MARQUETTE UNIVERSITY

Auditory Model-based Bionic Wavelet Transform

For Speech Enhancement

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree of

MASTER OF SCIENCE

Field of Electrical and Computer Engineering

by

Xiaolong Yuan, B.S.E.E.

Speech and Signal Processing Lab
Milwaukee, Wisconsin
May 2003

To Virginie Striebel

Abstract

In many speech processing applications, speech has to be processed in the presence of undesirable background noise like white Gaussian noise, colored noise, multi-talker babble noise, car interior noise and so on. Various methods have been applied for the suppression of ambient noise while minimizing the extent of speech distortion. Bionic wavelet transforms are a new time-frequency analysis method recently proposed in the biomedical engineering field (Yao 2001), based on an active cochlear model, which have proven to be meaningful in cochlear implant research. The standard wavelet transform resembles, to some degree, the way that the front end human auditory system, mainly the passive cochlea, processes speech, yet the bionic wavelet transform promises even more similarity to the active cochlea and hence more flexibility and efficiency. Therefore applying this method to speech enhancement may lead to a promising future in this field.

Spectral subtraction methods have been widely used in speech enhancement, but all are notorious for unexpected music tone artifacts. Wiener filtering and Ephraim Malah filtering methods have achieved good performance dealing with white Gaussian noise at different SNR levels. Wavelet-based methods using thresholding techniques are promising for coping with real life noise of various kinds. However many improvements have yet to be made to render this approach more flexible and robust. The bionic wavelet transform (BWT) was proposed in 2001 and to date no one has yet introduced it into speech enhancement other than cochlear implant research. Due to the integration of human auditory system model into the wavelet transform, the BWT has great potential in speech enhancement and may lead to a new path in wavelet-based speech processing. In the thesis, basic spectral subtraction, iterative Wiener filtering, Ephraim Malah filtering and traditional wavelet thresholding techniques have been

used as baseline methods for speech enhancement tests. Segmental signal-to-noise ratio (SSNR) and signal-to-noise ratio (SNR) are adopted for the objective speech quality evaluation and the mean opinion score (MOS) subjective measure is also employed. Bionic wavelet based thresholding is then implemented to enhance speech quality and comparisons are made among the performances of 5 different enhancement methods. Bionic wavelet-based thresholding showed significant advantage over traditional thresholding and Ephraim Malah filtering proved the best among all five algorithms.

Acknowledgements

I would like to thank my advisor Dr. Michael Johnson for entertaining my interests in speech signal processing and giving me the opportunity to explore speech enhancement research. He is the one who has been leading me into the wonderland of speech research and along the way he has always been supportive in giving advice and directions as I am learning about my ignorance of this world. He is the one who conscientiously helps me out of those baffling technical details and his down-to-earth attitude towards the quest of truth is always an example for me. I am also indebted to Dr. Kristina Ropella's biomedical signal processing courses that lay a sound foundation of my research attempt on wavelet analysis. She leads me to understand how important it is to broaden one's views in related fields when conducting interdisciplinary research. Special thanks go to Patrick Clemens who kindly helps me out in software related troubleshooting. I also owe a great deal to Dr. Jun Yao from Northwestern University who provided generous input on my understanding of the bionic wavelet transform.

The production of this work is an academic summary of my two years' graduate school life in Marquette. On the other side, several special friends made my two-year overseas life experience full of meanings. May this thesis a dedication to Virginie Striebel, Franz Glatzl, Mathieu Bohnert, and Mahmood Zia Khan.

I would also like to express my deepest gratitude to my parents Mr. Zhirun Yuan and Mrs. Qiuming Jee and my sister Rui Yuan, who provided love, support and encouragement.

Lastly but not the least, this thesis would never have been conceived without the faculty, students and administrators of the Department of Electronic and Computer Engineering at Marquette University.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Communication via speech is one of the essential functions of human beings. Humans possess varied ways to retrieve information from the outside world or to communicate with each other, and the three most important sources of information are speech, images and written text. For many purposes, speech stands out as the most efficient and convenient one. Speech not only conveys linguistic contents, but also communicates other useful information like the mood of the speaker. Language communication through speech is closely intertwined with the evolution of human civilization.

Speech processing is an interdisciplinary field that studies acoustic signals using both signal processing techniques and knowledge from hearing science, phonetics, linguistics, and psychology. Thanks to the explosive advances in digital signal processing, the ease and speed of representing, storing, retrieving and processing speech data has boosted the development of speech processing techniques to address different application areas: speech enhancement, speech synthesis, speech coding and speech recognition.

One of the most important branches of speech processing, speech enhancement focuses on finding an optimal (i.e., preferred by a human listener) estimate $\hat{s}(t)$, given a noisy measurement $y(t) = s(t) + n(t)$ (Van Compernolle 1993). Application areas include the reduction of noise for hearing purposes, the preprocessing of speech coding or recognition systems and hearing aid research. Noise reduction or speech enhancement has always been a non-trivial problem for engineers. The total removal of background noise is practically impossible and the distortion of the speech content is inevitable. The ease of implementation is

another consideration when real-time on-site performance is expected, for example in hearing aid devices.

Traditional speech enhancement algorithms include spectral subtraction, Wiener filtering, time varying speech model-based or state-based methods, and microphone array based techniques. The wavelet transform distinguishes itself in the analysis of non-stationary signals such as speech. In 1995, Donoho introduced wavelet shrinkage as a powerful tool in denoising signals corrupted by additive white noise. Wavelet shrinkage employs non-linear thresholding in the wavelet domain and has proved broad asymptotically near-optimal properties for a wide class of signals corrupted by additive noise (Donoho 1995). Recent efforts made for speech enhancement using wavelet shrinkage include those of (Seok 1997),(Chang 2002), (Bahoura 2001),(Sheikhzadeh 2001).

In this thesis, our investigation is focused on the integration of the bionic wavelet transform into wavelet based denoising for the single-channel noise reduction of speech corrupted by broadband white noise, pink noise, plane cockpit noise, car interior noise, and multiple talkers' noise. We are the first to introduce the auditory-based bionic wavelet transform into speech enhancement. The motivations are two fold. First, wavelet-based denoising method has proved its efficiency in speech enhancement. Secondly, the bionic wavelet transform, which adapts itself to the time varying speech signal, better emulates the front-end human auditory periphery than the standard wavelet transform and has better energy concentration and time-frequency trade-off.

This thesis is organized as follows: Chapter 2 gives a review of the state-of- the-art spectral based speech enhancement strategies that have been developed to date. Chapter 3 discusses the

wavelet based algorithm. The new algorithm with results and quantitative performance comparison is presented in Chapter 4. Chapter 5 gives summaries and future explorations.

Chapter 2 Background

For decades, speech enhancement algorithms have been mostly focused on additive noise removal (Deller 1994), although there are other non-additive noises involved such as multiplicative noise, reverberation and channel or speaker interference. Most speech enhancement algorithms attempt to achieve a compromise between the extent of noise suppression and the distortion of the speech contents, considering the fact that both ambient noise and the speech itself have great unpredictability in their signal characteristics. For the single channel case, where only a single microphone is available, methods have capitalized on the use of perceptual or statistical constraints and reached varying degrees of success, in terms of the goals and assumptions being made for each approach (Deller 1994).

This chapter first reviews the production of speech and its source/filter model and the characteristics of different noises that interfere with noises. Then state-of-the-art single-channel speech enhancement techniques are presented, and wavelet-based methods are discussed in more detail.

2.1 The production of speech and its acoustic aspects

Acoustic speech is commonly regarded as resulting from a combination of a source of sound energy (e.g., the larynx) modulated by a transfer function (filter) determined by the shape of the vocal tract. This model is often referred to as the *source-filter theory of speech production* and stems from the experiments of Johannes Müller (1848) in which a functional theory of phonation was tested by blowing air through larynx excised from human cadavers (Rubin

1998).For the unvoiced part of speech, the source can be seen as white noise while for the voiced part of speech, pulse train acts as the energy source. In Figure 1, A_N , A_V are the gains for the energy sources of the unvoiced and voice speech, respectively.

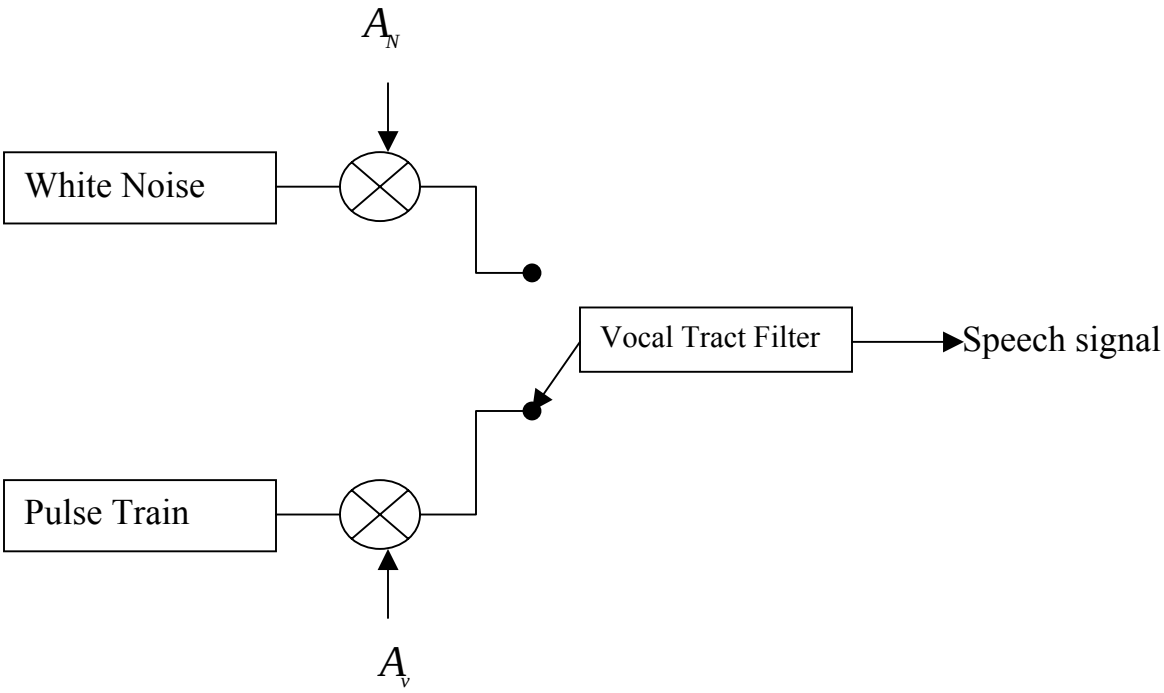

Figure 1 The source-filter model of speech production

Normally, for short segments or frames of speech (10~30 ms) the shape of the vocal tract remains relatively the same resulting the valid use of a linear time-invariant filter. Therefore the speech signal can be seen as a stationary random process, which allows the use of the short time Fourier Transform (STFT) or other stationary analysis techniques.

2.2 A Summary of Single Channel Speech Enhancement Approaches

The main purpose of speech enhancement is to reduce (additive) noise while minimizing the degree of distortion of the desired speech signal. For single-channel spectral-based techniques,

there are generally four approaches to separate speech from noise, although some methods tend to have combined characteristics (Lim 1979).

2.2.1 Methods Using Periodicity of the Voiced Speech

Some enhancement methods capitalize on the observation that waveforms of voiced sounds are periodic with a period that corresponds to the fundamental frequency (Lim 1979). Two typical methods are presented as follows.

- *Adaptive Comb Filtering*

Normally the energy of the speech is concentrated in narrow bands of frequency whereas the noise has energy spread across the whole spectrum. Comb filtering (Shields 1970) passes the harmonics of speech but rejects the frequency content between those harmonics, which are mainly from the noise. Due to the fact that pitch information (the perceived fundamental frequency) varies from speaker to speaker or even within speech from a single speaker, we require the comb filter to be adaptive, matching the changing pitch of the input waveform.

- *Adaptive Noise Cancellation Techniques*

ANC (Adaptive Noise Cancellation) is useful for processing speech that results from what is traditionally referred to as the "two microphone" problem. It assumes speech is corrupted by uncorrelated noise and that a reference noise source is available, which is correlated with the noise to be reduced. Adaptive filtering techniques are employed to reduce noise including LMS (linear mean square) algorithm (Haykin 1986) and steepest descent algorithm (Haykin 1986) that requires no *a priori* knowledge of the statistics of the noise In those cases where the reference noise channel is not available, we generate the reference noise input as follows.

$$y(n) = s(n) + d(n) \tag{1.1}$$

where $y(n), s(n), d(n)$ represent corrupted speech, clean speech, and additive uncorrelated noise, respectively. After delaying one pitch period of the voiced speech, we then subtract out the speech content and obtain the reference noise approximately from $r(n)$ where $r(n)$ is given by

$$r(n) = y(n) - y(n+T) \tag{1.2}$$

But still we face the same problem as comb filtering: do we have accurate pitch information? Although linear prediction, autocorrelation, and cepstrum analysis are all effective, to various extents, in pitch extraction from clean speech signals, we still lack powerful solutions to extract precise pitch information from degraded speech.

The drawbacks of those methods just described are that the intelligibility of the processed speech is reduced over a wide range of signal-to-noise ratios when white noise or a competing talker is involved (Lim 1979).

2.2.2 Methods Based on Formant

Formants refer to the resonant frequencies of the vocal tract. At different times in speech production, the vocal tract, as determined by the position of the articulators, assumes different shapes with each shape corresponding to a set of formant frequencies. The formants in the spectrum are usually denoted F1, F2… beginning with the lowest frequency and although there are actually a large number of formants in a given sound, we normally find 3-5 formants after Nyquist sampling (Deller 1994).

