 1998.

[18] N. Virag, "Speech enhancement based on masking properties of the auditory system . vol. 1, pp. 796-799. 1995," *ICASSP'95*, vol. 1, pp. 796–799, 1995.

[19] ——, "Single channel speech enhancement based on masking properties ofthe human auditory system," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, pp. 126–137, 1999.

[20] D. E. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, "Perceptual filters for audio signal enhancement," *J. Audio Eng. Soc,*, vol. 45, pp. 22–36, 1997.

[21] D. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 497–514, 1997.

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

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

[24] T. Gülzow, A. Engelsberg, and U. Heute, "Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement," *Signal processing*, vol. 64, pp. 5–19, 1998.

[25] C. Lu and H. Wang, "Enhancement of single channel speech based on masking property and wavelet transform," *Speech Communication*, vol. 41, pp. 409–427, 2003.

[26] Y.Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 59- 67, 2004.

[27] A. Teolis, *Computational Signal Processing with Wavelets.* Boston: Birkhauser, December 1996.

[28] C. Valens, *A really friendly guide to wavelets.* citeseer.ist.psu.edu/valens99really.html.

[29] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. on Pattern Anal. Machine Intell.*, vol. 11, pp. 674–693, 1989.

[30] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to wavelets and wavelet transforms: a primer.* Prentice-Hall International, 1998.

[31] R. R. Coifman, Y. Meyer, and M. V. Wickerhauser, "Wavelet analysis and signal processing," in *Wavelets and Their Applications.* Boston: Jones and Bartlett, 1992, pp. 153–178.

[32] S. Mallat, *A Wavelet Tour of Signal Processing, 2nd edition.* Academic Press, 1999.

[33] D. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, pp. 613–627, 1995.

[34] D. Donoho and Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika, Cambridge*, vol. 81, pp. 425–455, 1994.

[35] S. Chang, Y. Kwon, S. Yang, and I. Kim, "Speech enhancement for non-stationary noise environment by adaptive wavelet packet," *Proceedings of the 2002 IEEE International Conferenceon Acoustics, Speech, and Signal Processing (ICASSP 2002)*, vol. 1, pp. 561–564, 2002.

[36] J. Seok and K. Bae, "Speech enhancement with reduction of noise components in the wavelet domain," *Proceedings of the 1997 IEEE International Conference*

on Acoustics, Speech, andSignal Processing (ICASSP '97), vol. 2, pp. 1323–1326, 1997.

[37] H. Sheikhzadech and H. Abutalebi, "Improved wavelet-based speech enhancement system," *in Proc. 7th European Conference on Speech Communication and Technology (EuroSpeech), Aalborg, Denmark*, 2001.

[38] M. Li, H. G. McAllister, N. D. Black, and T. A. D. Perez, "Perceptual time-frequency subtraction algorithm for noise reduction in hearing aids," *IEEE Transactions on Biomedical Engineering*, vol. 48, pp. 979–988, 2001.

[39] C. Lu and H. Wang, "Speech enhancement using robust weighting factors for critical-band-wavelet-packet transform," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*, vol. 1, 2004, pp. 721–724.

[40] S. Chen and J. F. Wang, "Speech enhancement using perceptual wavelet packet decomposition and teager energy operator," *The Journal of VLSI Signal Processing*, vol. 36, pp. 125–139, 2004.

[41] P. Srinivasan and L. Jamieson, "High-quality audio compression using an adaptive wavelet packet decomposition and psychoacoustic modeling," *IEEE Transactions on Signal Processing*, vol. 46, pp. 1085–1093, 1998.

[42] Y. Hu and P. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," in *IEEE Signal Processing Letter*, vol. 11, 2004, pp. 270–273.

[43] B. Carnero and A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronizedfast wavelet packet transform algorithms," *IEEE Transactions on, Signal Processing*, vol. 47, pp. 1622–1635, 1999.

[44] M. Marzinzik, "Noise reduction schemes for digital hearing aids and their use for the hearing impaired," Ph.D. dissertation, University Oldenburg, Germany, 2000.

[45] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," *International Conference on Acoustics, Speech, and Signal Processing 1995*, vol. 1, pp. 153–156, 1995.

[46] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. of the IEEE conference on Acoustics, Speech and Signal Processing*, 1979, pp. pp.208–211.

[47] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal Select. Areas Commun.*, vol. Vol. 6, pp. 314–323, Feb. 1988.

[48] D. Sinha and A. H. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. Signal Processing*, vol. 41, pp. 3463–3479, Dec. 1993.

[49] I. M. Johnstone and B. W. Silverman, "wavelet threshold estimators for data with correlated noise," *Journal of the Royal Statistical Society B*, vol. 59, pp. 319-351, 1997.

[50] J. Rutledge, *Time-Frequency and Wavelets in Biomedical Signal Processing*, M. Akay, Ed. IEEE Press, 1997.

[51] M. Li, H. G. McAllister, N. D. Black, and T. A. D. Perez, "Wavelet-based nonlinear agc method for hearing aid loudness compensation," *Vision, Image and Signal Processing, IEE Proceedings*, vol. 147-6, pp. 502–507, 2000.

[52] E. Villchur, "Signal processing to improve speech intelligibility in perceptive deafness," *Journal of the Acoustical Society of America*, vol. 53, pp. 1646–1657, 1973.

[53] L. A. Drake, J. C. Rutledge, and J. Cohen, "Wavelet analysis in recruitment of loudness compensation," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3306–3312, 1993.

[54] D. Kim, Y. C. l Park, W. K. Kim, S. Park, W. Doh, S. W. Shin, and D. H. Youn, "Simulation of hearing impairment with sensorineural hearing loss," *International Conference of the IEEE, Engineering in Medicine and Biology society*, vol. 5, pp. 1986–1989, 1997.

[55] J. Chalupper and H. Fastl, "Simulation of hearing impairment based on the fourier timetransformation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 857–860, 2000.

[56] *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. (2001), ITU-T P.862, ed., International Telecommunication Union., Geneva, 2001.

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