The purpose of the formant enhancement process is to re-introduce the formant peaks in the spectral envelope and to suppress the noise at high frequency (Branson 1997). The cesptral speech enhancement algorithm developed by Conway(Branson 1997) from the Marquette

University speech laboratory employs formant enhancement, pitch enhancement and energy contour enhancement. A formant tracking function is used to estimate the formant frequencies and the formant peaks are re-introduced through the log magnitude spectral envelope. The amplitudes and bandwidths of the three formants peaks are computed using the estimated frequencies of the first three formants and their corresponding confidences. The spectral envelop of the enhanced speech signal is then obtained through the sample-by-sample product of the formant enhancement function and the noisy spectral envelop (Branson 1997).

A fair amount of work has been done on feature-based speech intelligibility enhancement in high noise levels (Conway 1994). The speech features used include the formant frequency, pitch frequency and voiced/unvoiced decision (Niederjohn 1995). This method demonstrates that with adequate extracted speech features, a significant intelligibility enhancement of noise-corrupted speech is possible although with a deterioration in speech quality (Conway 1990).

2.2.3 Methods Based on Subspace

The basic idea of subspace method is to decompose noisy speech into the signal subspace (signal plus noise) and the noise subspace (only noise) as follows:

$$\mathbf{R_x} = \mathbf{Q}\boldsymbol{\Lambda}_\mathbf{x}\mathbf{Q}^\mathbf{H} = [\mathbf{Q_1} \quad \mathbf{Q_2}] \begin{bmatrix} \boldsymbol{\Lambda}_\mathbf{x} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Q_1^H} \\ \mathbf{Q_2^H} \end{bmatrix} \tag{1.3}$$

where $\mathbf{R_x}$ is the noisy correlation matrix, $\boldsymbol{\Lambda}_\mathbf{x} = \mathrm{diag}(\lambda_{x1,}...,\lambda_{xK})$ denoting non-zero

eigenvalues of $\mathbf{R_x}$, $\mathbf{Q_1}$ is the signal subspace and $\mathbf{Q_2}$ is the noise subspace with $(\bullet)^H$ denoting vector conjugate transpose.

Removing the noise subspace as well as estimating the clean speech from the signal space will lead to speech enhancement. The Karhunen-Loeve expansion (KLT), a theoretically optimal transform, serves as the decomposition tool and in practice, the discrete Fourier transform (DFT), discrete cosine transform (DCT), discrete Hartley transform (DHT), and discrete W transform are used to approximate the KLT.

Linear estimation of the clean speech from the signal subspace alone is the core of the subspace speech enhancement. The estimation is performed on a frame-by-frame basis assuming additive noise uncorrelated with speech. The residual signal represents signal distortion and residual noise. As two factors cannot be simultaneously minimized, how to minimize them in an optimal way leads to two basic routes (Ephraim 1995).

- *Time domain constraints*

Defining the energies of the signal distortion as ε_x^2 and the energies of residual noise as ε_n^2, we obtain the optimum linear estimator by solving the following time-domain constrained optimization problem:

$$\min_H \varepsilon_x^2 \text{ subject to: } \quad \frac{1}{k}\varepsilon_n^2 \le \alpha\sigma^2 \text{ where } 0 < \alpha < 1 \tag{1.4}$$

- *Spectral domain constraints*

Suppose that k-th spectral component of the residual noise is given by $v_k^T \varepsilon_n$ where v_k^T is the k-th column vector of the eigenvector matrix $\mathbf{\Sigma} = \mathbf{R_n^{-1}R_x}$ with $\mathbf{R_n}$ denoting the covariance matrix of the noise:

$$\min_H \varepsilon_x^2 \text{ subject to : } \quad E\{|v_k^T \varepsilon_n|^2\} \le \alpha_k \qquad \text{when k=1,…,M} \tag{1.5}$$

$$E\{|v_k^T \varepsilon_n|^2\} = 0 \qquad \text{when k=M+1,…,K} \tag{1.6}$$

The general implementation steps of the subspace method include:

1) Compute the covariance matrix of the noisy signal $\mathbf{R}_y$ and the covariance matrix of the estimated noise $\mathbf{R}_n$ and then estimate the matrix $\sum$ where

$$\sum = \mathbf{R}_n^{-1}\mathbf{R}_x = \mathbf{R}_n^{-1}(\mathbf{R}_y - \mathbf{R}_n) = \mathbf{R}_n^{-1}\mathbf{R}_y - \mathbf{I} \tag{1.7}$$

where the underlying assumption is that the speech signal and the noise are uncorrelated.

2) Assuming the eigenvalues of $\sum$ are ordered as $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_k$ to estimate the dimension of the speech signal subspace as follows:

$$M = \arg\max_{1 \leq m \leq k} \{\lambda_m > 0\} \tag{1.8}$$

3) Using different optimal linear estimators obtained through corresponding time or spectral constraints, we solve for $\mathbf{H}_{\mathbf{optimal}}$.

4) Estimate the enhanced speech through $\hat{\mathbf{X}} = \mathbf{H}_{\mathbf{opt}}\mathbf{Y}$.

2.2.3 Methods Using the Short Time Spectral Amplitude (STSA) of Speech

Spectral subtraction and Wiener filtering are two typical methods that fall into this category of speech enhancement, (Ephraim 1984) and we provide a detailed discussion on each in the following sections. They both require noise estimation from non-speech segments. These methods are motivated from the fact that in terms of speech intelligibility and quality, short-time spectral amplitude (STSA) information is more important than the phase information. The spectra of the background noise and of the noisy speech are estimated in order to reconstruct the enhanced speech. Ephraim Malah filtering, elaborated in Section 2.5,

capitalizes on the minimum mean square (MMSE) STSA estimator to provide enhanced speech, with a significant reduction of the noise.

2.3 Spectral Subtraction

Spectral subtraction (Boll 1979) has long been the most commonly adopted method due to its simplicity and effectiveness in reducing broadband and stationary additive background noise. As human perception is far more sensitive to the magnitude of speech than the phase property, basic spectral subtraction simply requires that an estimate of the noise magnitude spectrum is subtracted from the instantaneous input magnitude spectrum. The greatest asset of spectral subtraction exactly lies in its simplicity since all that is required is an estimate of the mean noise power, and most importantly it does not need any assumptions of the speech signal. Also spectral subtraction is computationally efficient, demanding about the same computation intensity as a high-speed convolution (Boll 1979).

2.3.1 Basic Spectral Subtraction

We assume speech is short-time stationary, and that the noise $n(t)$ is uncorrelated with the speech signal $s(t)$. Then the corrupted speech $y(t)$ can be represented as:

$$y(t) = s(t) + n(t) \tag{2.1}$$

where $Y(\omega), S(\omega), N(\omega)$ represent the Discrete Fourier Transform of $y(t), s(t), n(t)$, respectively. The frequency domain version of Eq. 2.1 is as follows:

$$Y(\omega) = S(\omega) + N(\omega) \tag{2.2}$$

$$|Y(\omega)|^2 = |S(\omega)|^2 + |N(\omega)|^2 \tag{2.3}$$

Due to the assumption of uncorrelated noise and speech, there are not any cross power terms of speech and noise. We can estimate the power spectrum of clean speech through the estimation of noise power $|\hat{N}(\omega)|^2$ due to the assumption that noise is pseudo stationary so the power spectrum in the silence region is the same as that in the speech region. We obtain the spectral power estimate as follows:

$$|\hat{S}(\omega)|^2 = |Y(\omega)|^2 - |\hat{N}(\omega)|^2 \tag{2.4}$$

where ^ indicates the estimated value. Typically, if we have negative estimated values of the clean speech power spectrum, the values are set to zero, which is referred to as half wave rectification.

Past research shows that the presence of noise in the phase information does not contribute immensely to the degradation of the speech quality, especially if SNR is greater than 5dB (Schroeder 1975) while at SNR<0 the noisy phase can lead to a perceivable roughness in the speech contributing to the reduction of speech quality. However, phase estimation of clean speech is not a trivial problem and the magnitude spectrum presents a far bigger threat to speech quality distortion than phase estimation (Boll 1979). Hence, normally we simply use the phase information of the corrupted speech for the reconstruction of the enhanced speech.

$$S(\omega) = |S(\omega)| \exp(j\phi_y(k)) \tag{2.5}$$

2.3.2 Musical Tones

It is obvious that the estimate of the clean speech relies on obtaining an accurate spectral estimate of the noise signal. However, since the noise spectrum is not available in practice, an averaged estimate of the noise is significantly different than the actual noise contents in the instantaneous speech spectrum. If we just simply subtract a smoothed estimate of noise spectrum, it causes the presence of some sinusoidal energy at various frequencies in the

enhanced speech, which sounds like synthetic musical tones. This metallic sound distracts the attention of the listener, and greatly affects the quality of the enhanced speech.

2.3.3 Modified Spectral Subtraction

Broadband noise spreads across the whole spectrum and its energy remains constant, while voiced sounds such as vowels tend to have more energy concentration than unvoiced sounds like fricatives. It is therefore sensible to subtract the noise power differently on a frame-by-frame basis. (Berouti 1979) introduced an over-subtraction factor β ($\beta \geq 1$) that varies frame-by-frame according to the signal-to-noise ratio and also a preset spectral flooring $\eta \,|\, \hat{N}_i(\omega) \,|^2$ ($0 \leq \eta \leq 1$) after subtraction.

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \beta \,|\hat{N}(\omega)|^2 & \text{if } |\hat{S}(\omega)| > \eta \,|\hat{N}(\omega)|^2 \\ \eta \,|\hat{N}(\omega)|^2 & \text{otherwise} \end{cases} \tag{2.6}$$

This weighted subtraction method has been shown to reduce musical tones significantly with very little compromise on the intelligibility of the speech (Berouti 1979). To make spectral subtraction even more flexible, an adjustable parameter α has been included:

$$|\hat{S}(\omega)|^{\alpha} = |Y(\omega)|^{\alpha} - \beta \,|\hat{N}(\omega)|^{\alpha} \tag{2.7}$$

Past experiments show that adjustments of α and β achieve better enhancement performance than basic spectral subtraction.

2.3.4 Limitations of Spectral Subtraction

Besides the fact that musical tones are impossible to remove completely, the main limitation to spectral subtraction lies in the degradation of the intelligibility of the enhanced speech, especially at low SNR levels (Boll 1979). Efficiency could be improved if we can significantly improve the performance of the speech silence detection algorithm. Taking into account the

possibility of non-stationarity of the noise signal, this method can be extended to non-stationary noise suppression using adaptive spectral noise estimation.

2.4 Wiener filtering

Under the assumption of stationary noise uncorrelated with the target speech signal, we have an alternative stochastic optimization method to suppress noise, based on minimizing the mean square error between estimated object signal value $\mathrm{Y}(\omega)$ and the original signal value $\mathrm{S}(\omega)$. The formulation of the optimal Wiener filter is as follows.

$$H_y(\omega) = \frac{S_s(\omega)}{S_s(\omega) + S_n(\omega)} \tag{2.8}$$

where $S_s(\omega)$, $S_n(\omega)$ represent the estimated power spectra of the object signal and the background noise, which are assumed uncorrelated and stationary. In the pseudo stationary case of speech, we again resort to the frame based analysis where for each frame, the transfer function of the Wiener filter is calculated and the speech is recovered through:

$$\hat{S}(\omega) = Y(\omega)\hat{H}_y(\omega) \tag{2.9}$$

As the Wiener filter is non-causal and zero-phase, the original phase of $Y(\omega)$ is again employed in the reconstruction. Obviously the Wiener filter requires prior knowledge of both speech and noise statistics and they have to be estimated in real practice.

Similar to modified spectral subtraction in Eq.2.7, Wiener filters have the following generalized form:

$$H_y(\omega) = \left(\frac{S_s(\omega)}{S_s(\omega) + \beta S_n(\omega)} \right)^{\alpha} \tag{2.10}$$

where $H_y(\omega)$ represents the Wiener filter transfer function, $S_s(\omega)$ is the power spectrum estimate of the speech of each frame, and $S_n(\omega)$ is the noise power estimate of the noise, and β is the noise suppression factor.

2.4.1 Iterative Wiener filtering

In the single-channel case, noise statistics normally have to be estimated during silence frames and *a prior* knowledge of the speech statistics have to be obtained via an iterative estimation process. Lim and Oppenheim(Lim 1979) proposed the estimation of speech parameters in an all–pole model corrupted by white Gaussian noise. And later Hansen and Clements (Hansen 1985) generalized for the colored noise case.

Speech can be modeled as a random process using autoregressive–moving average (ARMA), autoregressive (AR) or moving average (MA) models. Although a pole-zero modeling of the vocal tract is closer to the truth of our vocal tract, the all-pole model offers an advantage in terms of computation. Lim and Oppenheim (Lim 1979) applied maximum *a posteriori* estimation (MAP) to maximize the *a posteriori* probability density of the linear prediction coefficients vector $\mathbf{a}$, given the noisy speech vector $\mathbf{y}$ for each speech frame:

$$\hat{\mathbf{a}} = \arg\max_{a} \quad p(\mathbf{a} \mid \mathbf{y}) \tag{2.11}$$

For the speech-in-noise case, the solution to the MAP problem requires solving a set of non-linear equations (Quatieri 2001). This problem has been intuitively reformulated into an iterative approach that requires a linear solution for each iteration (Lim 1979).

Essentially, we wish to maximize a joint conditional likelihood of the linear prediction coefficients vector $\mathbf{a}$, and the clean speech $\boldsymbol{x}$ with respect to the noisy speech vector $\mathbf{y}$: $p(\mathbf{a}, \mathbf{x} \mid \mathbf{y})$. Lim and Oppenheim consider a sub-optimal approach, termed linear MAP (LMAP), which begins with an initial guess of the $\hat{\mathbf{a}}^{\mathbf{0}}$. The speech $\mathbf{x}$ is estimated as the conditional mean of the noisy speech $\mathbf{y}$: $E[\mathbf{x} \mid \hat{\mathbf{a}}^{\mathbf{0}}, \mathbf{y}]$. The $\hat{a}_k(i)$, k=1,2…p, are estimated through the autocorrelation method of linear prediction. In each iteration, a new set of linear prediction parameters is estimated and the speech estimate is updated accordingly:

$$\hat{S}_s(\omega) = \frac{A^2}{\left| 1 - \sum_{k=1}^{p} \hat{a}_k(i) \exp(-j\omega k) \right|^2} \tag{2.12}$$

This estimation procedure is linear at each iteration and continues until some criterion is satisfied. Lim and Oppenheim showed that this technique, under certain conditions, increases the joint likelihood $p(\mathbf{a}, \mathbf{x} \mid \mathbf{y})$ at each iteration (Quatieri 2001). With further simplifying assumptions, it can also be shown MAP estimation of speech $\mathbf{x}$ is equivalent to non- causal Wiener filtering of the noisy speech $\mathbf{y}$.

$$H_y(\omega) = \left(\frac{\hat{S}_s(\omega)}{\hat{S}_s(\omega) + \beta \hat{S}_n(\omega)} \right) \tag{2.13}$$

where $\hat{S}_s(\omega)$ represents the estimated power spectrum of the speech for each iteration and $\hat{S}_n(\omega)$ is the estimate of the noise power spectrum.

2.5 Ephraim-Malah filtering

2.5.1 Fundamentals

Spectral subtraction and wiener filtering, discussed above, can be regarded as application of a suppression rule in the frequency domain of the corrupted signal. More precisely,

those methods capitalize on the importance of the short time spectral amplitude (STSA) of the speech signal in its perception (Ephraim 1984). However, in the spectral subtraction algorithm, the STSA is being optimally estimated in the maximum likelihood sense (ML), from the square root of the short time Fourier Transform (STFT) estimator. For Wiener filtering, the STSA estimator is derived from the optimal minimum mean square (MMSE) short time Fourier Transform (STFT) estimator. Neither of those two estimators are *optimal* spectral amplitude estimators under the assumed Gaussian model and statistical criterion (Ephraim 1984).

Directly derived from the noise observation, Ephraim and Malah proposed an optimal minimum mean-square error (MMSE) short-time spectral amplitude estimator (STSA) for speech enhancement in 1984. This spectral amplitude estimator is obtained optimally, in the maximum likelihood sense, via modeling speech and noise spectral components (i.e. Fourier coefficients) as statistically independent Gaussian random variables. Although more complex to derive theoretically than spectral subtraction or Wiener filtering, the estimator itself can be implemented fairly straightforwardly according to the formula in Ephraim's original paper, discussed below.

2.5.2 Motivation

We cannot expect to obtain the probability distribution of the speech signal simply through observing the long time behavior of this non-stationary and non-ergodic random process [(Ephraim 1984). The background noise in the real world is most often non-stationary and non-ergodic as well. In order to model speech and noise in a reasonable sense, we turn to look for the asymptotic statistical properties of their Fourier Transform coefficients.

According to the central limit theory, we can model the FT coefficients of each observation frame of speech and noise as independent Gaussian variables. This relies on the fact that each FT coefficient is nothing but a weighted sum of random variables and the correlation between coefficients reduces as the length of the observation frame increases.

When combined with the derived optimal phase estimator θ_k for the observed phase of the corrupted speech signal, the transfer function of the time varying Malah filter takes the following form:

$$H(k) = \frac{\sqrt{\pi v_k}}{2\gamma_k}\left[(1+v_k)I_0(\frac{v_k}{2}) + v_k I_1(\frac{v_k}{2})\right]\exp(\frac{-v_k}{2}) \tag{2.14}$$

where $I_0(\bullet)$ and $I_1(\bullet)$ represent zero and first order modified Bessel functions respectively. Furthermore v_k is given by:

$$v_k \overset{\Delta}{=} \frac{\xi_k}{1+\xi_k}\gamma_k \tag{2.15}$$

where ξ_k and γ_k represent the *a priori* and *a posteriori* signal-to-noise ratios for the kth spectral component. During iterative enhancement, the estimator ξ_k is obtained with knowledge of the previously enhanced spectral components.

2.5.3 Implementation

To be consistent with the Ephraim and Malah's original paper (Ephraim 1984), we denote X_k, D_k, Y_k as the kth Fourier coefficient of the speech signal, the noise process and the noisy signal observation respectively, given the analysis interval [0,T], with $X_k = A_k e^{j\alpha_k}$ and $Y_k = R_k e^{j\theta_k}$. Also $\lambda_d(k)$ represent the variance of the kth short time spectral amplitude of the noise and

equals $E\left[|D_k^2|\right]$ and $\lambda_x(k)$ represent the variance of the Kth short time spectral amplitude of the speech and equals $E\left[|X_k^2|\right]$.

Initially $\lambda_d(k)$ is calculated through the silence regions, and then an *a posteriori* SNR estimator is obtained through $\gamma_k \overset{\Delta}{=} \frac{R_k^2}{\lambda_d(k)}$ on a frame by frame basis calculation with the noisy speech FT coefficients.

For the estimation of noise statistical information, we generally desire to use the non-speech frame most adjacent in time to the specific analysis frame but this generally involves speech detection that can lead to significant errors. For the stationary white Gaussian noise, we simply use the estimation from the first several silent frames. The *a priori* SNR $\hat{\xi}_k$ is initialized according to the following formula:

$$\hat{\xi}_k = \alpha + (1-\alpha)P[\gamma_k(n)-1], 0 \le \alpha \le 1 \tag{2.16}$$

where the P function is an operator which implements the same flooring mechanism as in the spectral subtraction algorithm, to ensure the STSA estimator is positive even if $[\gamma_k(n)-1]$ is negative; α is the smoothing constant with typical default value of 0.98.

To iteratively calculate the STSA estimator of the kth speech spectral component in each analysis frame $\hat{A}_k$, we use:

$$\hat{A}_k = \frac{\xi_k}{1+\xi_k}\exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right\}R_k \tag{2.17}$$

For each frame, we update the *a priori* SNR "decision-directed" estimator $\hat{\xi}_k$ using the short time spectral amplitude estimator $\hat{A}_k(n-1)$ obtained from the previous frame:

$$\hat{\xi}_k = \alpha \frac{\hat{A}_k^2(n-1)}{\lambda_d(k,n-1)} + (1-\alpha)P[\gamma_k(n)-1], 0 \le \alpha \le 1 \tag{2.18}$$

Unlike spectral subtraction, where the spectral magnitude is averaged regardless of whether the frame contains speech or noise or both, the Ephraim Malah MMSE STSA applies nonlinear smoothing mostly when the analysis frame is predominantly noise. Interestingly, Ephraim and Malah found that this optimal MMSE STSA estimator provides colorless residual noise rather than the musical tone artifacts often incurred by spectral subtraction methods while drastically reducing the noise without significant distortion of speech contents.

2.5.4 Limitations of Ephraim Malah filtering

Ephraim Malah has the theoretical advantage of finding the optimal solution according to the short time spectral amplitude changes and the implementation is quite straightforward despite the complex mathematical basis. As with spectral subtraction and Wiener filtering, the basic problem encountered in applying Ephraim Malah filtering is again how to estimate the noise statistics to the most precise extent possibly. Speech pause detection algorithms may often introduce significant errors and lead to unsatisfactory final performance. Furthermore in dealing with non-stationary noise, the Ephraim Malah filter possesses significant advantages over other classical spectral based methods.

Chapter 3 Wavelet Based Methods

So far all the speech enhancement techniques discussed are based on the spectral information obtained through the short time Fourier transform analysis of the target signal. These are all frequency-based methods intending to preserve the slow-varying short time spectral characteristics of the speech such as the low-frequency harmonics of vowels, which is still not enough to maintain speech quality after the processing. We also wish the speech enhancement algorithm to preserve instantaneous properties such as the attack of the plosives (i.e., the stop consonants like b, d, g, p, t, k. that are transient, non-continuant sounds produced by building up pressure behind a total constriction somewhere along the vocal tract, and suddenly releasing this pressure (Deller 1994)). As a powerful time-frequency tool, the wavelet transform has established a reputation as a tool for signal analysis: having high frequency-resolution (and low time-resolution) for the low frequency content of the signal while having low frequency-resolution (and high time-resolution) for the high frequency content of the signal. The wavelet transform can be regarded as a bank of band-pass filters with constant Q factor (the ratio of the bandwidth and the central frequency). Through appropriate choice of a mother wavelet that both has finite effective support width in the time domain and concentrating property in the frequency domain, the wavelet analysis has a distinct ability to detect local features of the signal in both time and frequency, such as the plosive fine structures of the speech and other transient, instantaneous and dynamic speech components that contribute significantly to the quality of the speech (Quatieri 2001). We will first introduce the basic concepts of the classic wavelet transform and its relationship to the Fourier transform. The fundamentals of wavelet thresholding will be explained in Section 3.2. In Section 3.3 we introduce the bionic wavelet transform, recently invented.

3.1 Continuous Wavelet Transform

The Fourier transform has long been the most important underpinning for frequency-domain signal processing. The theory on wavelet transform, which originated as a branch of applied mathematics in the 1980's, was first introduced into the signal processing field thanks to the efforts of French mathematicians I. Daubechies and S. Mallat. Today, intertwined with multi-resolution and filter bank theory, wavelets analysis plays an important role in time-frequency analysis.

3.1.1 Wavelet Introduction

The word "wavelet" literally means " a small wave". A wavelet is a function that has finite energy and zero mean. It is a powerful tool for the analysis of transient, non-stationary characteristics such as drift, trends, abrupt changes, beginning and ends of events, breakdown points, and discontinuities in higher derivatives and self-similarity. We have available many kinds of wavelets: Haar, Mortlet, Daubeshies, etc.; they look different and have different properties: orthogonal, bi-orthogonal, normalized etc. For example, the Morlet wavelet is illustrated in Figure 2, with a solid line as its real part and a dashed line as its imaginary part. It is a complex exponential function at frequency ω_0 with Gaussian

envelope: $\varphi(t) = e^{\frac{-t^2}{2}} e^{j\omega_0 t}$.

Wavelet analysis is one way to localize events in time (or space) and frequency. The goal of wavelet analysis is to create a set of basis functions (i.e., expansion functions) so that the transform will give an informative, efficient and useful description of the target signal. In a

nutshell, the continuous wavelet transform (CWT) is nothing but a set of the inner products of the observed signal $f(t)$ with the shifted and scaled mother

wavelets $\varphi_{a,\tau}(t) = \frac{1}{\sqrt{a}}\varphi\left(\frac{t-\tau}{a}\right)$ where τ and a represent the time shift and scale variables.

$$< f(t), \varphi_{a,\tau}(t) >= WT_f(a,\tau) = \frac{1}{\sqrt{a}}\int f(t)\varphi^*\left(\frac{t-\tau}{a}\right)dt \tag{3.1}$$

If $\varepsilon = \int |\varphi(t)|^2 \, dt$ is the energy of the basic mother wavelet, the shifted and dilated wavelets $\varphi_{a,\tau}(t) = \frac{1}{\sqrt{a}}\varphi\left(\frac{t-\tau}{a}\right)$ maintain the same energy due to the scaling factor $\frac{1}{\sqrt{a}}$:

$$\varepsilon' = \int \left|\frac{1}{\sqrt{a}}\varphi(\frac{t-\tau}{a})\right|^2 dt = \frac{1}{a}\int \left|\varphi(\frac{t-\tau}{a})\right|^2 dt = \varepsilon \tag{3.2}$$

In order to have an inverse transform, any mother wavelet chosen must satisfy the admissibility condition that means:

$$c_\varphi = \int_0^{+\infty} \frac{|\Gamma(\omega)|^2}{\omega} d\omega < +\infty \tag{3.3}$$

where $\Gamma(\omega)$ denotes the mother wavelet in the frequency domain. This condition implies at least two things about a valid mother wavelet: 1) $\Gamma(\omega)$ has band-pass property ($\Gamma(0) = 0$); 2) $\varphi(t)$ has an oscillatory characteristic.

After satisfying the admissibility condition, the inverse transform is given by:

$$f(t) = \frac{1}{C_\varphi}\int_0^{\infty}\int_{-\infty}^{\infty} WT_f(a,\tau)\varphi(t)d\tau\frac{da}{a^2} \tag{3.4}$$

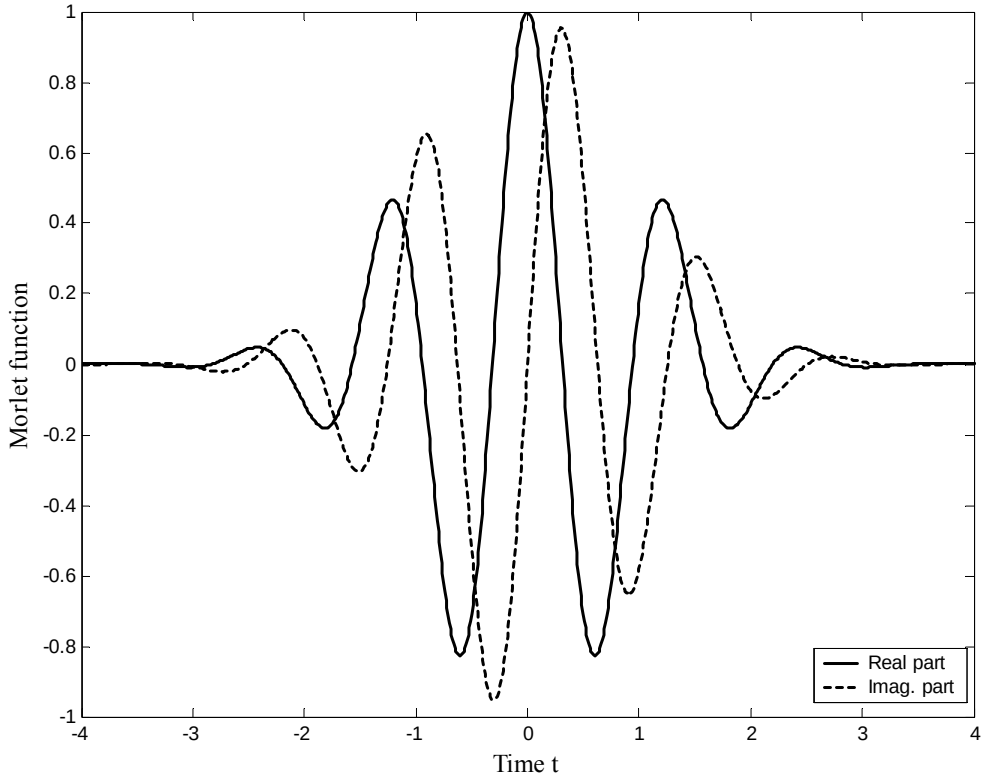

Figure 2 The Morlet wavelet in the time domain.

3.1.2 Comparison with Short Time Fourier Transform (STFT)

To understand the major advantages of wavelet transforms, let us first review the short time Fourier transform (STFT) that is the most used spectral analysis method in speech signal processing.

$$F(\omega,\tau) = \int_{-\infty}^{\infty} f(t)w(t-\tau)e^{-j\omega t}dt \tag{3.5}$$

where $f(t)$ is the target signal and $w(t-\tau)$ is the moving window. The limitation of the standard Fourier transform is that it extracts the frequency content of the signal only but not the frequency changes with respect to time. This is partially solved through the STFT by using sliding analysis windows. However the STFT uses a fixed window length and still

cannot always simultaneously resolve short-lived events and closely spaced long-duration tones in speech (Quatieri 2001). This drawback is rooted in the well-known uncertainty principle that limits time-frequency resolution: $D(x)B(x) \geq \frac{1}{4}$ where the product of time duration D(x) and bandwidth B(x) of a signal x must exceed a constant.

The wavelet transform minimizes the limitation of the uncertainty principle by varying the length of the moving window with variant scaling factor a. Ideally, long windows are employed on low frequency parts of the speech signal for good frequency resolution and short windows are employed on high frequency components of the speech signal, say the attack of the glottal pulse and plosives of speech, for good time resolution. The wavelet transform succeeds in adjusting time and frequency resolution without defeating the uncertainty principle.

3.1.3 Implementation of Continuous Wavelet Transform

To calculate the inner product of the CWT, normally we need to resort to numerical integration using computers. The simplest way is to discretize time and shift as follows: $t = nT_s$ and $\tau = kT_s$ and T_s is the sampling interval. Then Eq. 3.1 becomes:

$$WT_f(a, kT_s) = \frac{T_s}{\sqrt{a}} \sum_n f(nT_s) \varphi \left[\frac{(n-k)T_s}{a} \right] \tag{3.6}$$

For each value of the scale a, we obtain a set of wavelet coefficients under this specific scale. There are some other existing fast algorithms for the continuous wavelet transform such as algorithm a'trous (Holschneider 1989), chirp-z transform (Jones 1991), Mellin transform (Bertrand 1990).

Under the admissibility condition:

$$c_{\varphi} = \int_{0}^{+\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega < +\infty \tag{3.7}$$

the two-dimensional wavelet coefficients $WT_f(a, kT_s)$ are a complete, stable yet redundant representation of the one dimensional signal. In order to speed up computation and save memory, we wish to discretize the scale a and shift τ in an efficient way to form a new set of wavelet coefficients.

3.1.4 Discrete Wavelet Transform

One drawback of the CWT is that the representation of the signal is often redundant. Unlike the continuous wavelet transform, which can operate on every scale, the discrete wavelet transform (DWT) chooses a subset of scales and positions to calculate. A sample version of the wavelet coefficients $WT_f(a, \tau)$ can reconstruct the original signal in an efficient way if the family of dilated and shifted mother wavelets of selected *a* and τ constitute an orthonogonal and complete basis (Daubechies 1992). A common sampling practice is that for each scale $a_m = a_0^m$ for m=0,1,2,3…N, the sampling interval is $\tau_m = \tau_0 a_0^m$ for m=0,1,2,3…N. One particular natural case is when a_0=2 so that the sampling rate of the shift decreases by a factor of two as the scale increases by a factor of two (Quatieri 2001). This is so called *dyadic or octave* sampling and it allows the implementation of a fast dyadic wavelet transform and its inverse with filter banks. High-pass filter removes the low-frequency components of the signal and the corresponding filter parameters become the *detailing* part of the wavelet coefficients. Low-pass filter removes the high frequency components of the signal and the corresponding filter parameters become the *smoothing* part of the wavelet coefficients. Partly due to the efficient implementation and auditory and visual cortex-like properties of dyadic wavelets, a large part of wavelet theory has involved finding

dyadic wavelet bases that are orthogonal and that are useful in a variety of applications (Mallat 1998).

3.1.5 Multi-resolution Analysis of Discrete Wavelet Transform

The multi-resolution analysis concept was initiated by Meyer (Meyer 1992) and Mallat (Mallat 1989) and provides a natural framework for the understanding of wavelet bases. In the dyadic wavelet transform, the basis functions are two parts: the scaling functions $\psi(t)$ and the wavelet functions $\varphi(t)$.

$$\psi_{m,\tau}(t) = 2^{-\frac{m}{2}}\psi^0(2^{-m}t-\tau) \text{ where } m \in Z, \tau = n*2^m \in Z \qquad (3.8)$$

$$\varphi_{m,\tau}(t) = a_0^{-\frac{m}{2}}\varphi^0(2^{-m}t-\tau) \qquad (3.9)$$

The scaling function $\psi^0(t)$ can be obtained as a sum of copies (dilated, shifted, scaled versions) of itself as illustrated in Eq.3.10,

$$\psi^0(t) = \sum_{\tau=0}^{L} C_\tau \psi(2t-\tau) \qquad (3.10)$$

and the wavelet function $\varphi^0(t)$ can be then obtained from the scaling function $\psi^0(t)$ as follows:

$$\varphi^0(t) = \sum_{\tau=0}^{L} (-1)^\tau C_{1-\tau}\psi^0(2t-\tau) \qquad (3.11)$$

where C_τ can be seen as the low-pass filter coefficients and $C_{1-\tau}$ can be seen as the high-pass filter coefficients and where L–1 is related to the number of vanishing moments in the scaling function $\psi^0(t)$. They two together constitute a quadrature mirror filter (QMF) and an

extensive study of the QMF can be found in (Monzon May,1994). The simple relation of two filter coefficients is as follows:

$$C_{\tau}(\tau) = (-1)^{\tau} C_{1-\tau}(L-1-\tau) \tag{3.12}$$

Having the basis for decomposition, we can write the dyadic wavelet transform as follows:

$$f(t) = \sum_{\tau} c_{J,\tau} \psi^{0}(t-\tau) + \sum_{\tau} \sum_{m=1}^{J} d_{j,k} \varphi^{0}(a_0{}^{m} * t - \tau) \tag{3.13}$$

where ϕ is the scaling function and φ is the wavelet function , $a_0 = 2,\ \ m = 1, 2, ...N,$ and $\tau = \tau_0 * a_0^m$. The above equation shows how a signal can be decomposed into the summation of approximations (low frequency components of the signal) and details (high frequency components of the signal) at different resolutions.

3.2 Wavelet Thresholding

As wavelet analysis has its basis emulating the front-end auditory periphery (Mallat 1998), efforts have been made to take advantage this signal-processing tool for speech enhancement. The most used approach is based on the non-linear thresholding of the wavelet coefficients (Donoho 1995), which bridges the multi-resolution analysis and non-linear filtering.

3.2.1 Principle

Donoho proposed this powerful wavelet-based approach as follows:

Let y be a finite length observation sequence of the signal x that is corrupted by zero-mean white Gaussian noise n with variance σ^2:

$$y = x + n \tag{3.13}$$

In the wavelet domain, this gives:

$$W_y = W_x + W_n \tag{3.14}$$

The clean signal x can be estimated in the following way:

$$x = W^{-1} X_{estimation} = W^{-1} Y_{thresh} \tag{3.15}$$

where Y_{thresh} represents the wavelet coefficients after thresholding and W^{-1} denotes the inverse wavelet transform. The approach capitalizes on the fact that an appropriate transform (i.e., wavelet transform) projects the signal onto the transformed domain where the signal energy is concentrated in a small number of coefficients, while the noise is evenly distributed across the transformed domain. There are generally two ways of thresholding: one is called hard thresholding (Eq.3.16) and the other is called soft thresholding (Eq.3.17). Figure 3 is an illustration of this technique.

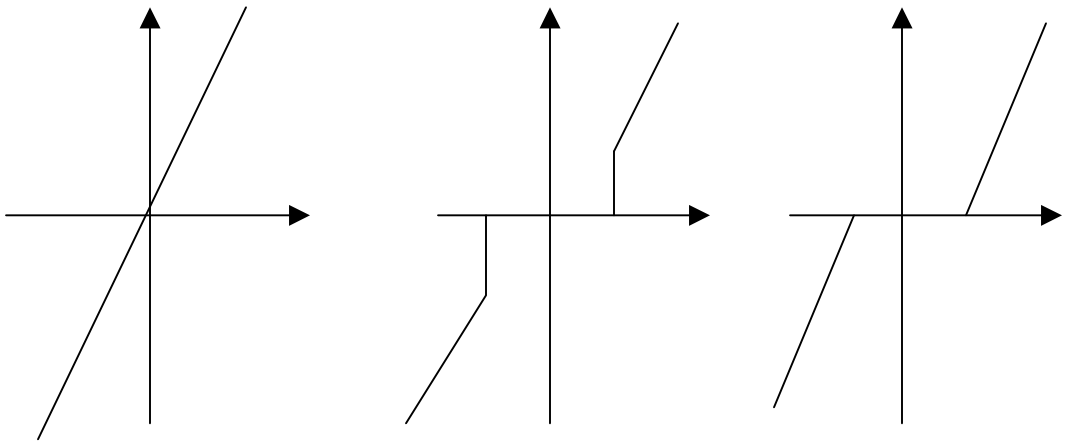

Figure 3 Without thresholding, hard thesholding, soft thresholding

$$Thr_{Hard}(X,T) = \begin{cases} X & |X| > T \\ 0 & |X| < T \end{cases} \tag{3.16}$$

$$Thr_{Soft}(X,T)=\begin{cases} Sgn(X)(|X|-T) & |X|>T \\ 0 & |X|<T \end{cases} \tag{3.17}$$

where X represents the wavelet coefficients before thresholding and T is the threshold. Both of these two methods suffer from distortion of the speech because they set coefficients to zero that may carry useful information, resulting in observable sharp time frequency discontinuities in the speech spectrogram. Various modifications have been made. For example, Sheikhzadeh (Sheikhzadeh 2001) proposed using an exponential function to attenuate coefficients that are smaller than the threshold value in a nonlinear manner to avoid creating abrupt changes. Other data compression functions can also be chosen such as the μ-law:

$$Thr(X,T)=\begin{cases} X & |X|>T \\ T\left(\dfrac{\left[(1+\mu)^{|X/T|}-1\right]}{\mu}\operatorname{sgn}(X)\right) & |X|<T \end{cases} \tag{3.17}$$

where X is the wavelet coefficients and T is the threshold value.

3.2.2 How to Choose the Threshold

The choosing of the threshold value can be determined in many ways. Donoho derived the following formula based on white Gaussian noise assumption:

$$T=\sigma\sqrt{2\log(N)} \tag{3.18}$$

where T is the threshold value, N is the length of the noisy signal y, and σ=MAD/0.6745, with MAD denoting the absolute median estimated on the first scale of the wavelet coefficients.

Johnstone and Silverman (Johnstone 1997) proposed the level dependent threshold method to deal with correlated noise, where for each frequency interval the threshold is proportional to the standard deviation of the noise in that interval.

$$\lambda_a = \sigma_a \sqrt{2\log(N_a)} \tag{3.19}$$

with $\sigma_a = MAD_a / 0.6745$, N_a is the number of samples in scale a, and MAD_a is the absolute median estimated at scale a.

3.2.3 Four Types of Threshold Selection Rules

- Threshold selection rule based on Stein's unbiased estimate of the risk

Different estimation rules could be compared on the basis of their resulting mean-square error (MSE) or more formally, the risk: $\text{Risk}(s, \text{Threshold}) = E\left\{\left\|s - \hat{s}\right\|^2\right\}$. Stein (Stein 1981) has, under quite general conditions, derived an unbiased estimator of such a risk for a Gaussian estimator.

- Heuristic threshold selection rule

This is a heuristic variant of the first option (Mathworks 1998).

- Fixed form threshold selection rule

This rule uses the universal threshold shown by Eq.3.18.

- Minimax performance threshold selection rule

The minimax rule uses a fixed threshold chosen to yield minimax performance for mean square error against an ideal procedure. The derived formula is as follows (Guo 2000):

$\text{Threshold} = 0.3936 + 0.1829 * \dfrac{\log(N)}{\log(2)}$ where N is the length of the signal.

3.2.4 Limitation of the Wavelet based Thresholding

Although the wavelet based method does not require a speech or noise model and can be applied to a broader class of signals, merely a general thresholding on the wavelet coefficients does not guarantee a good performance as will be shown later by our tests in Chapter 4. The bionic wavelet transform we are presenting as follows has a better signal energy concentration property and time-frequency selectivity and this will be expected to yield far more efficient thresholding performance.

3.3 Bionic Wavelet Transform

J. Yao and Y.T. Zhang proposed the bionic wavelet transform (BWT) as a new time frequency method based on an auditory model (Yao 2001). “Bionic” means that it is rooted in an active biological mechanism. The BWT is distinguished from the standard wavelet transform (WT) in that the resolution in the time-frequency domain achieved by the BWT can be adaptively adjusted not only by the signal frequency changes but also by the signal’s instantaneous amplitude and its first-order differential (Yao 2001). It is an adaptive mother wavelet that makes a wavelet transform adaptive, and in the BWT, it is the active control mechanism in the human auditory model that adjusts the mother wavelet according to the signal to be analyzed.

3.3.1 Human Auditory Front-end Periphery

The outer ear, middle ear and inner ear constitute the human peripheral auditory system. Sound travels through the outer ear into the ear canal and results in vibrations of the eardrum in the outer ear that connects to the middle ear. Some frequencies of the incoming sound are being

attenuated more than others. The middle ear transforms the air vibrations from the outer ear into the vibrations in the fluid of the inner ear. Through the impedance matching of the middle ear, the energy from the outer ear is transferred into the energy of the inner ear. Other than being a transformer, the middle ear also acts as a low-pass filter (Guinan 1967).

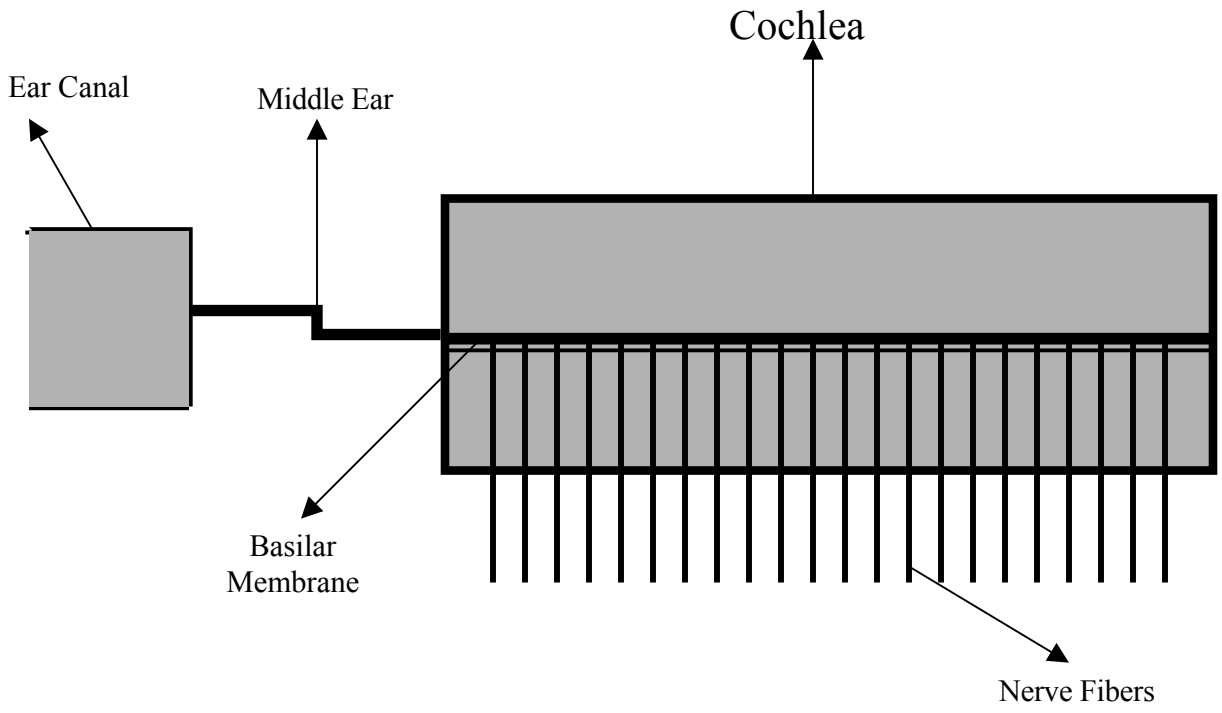

Figure 4 Simplified illustration of front-end human auditory system

The inner ear performs most of the signal processing tasks of acoustic signals. Its organs include the cochlea, basilar membrane, the inner and outer hair cells and the auditory nerves. The major component of the inner ear is the cochlea that can be viewed as a spatial frequency analyzer (Quatieri 2001). The cochlea is a snail-shaped coiled tube full of fluid, with the basilar membrane running midway along it. The simplified schematic of an uncoiled cochlea is illustrated above [Figure 4]. The basilar membrane is to transform fluidic stimuli into electrical

stimuli. Due to the stiffness gradient of the basilar membrane, the vibrations of a specific tone frequency propagate along it like a traveling wave in the direction toward the apex from the base, and at one specific point of the basilar membrane the amplitude of the wave reaches its highest point before decaying. This specific tone frequency is referred to as the characteristic frequency of the cochlea. Furthermore, the maximum displacement stimulates the inner hair cells (IHCs) that are sensitive to any displacement of the basilar membrane. The neurons connected to the inner hair cells and outer cells undergo a firing, which causes impulses to be sent to the brain via nerve fibers.

As the basilar membrane is thinning from the base to the apex, lower frequency propagates further towards the apex of the cochlear. As such, the cochlea segregates different incoming frequencies into different spatial locations along its length (Yang 1992). A popular functional view of the cochlea is to think of it as a bank of band pass filters with a constant Q factor (the ratio of the center frequency and the bandwidth remains constant). For above 800Hz in humans the transfer function of those cochlear filters remain approximately invariant except for a translation or time-shift while the center frequency of those filters decreases along the base to the apex (Yang 1992). We can see that the spatial axis of the cochlea is akin to the scale axis and the outputs of the cochlear from each stimulus act as those of a standard wavelet transform.

However, by making this signal processing abstraction we have ignored several nonlinear phenomena that achieve high sensitivity and frequency selectivity of the cochlea. Among them is the so called “ active cochlear mechanism”(Deng 1988). The bionic wavelet transform elaborated hereafter is originated from this active cochlear mechanism that has been qualitatively reflected in the modified Giguere’s auditory model.

3.3.2 Modified Giguere's Auditory Model

Researchers have attempted the modeling of human auditory system for a long time. There are four stages involved in this modeling, namely the outer ear and the middle ear, the cochlea of the inner ear, the inner and outer hair cell responses, and high-level brain processing.

Giguere and Woodland (Giguere 1994) put separate auditory model stages together and further constructed a computational model from the external ear to the inner hair cells. Three major parts are included in Giguere's model, namely the auditory canal, the middle ear and the cochlea (Giguere 1994). Motivated by the analogy of electronics and acoustics, the model is a typical electrical network based on the compact electro-acoustic theory, and each component in the model represents a relevant anatomical part of the auditory system. The auditory canal is modeled using four T sections in cascade, which is the discretized form of a uniform transmission line (Giguere 1994). See Figure 5. The middle ear and cochlea is coupled with an ideal transformer that simulates the acoustics flow from the eardrum to the oval window. The cochlea is modeled as a nonuniform and nonlinear transmission line (Giguere 1994) (Furst 1988), which is divided into N sections from its base to apex, and each section is emulated by a typical RLC resonant circuit. Refer to Figure 6. In each section, the series inductor L_{sn} represents the acoustic mass of the cochlear fluid, and the shunt resistor R_n, indicator L_n and capacitor C_n represent the acoustic resistance, mass and stiffness of the basilar membrane respectively (Zheng 1999). Active sources are used and seen as the mobile factor that relates to the sensitivity and frequency selectivity of the cochlear. When the input sound pressure is very low, the active source increases the frequency selectivity of a segment by reducing the segment's damping. Therefore each RLC circuit section has the resonant frequency

corresponding to the characteristic frequency of that segment of basilar membrane, and can be regarded as a band pass filter with nonlinear damping.

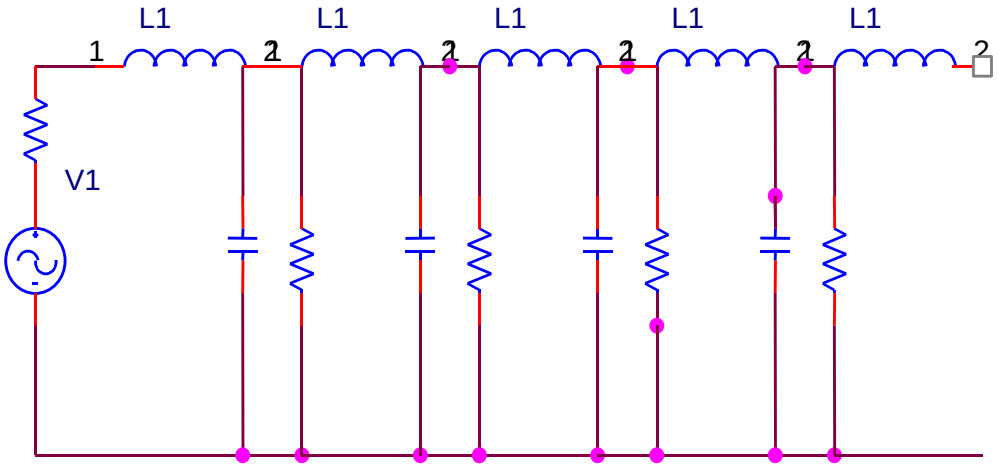

Figure 5 Schematic of the auditory canal (Adopted from [Giguere, 1994])

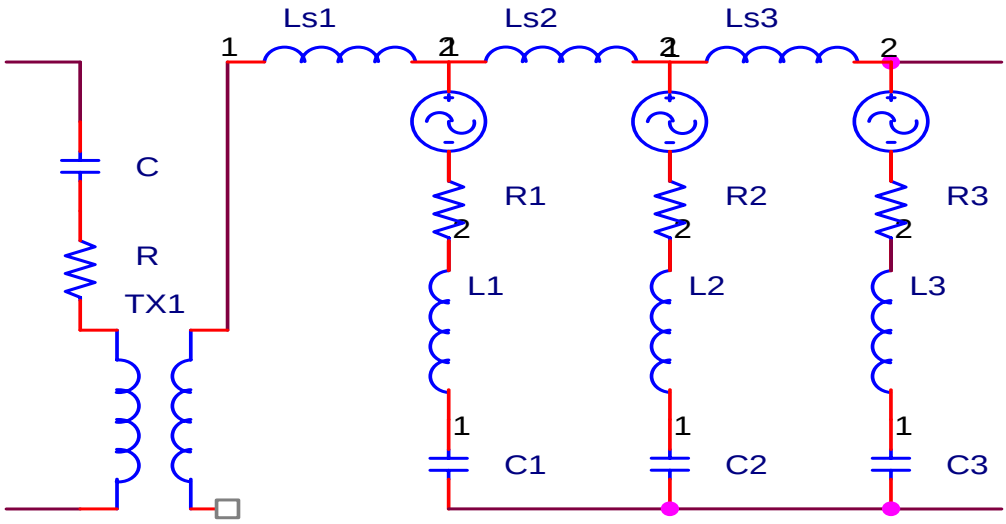

Figure 6 Schematic of the cochlea (Adopted from [Giguere, 1994])

Giguere's auditory model has been proven to be helpful to the study of OAEs (otoacoustic Emissions), which are low level sounds generated from the inner ear either spontaneously or due to an acoustic stimulus that can be detected in the outer ear canal by a sensitive microphone (Whitehead 1994), (Kemp 1978), (Probst 1991). OAEs have long been used for better

understanding of cochlear functions and the auditory periphery. Zheng (Zheng 1999) has shown clearly that the generation of transient evoked OAE signals (TEOAEs) is related to the same active mechanism as the cochlear sharp frequency selectivity. The experimental evidences of the spontaneous OAEs (SOAEs) from (Yao 1999) have further proved that the cochlea is nonlinear and active.

Physiological research has found that the activities of inner hair cells (IHCs) and outer hair cells (OHCs) cause the quality factor Q of the cochlear filter banks to be changeable according to the basilar membrane displacement (Gelfand 1998), which means that the standard wavelet transform abstraction of the cochlea only shows the passive functioning side of it. Recent work (Yao 2001) has adopted Giguere's model for the qualitative description of this active phenomenon of the cochlea. In Giguere's model, at one point of the basilar membrane, we qualitatively model the displacement of the basilar membrane d as a function of the tone frequency ω_0 by the following equation:

$$\ddot{d}(x,t) + R_{eq(x,d)}\,\dot{d}(x,t)/L(x) + \omega_0^2(x)d(x,t) = P \qquad (3.20)$$

where x represents the distance along the basilar membrane from the basal end and t represents time; d stands for the displacement of the basilar membrane and $\ddot{d}$ and $\dot{d}$ are the first and second-order differentials of d in terms of t respectively. P is the pressure difference across the basilar membrane and ω_0 characterizes the tone frequency at that point of the basilar membrane and equals $1/\sqrt{L(x)C(x)}$, where $L(x)$ and $C(x)$ represent the acoustic mass and compliances, respectively and the parameters L, C, ω_0 are constant on a specific point of basilar membrane. The equivalent resistance R_{eq} in Eq.3.21 is given by:

$$R_{eq} = R(x) - G_1(x)\frac{d_{1/2}}{d_{1/2} + |\, d(x,t)\,|} R(x) \tag{3.21}$$

where $R(x)$ is passive resistance corresponding to the acoustic resistance and is constant on a specific point of basilar membrane. $d_{1/2}$ is a saturation factor and $G_1(x)$ is the active gain factor whose value is related to the activity of the corresponding outer hair cell group. The term $G_1(x)\frac{d_{1/2}}{d_{1/2} + |\, d(x,t)\,|} R(x)$ hence represents the active control functions of the outer hair cells (OHCs), the output of which was found to be proportional to the basilar membrane displacement (Gelfand 1998).

Recent research results, however, have shown that nonlinear damping alone is not enough to describe the active mechanism of the cochlea and that nonlinear compliance is also necessary (Hubbard 1995). (Gelfand 1998) pointed out that the output of the inner hair cells (IHCs) is proportional to the velocity of the displacement of the basilar membrane. We hence further introduce the nonlinear capacitance as follows:

$$C_{eq}(x) = \left(1 + G_2(x)\,|\,\frac{\partial[d(x,t)]}{\partial t}\,|\right)^2 C(x) \tag{3.22}$$

where $G_2(x)$ is the active factor that relates to the effect of the active mechanism on the compliance of a point on the basilar membrane. This changes the tone frequency ω_0 to $1/\sqrt{L(x)C_{eq}(x)}$ and makes it adaptive rather than fixed.

As R_{eq} and C_{eq} vary according to the displacement of the basilar membrane, the pervious signal processing abstraction of the cochlea as having constant quality factor $Q = R^{-1}\sqrt{L/C}$ is no longer valid. Based on Eq.3.21 Eq.3.22 it becomes: $Q_{eq} = R_{eq}^{-1}\sqrt{L/C_{eq}}$. It is this

adaptable new quality factor Q_{eq} that motivates us to arrive at a new wavelet transform with adaptive mother wavelet according to not only the frequency changes of the target signal but also to the amplitude of it.

3.3.3 The Origination of Bionic Wavelet Transform (BWT)

Basically, the idea of the BWT is to make the envelope of the mother wavelet time varying according to the characteristics of the target signal by introducing a time varying T function. Let us first review some basics of the wavelet transform (WT) again.

Any mother wavelet $\varphi(t)$ of the continuous wavelet transform must satisfy the admissibility condition to be invertible. The implies $\varphi(t)$ has oscillations as follows:

$$\varphi(t) = \frac{1}{\sqrt{a}} \tilde{\varphi}(t) \exp(j\omega_0 t) \tag{3.23}$$

where $\tilde{\varphi}(t)$ is the envelope of $\varphi(t)$. To apply the WT definition on a signal $x(t)$, which simply means the inner product between the shifted and dilated versions of mother wavelet and the signal itself, we have:

$$WT_f(a,\tau) = \frac{1}{\sqrt{a}} \int x(t) \tilde{\varphi}^*(\frac{t-\tau}{a}) \exp(-j\omega_0(\frac{t-\tau}{a}))dt \tag{3.24}$$

where a is scale and τ is the time shift.

To emulate the active control function of the hair cells of the auditory system, we introduce the T function that will adjust the mother wavelet $\varphi(t)$ in the following way:

$$\varphi(t) = \frac{1}{T\sqrt{a}} \tilde{\varphi}(\frac{t}{T}) \exp(j\omega_0 t) \tag{3.25}$$

the first T is just the scaling factor to insure the energy is the same for every mother wavelet and the second T adjusts the envelope of the mother wavelet $\varphi(t)$ without adjusting its center frequency. Applying the WT definition with our new adaptive mother wavelet, we have the bionic wavelet transform:

$$BWT_f(a,\tau) = \frac{1}{T\sqrt{a}} \int f(t)\tilde{\varphi}^*(\frac{t-\tau}{Ta}) \exp(-j\omega_0(\frac{t-\tau}{a}))dt \tag{3.26}$$

3.3.4 T function

Before we derive the formula for the calculation of the T function, let us first review the difference of the WT and the BWT.

Given T as a constant in a short enough period, the Fourier transforms of $\varphi(t)$ and $\varphi_T(t)$ are as follows:

$$\Gamma(\omega) = \frac{1}{\sqrt{a}} \tilde{\Gamma}[(\omega - \omega_0)] \tag{3.27}$$

$$\Gamma_T(\omega) = \frac{1}{\sqrt{a}} \tilde{\Gamma}[T(\omega - \omega_0)] \tag{3.28}$$

where Γ and Γ_T represents the Fourier transforms of $\varphi(t)$ and $\varphi_T(t)$ respectively. It is clear that

$$Q_T = TQ_0 \tag{3.29}$$

After substituting R_{eq} [3.21] and C_{eq} [3.22] into $Q_{eq} = R_{eq}^{-1}\sqrt{L/C_{eq}}$, we have:

$$Q_{eq} = \left(1 - G_1 \frac{d_{1/2}}{d_{1/2} + |d(x,t)|}\right)^{-1} \times \left(1 + G_2 |\partial(d(x,t)/\partial(t)\right)^{-1} Q \tag{3.30}$$

Comparing [3.29] with [3.30], we conclude:

$$T(\tau+\Delta\tau)=\left(1-G_1\frac{BWT_s}{BWT_s+|BWT_f(a,\tau)|}\right)^{-1}\times\left(1+G_2\,|\,\partial(BWT_f(a,\tau)/\partial(t)\,|\right)^{-1} \quad (3.31)$$

where G_1 and G_2 are active factors which we discussed in the previous section, $BWT_f(a,\tau)$ is the coefficient of the BWT at time τ and scale a that maps the displacement of the basilar membrane $d(x,t)$ in Eq.3.20, and BWT_s maps the saturation constant factor $d_{1/2}$ in Eq.3.20. In the implementation, T is a constant in the period of a calculation step $\Delta\tau$. Obviously T is related to signal instantaneous amplitude and its first-order differential. The values of G_1 and G_2 and BWT_s depend on the target signal properties and other parameter settings, and we chose them to be 0.87,45,0.8 in our experiments according to (Yao 2001).

3.3.5 K Factor

The BWT is adaptively adjusted according to the signal's instantaneous amplitude and its first-order differential in that the T function is obtained step by step as in Eq.3.31 through numerical integration. However, the direct implementation of this algorithm leads to huge computation costs that make the practical use almost impossible. The affinity of the WT coefficients and the BWT coefficients motivated us to discover the linear relationship between them (Yao 2002):

According to Moyal theorem, we have the following equation:

$$< f_1(x),\, f_2(x) >= \frac{1}{C_\varphi} < WT_{f1}(a,\tau), WT_{f2}(a,\tau) > \quad (3.32)$$

where $C_\varphi=\int_0^{+\infty}\frac{|\Gamma(\omega)|^2}{|\omega|}d\omega<+\infty$, $\Gamma(\omega)$ is the Fourier transform of $f_1(x)$ and

$WT_{f1}(a,\tau), WT_{f2}(a,\tau)$ are the wavelet coefficients of the signals f1,f2 respectively.

We then let $f_2(x) = \varphi_{a,\tau}^{bwt}(t) = \frac{1}{T\sqrt{a}}\tilde{\varphi}\left(\frac{t-\tau}{aT}\right)\exp\left(j\omega_0\left(\frac{t-\tau}{a}\right)\right)$ the left hand side of Eq.3.32

becomes

$$< f(t), \varphi_{a,\tau}^{bwf}(t) >= BWT_f(a,\tau) =< WT_f(a,\tau), WT_{\varphi^{bwt}}(a,\tau) > \quad (3.33)$$

Meanwhile

$$WT_{\varphi^{bwt}}(a,\tau) =< \varphi^{bwt}(t), \varphi_{a,\tau}(t) >= \frac{1}{aT}\int \varphi\left(\frac{t-\tau}{aT}\right)\varphi\left(\frac{t-\tau}{a}\right)dt \quad (3.34)$$

$$= \frac{1}{aT}\int \tilde{\varphi}\left(\frac{t-\tau}{aT}\right)\exp\left(j\omega_0\left(\frac{t-\tau}{a}\right)\right)\tilde{\varphi}\left(\frac{t-\tau}{a}\right)\exp\left(j\omega_0\left(\frac{t-\tau}{a}\right)\right)dt \quad (3.35)$$

Let $\hat{t} = \frac{t-\tau}{a}$; then Eq.3.34 becomes

$$< \varphi^{bwt}(t), \varphi_{a,\tau}(t) >= \frac{1}{T}\int \tilde{\varphi}_{a,\tau}\left(\frac{\hat{t}}{T}\right)\exp\left(j\omega_0\left(\hat{t}\right)\right)\tilde{\varphi}_{a,\tau}\left(\hat{t}\right)\exp\left(j\omega_0\left(\hat{t}\right)\right)d\hat{t} = K' \quad (3.36)$$

where $\hat{t}$ is the time integration variable and T is a constant for certain scale and time. The Eq.3.36 is therefore a constant that can be represented by K'. Hence the right hand side of Eq.3.32 satisfies:

$$< WT_f(a,\tau), WT_{\varphi^{bwt}}(a,\tau) >=< WT_f(a,\tau), K^{'} >= \frac{K'}{C_\varphi}WT_f(a,\tau) \quad (3.36)$$

Substituting 3.36 into 3.33, we have: $BWT_f(a,\tau) = K * WT_f(a,\tau)$ where K is solely dependent on the T function. In the case where Morlet wavelet is the mother wavelet for the bionic wavelet transform, $\varphi(t) = \exp\left(-\frac{t}{T_0}\right)^2 \exp\left(j\omega_0 t\right)$, the K factor can be approximated by $\frac{1.7725T_0}{\sqrt{T^2+1}}$ (Yao 2002).

Chapter 4 Implementation and Performance Evaluation

This chapter describes the implementation details and performance evaluation of the proposed bionic wavelet transform denoising algorithm compared to four other traditional speech enhancement methods. Evaluation of a speech enhancement algorithm is not trivial. Speech quality and speech intelligibility are both important for the assessment. Speech quality indicates how well or natural the speech sounds and speech intelligibility measures how well the information carried via speech can be understood by human listeners. While objective quality assessment methods can somehow reflect speech quality based on mathematical measures, some of them may not be consistent with human perception. Subjective measures of intelligibility and quality are thus often required. In the thesis, speech quality evaluation is our focus. Section 4.1 describes the implementation of the proposed algorithm and the speech material used to test the algorithms. Section 4.2 explains the objective measures that were used to evaluate the algorithms. Section 4.3 explains the subjective measures that were used to evaluate the algorithms.

4.1 Implementation

Ten different utterances (see Table 1) from the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus database were used to evaluate the proposed algorithm and compare with other conventional algorithms. Every sentence has a silence segment in the beginning that lasts more than 100ms.We corrupted them using white Gaussian noise at the following SNR levels in dB: -10, -5,0,5,10. We also corrupted those utterances at 0dB SNR with F-16 cockpit noise, Volvo car interior noise, pink noise, and multiple talkers' noise

(babble noise) in order to investigate how those methods deal with non-stationary real-life noises. The power spectra of the different noises are illustrated in Figure 7. Our speech enhancement algorithms have been tested on those 10 corrupted utterances respectively.

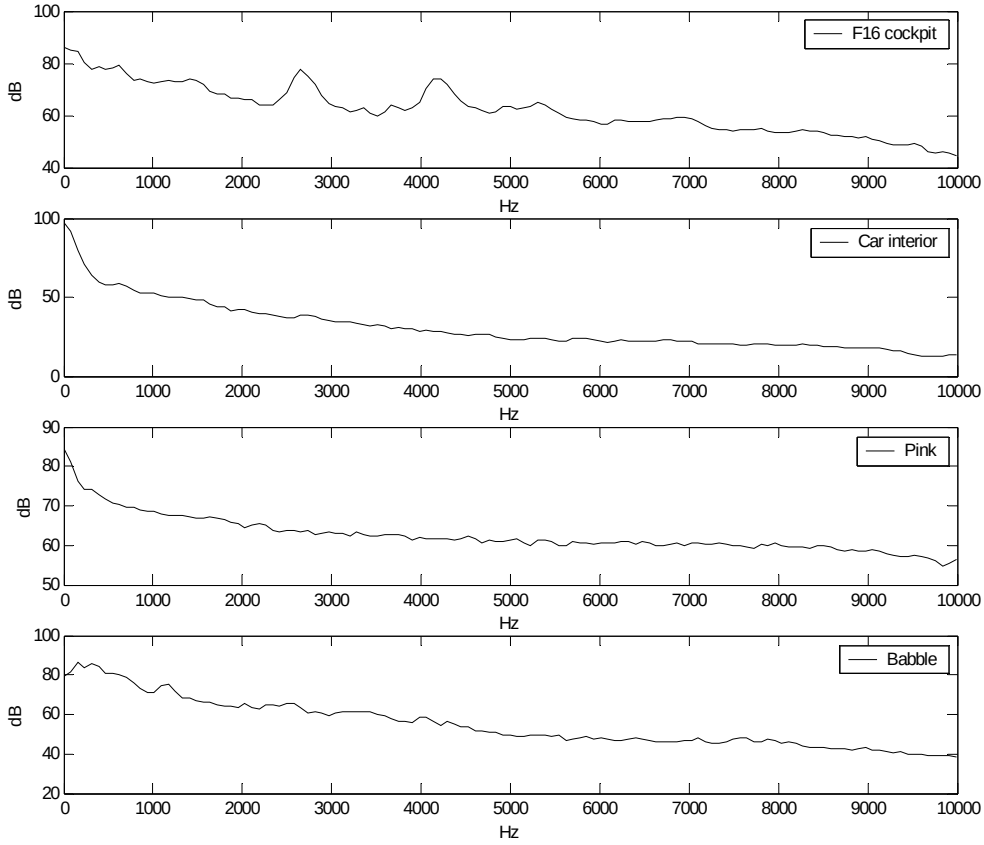

Figure 7 The power spectra of the different noises

For the bionic wavelet thresholding case, as illustrated in Figure 7, the Morlet function is chosen as the mother wavelet, which does not have scaling functions and therefore does not support the fast discrete wavelet transform. The support length of Morlet wavelet is chosen as [-4,4], and 2.5π is chosen as its oscillatory frequency. The central frequencies for bionic

wavelet analysis range from 221 - 5291 Hz (1~22 scales). We apply the soft thresholding on the bionic wavelet coefficients using the four selection rules: Stein's unbiased estimate of the risk rule, heuristic threshold selection rule, fixed selection rule, and minimax-performance threshold selection rule. The modified bionic wavelet coefficients are divided by the K factor to become the wavelet coefficients and the inverse continuous wavelet transform is employed to reconstruct the speech signal.

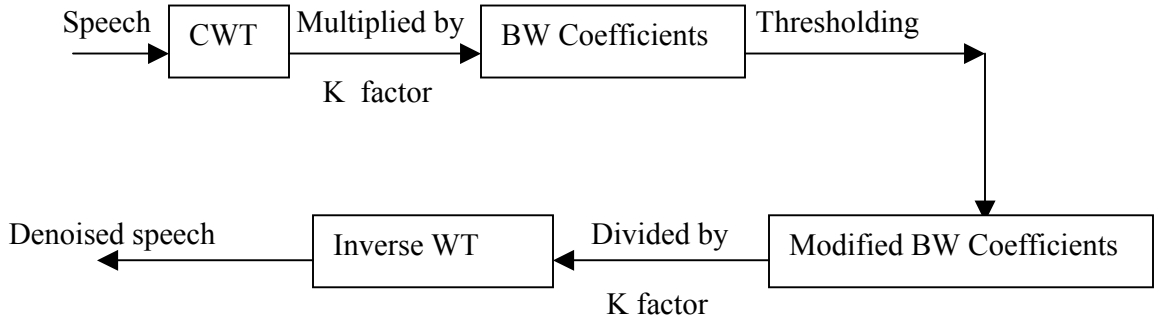

Figure 8 The procedure of the bionic wavelet transform denoising technique

For spectral subtraction, Wiener filtering, and Ephraim Malah filtering, the speech signal is first Hanning-windowed using a 25-ms window and a 12-ms overlap between frames. The windowed speech frame is then analyzed using the Fast Fourier Transform (FFT). Noise estimation is done through the first 3 analysis frames at the beginning of the data. For the traditional wavelet thresholding method, the Daubechies-1 mother wavelet is used (as the Morlet mother wavelet does not support discrete wavelet theory) with 3-level dyadic discrete wavelet decomposition and the multiplicative threshold rescaling is employed to estimate noise at each level and the details coefficients are soft thresholded based on four selection rules: Stein's unbiased estimate of the risk, heuristic threshold selection rule, fixed form rule, and minimax-performance threshold selection rule.

File	Sentence
1	First add milk to the shredded cheese.
2	Oh Mother, I saw them!
3	Heat's bad for frostbite.
4	The singer's finger had a splinter.
5	Run-down, iron-poor.
6	It's never wrong if love is real.
7	Did dad do academic bidding?
8	They'll roll off in another day.
9	Primitive tribes have an upbeat attitude.
10	Does Creole cooking use curry?

Table 1 List of sentences used from the TIMIT database for objective and subject performance evaluation

4.2 Objective measure

The signal-to-noise ratio (SNR) is the most widely used objective measure of speech quality. Variations include classical SNR, segmental SNR and frequency weighted segmental SNR.

The classical SNR measure (in dB) is obtained as

$$SNR = 10\log_{10}\frac{\sum_{n} s^2(n)}{\sum_{n}[s(n)-\hat{s}(n)]^2} \tag{4.1}$$

where $s(n)$ represents the clean speech which is only accessible in experiments where noise is artificially added, $\hat{s}(n)$ represents the enhanced speech. Despite its mathematical simplicity, this method is a poor estimate of speech quality for a broad range of speech distortions (Deller 1994). However, measuring SNR on frame-by-frame basis and averaging the results leads to an improved version of the classic SNR method, called segmental SNR:

$$SSNR = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\sum_{n=m_j-N+1}^{m_j} \frac{s^2(n)}{[s(n)-\hat{s}(n)]^2}\right] \tag{4.2}$$

where M is the number of frames, each of which is length N. For each frame (typically 15-25 msec), an SNR is calculated and then after averaging all these measurements over all segments of the speech signal, we obtain the final SNR measurement having assigned equal weight on loud and soft portions of the speech. Another factor that affects the SNR measures' performance is the silent regions of the speech, so it is desirable that we could effectively remove them, if possible.

An important aspect of using objective measures is that we have to make sure the clean speech and processed speech are synchronized during the calculation because there are typically time delays after a signal is passed through the enhancement system.

4.2.1 Experimental Results

In the white Gaussian noise case, the wavelet thresholding using four threshold selection rules have been compared. Wave Risk, Wave Fixed, Wave Heuristic, Wave Minimax represent Stein's unbiased estimate of the risk, heuristic rule, fixed rule, minimax-performance rule, respectively. Their performances are compared in the following Figure 8 and Figure 9.

From Figure 8 and Figure 9, we find that four types of threshold selection rules yield fairly close performances with Wave Risk (Stein's unbiased estimate of the risk selection rule) being the best. Therefore, we choose Wave Risk for final comparisons with other enhancement algorithms.

In the white Gaussian noise case, the bionic wavelet thresholding using the same four threshold selection rules have been tested and their performances have been compared in the following Figure 10 and Figure 11.

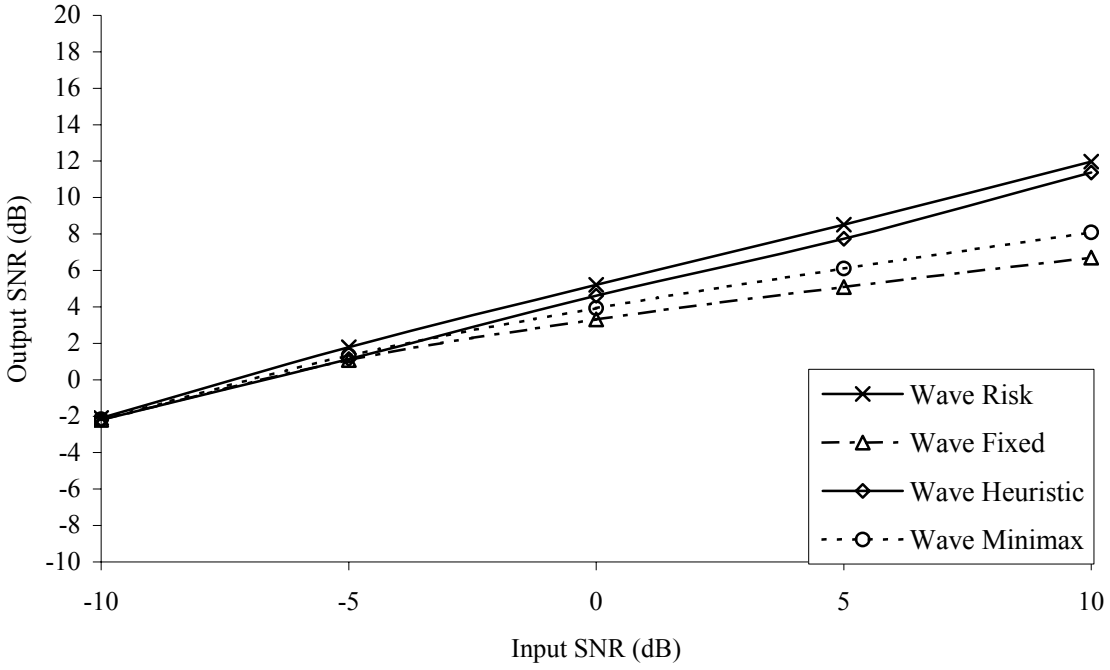

Figure 9 SNR results for wavelet thresholding in white Gaussian noise case

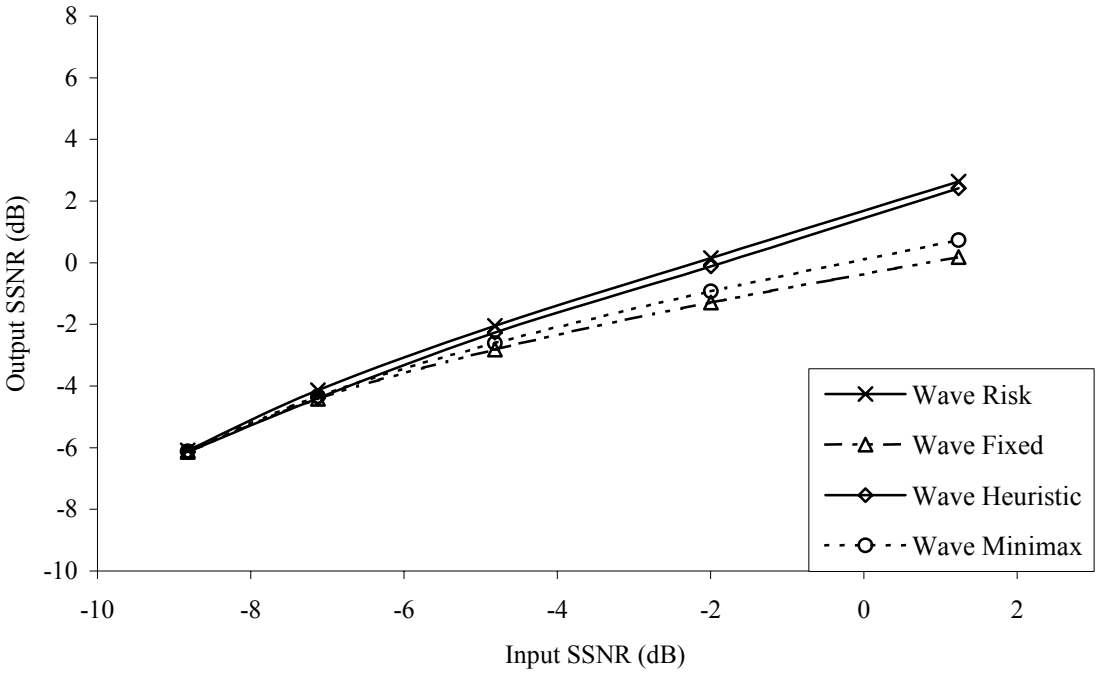

Figure 10 SSNR results for wavelet thresholding in white Gaussian noise case

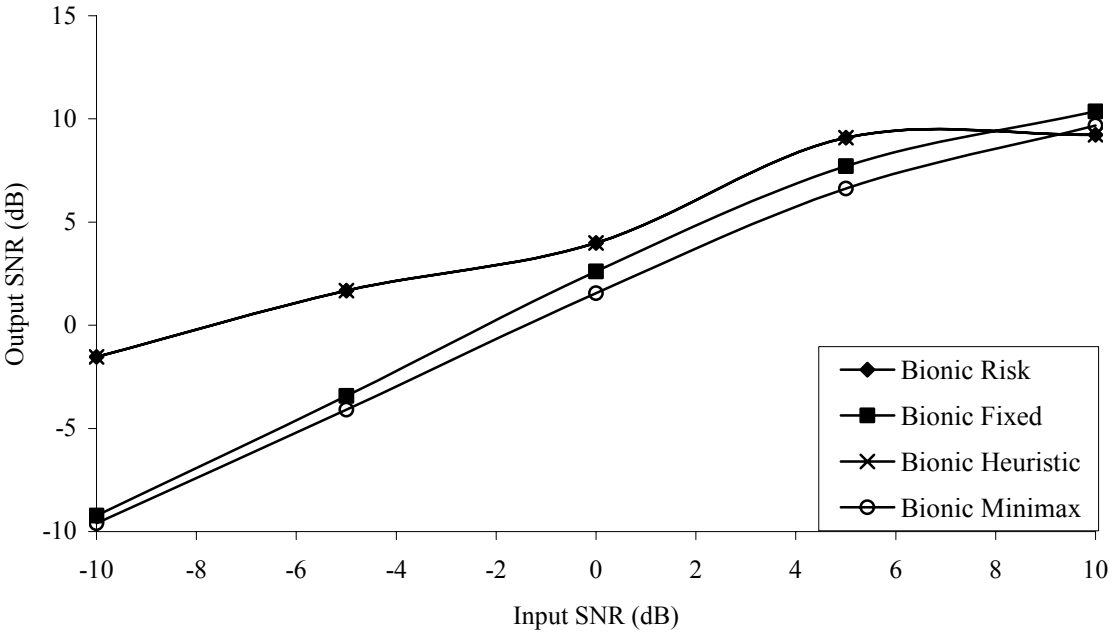

Figure 11 SNR results for bionic wavelet thresholding in white Gaussian noise case

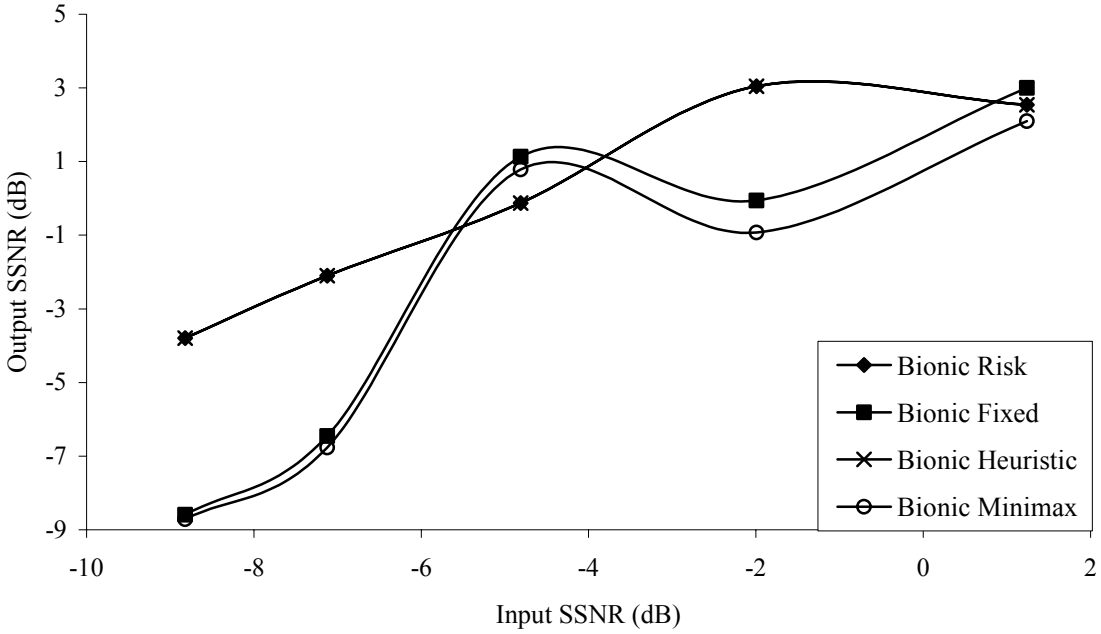

Figure 12 SSNR results for bionic wavelet thresholding in white Gaussian noise case

From Figure 10 and Figure 11, we find that Bionic Risk yields the same results as Bionic Heuristic and both have consistently better performances than the other two rules. We therefore choose Bionic Risk for final comparisons with other enhancement methods.

In the white Gaussian noise case, the comparison between traditional methods and the bionic wavelet thresholding method is to be shown in the following Figure 12 and Figure 13.

Form Figure 12 and Figure13, Ephraim Malah filtering, as expected, proves its efficiency as the best and noticeably the bionic wavelet thresholding method using Stein's unbiased estimate of the risk rule performs better than the corresponding wavelet thresholding method. The bionic wavelet thresholding using Stein's unbiased estimate of the risk rule also outperforms iterative wiener filtering and spectral subtraction.

In the real-life noise case where f16 cockpit noise, car interior noise pink noise and multiple-talker babble noise are all at 0dB input SNR level, we first compare the wavelet thresholding using four threshold selection rules in the following Figure 14 and Figure 15.

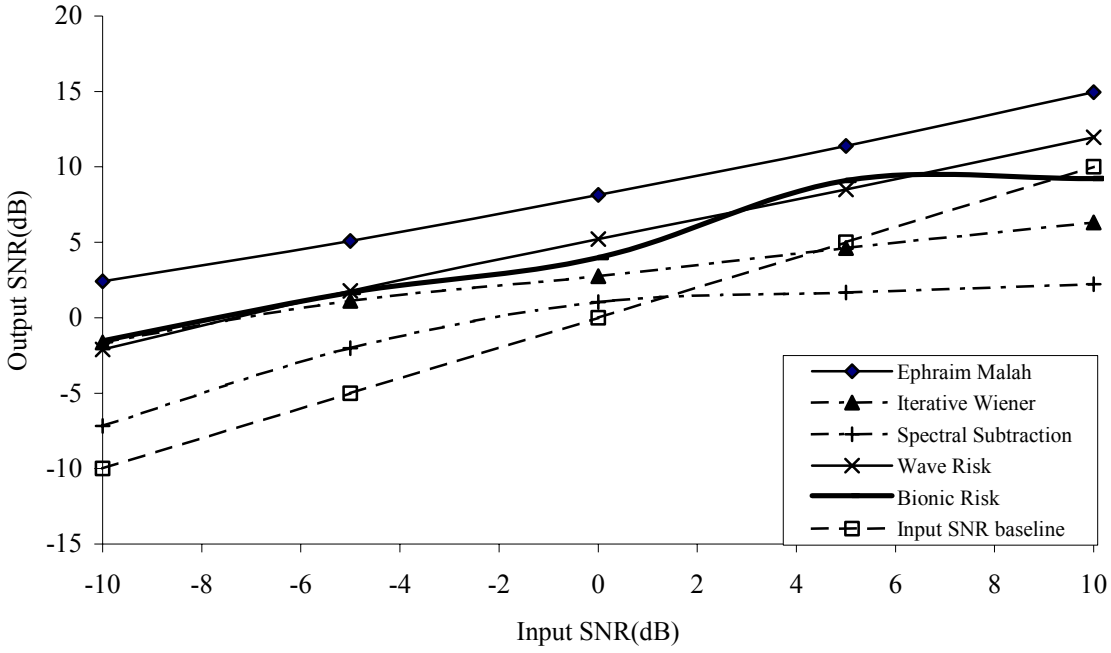

Figure 13 SNR comparisons of five methods for white Gaussian noise

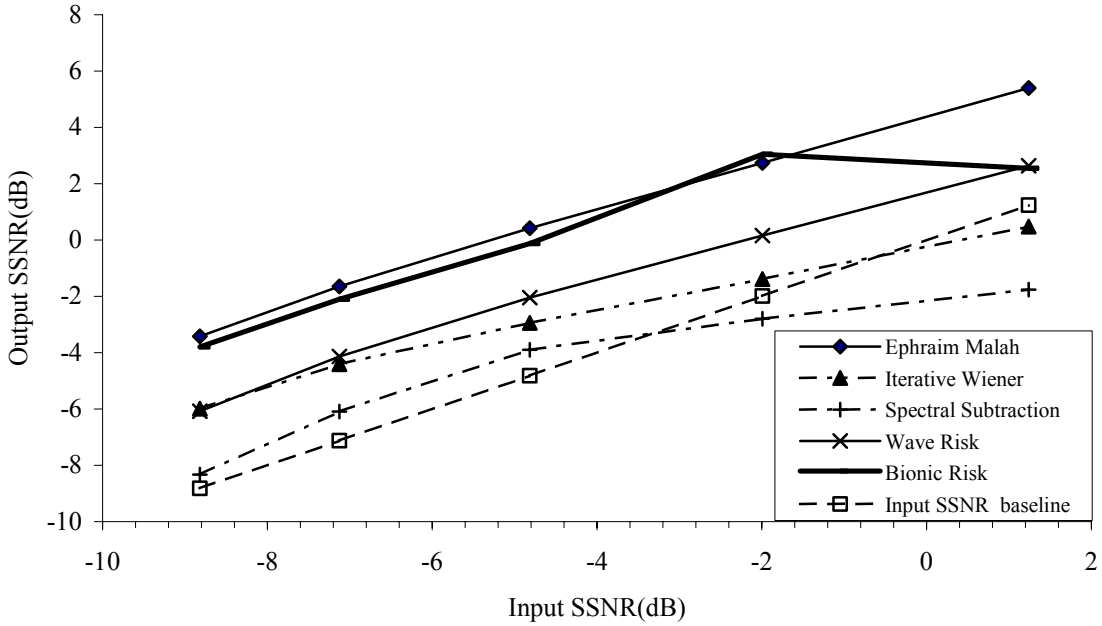

Figure 14 SSNR comparisons of five methods for white Gaussian noise

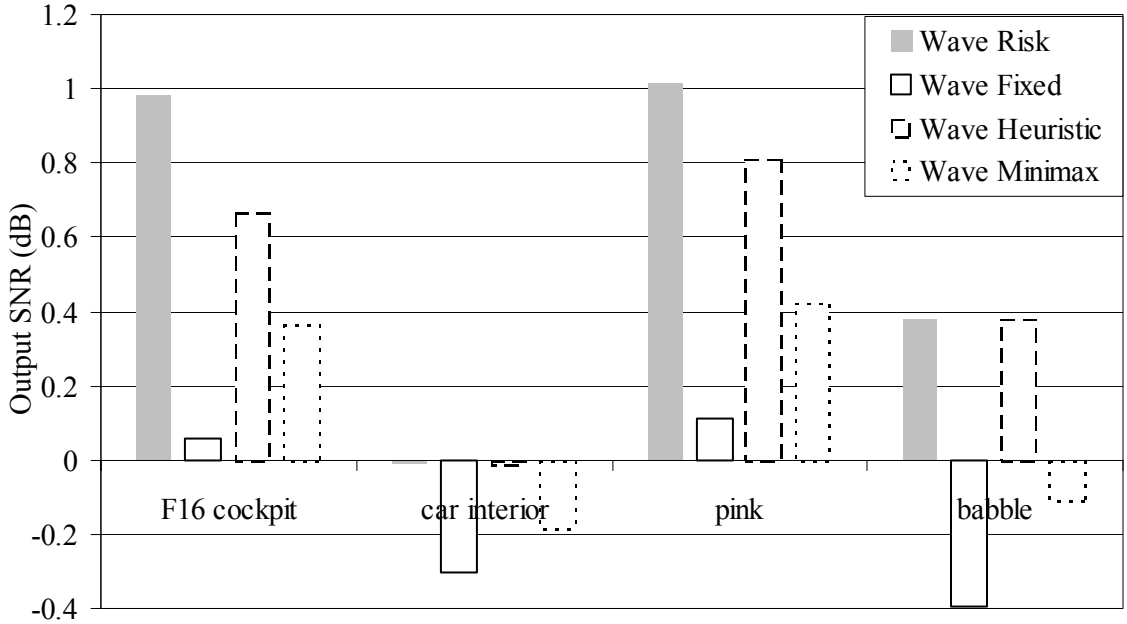

Figure 15 SNR comparison for wavelet thresholding in the real-life noise case

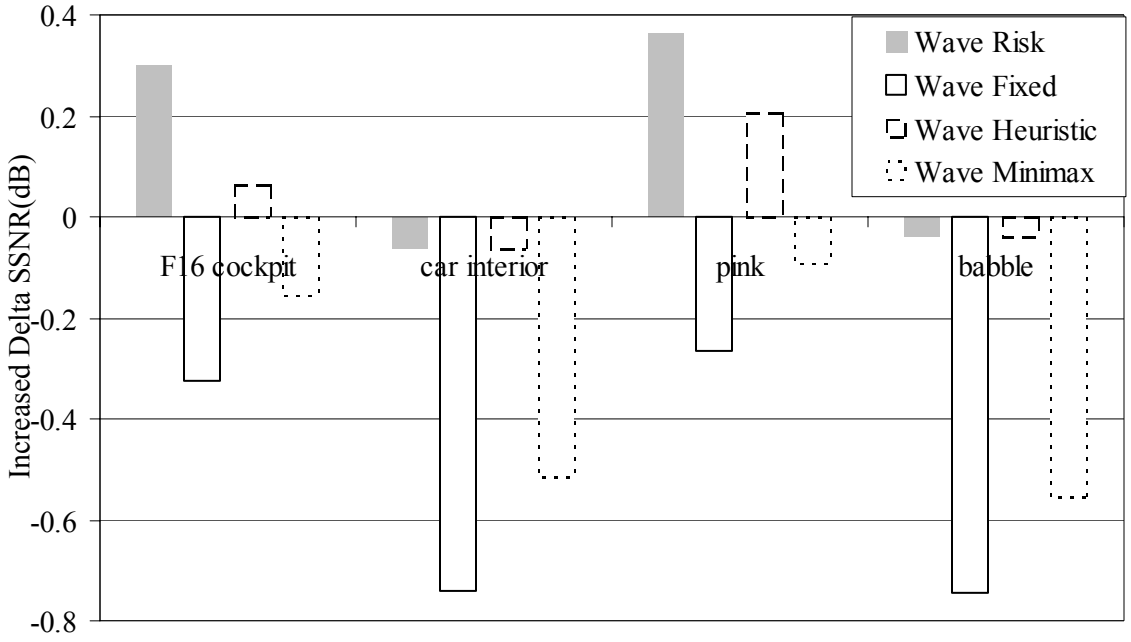

Figure 16 SSNR comparisons for wavelet thresholding in the real-life case

From Figure14 and Figure15 we find that Wave Risk again shows the best performance, as expected. We therefore choose Wave Risk for final comparisons with other types of enhancement algorithms.

In the real-life noise case, we compare the bionic wavelet thresholding using four selection rules in the following Figure 16 and Figure17.

From Figure 16 and Figure 17, we notice that the bionic wavelet thresholding using Stein's unbiased estimate of the risk rule and its heuristic variant option both perform the best on three types of noise and the worst on car interior noise. For the convenience of comparison, we still choose Bionic Risk for final comparisons with other types of enhancement methods.

In the real life noise case, the comparison between traditional methods and the bionic thresholding using Stein's unbiased estimate of the risk rule is to be shown in the following Figure 18 and Figure 19.

From more reliable SSNR comparisons in Figure19, on four kinds of real life noise, we notice Ephraim Malah filtering, as expected, proves its efficiency as the best and noticeably the bionic wavelet thresholding method using Stein's unbiased estimate of the risk rule again performs better than the corresponding wavelet thresholding method. Bionic wavelet thresholding using the risk rule also outperforms iterative Wiener filtering and spectral subtraction.

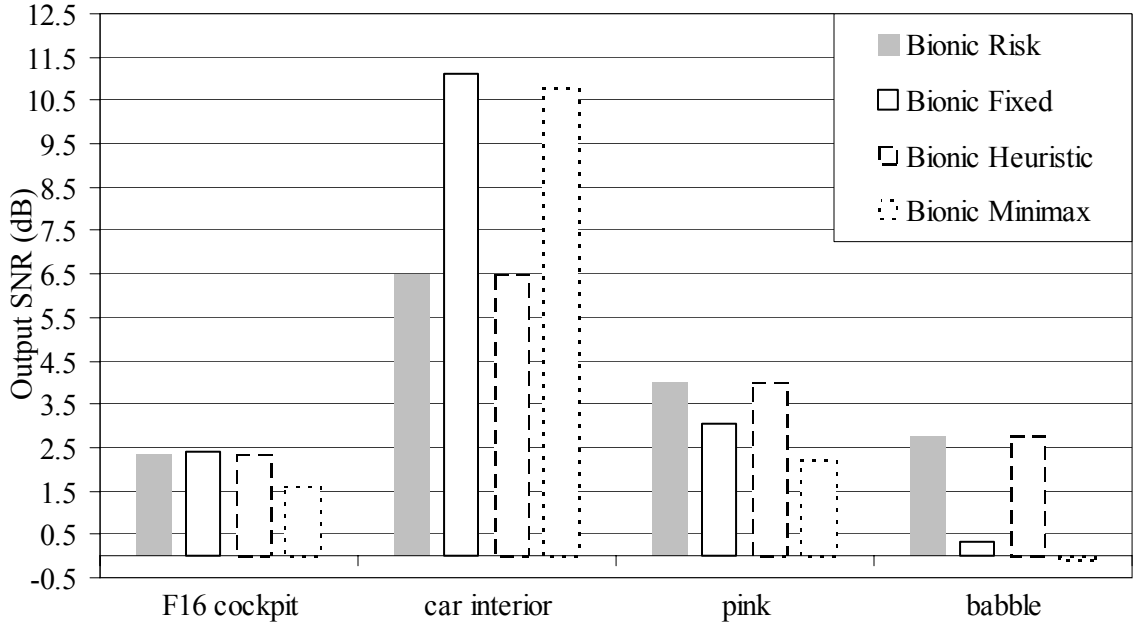

Figure 17 SNR results for bionic wavelet thresholding in the real life noise case

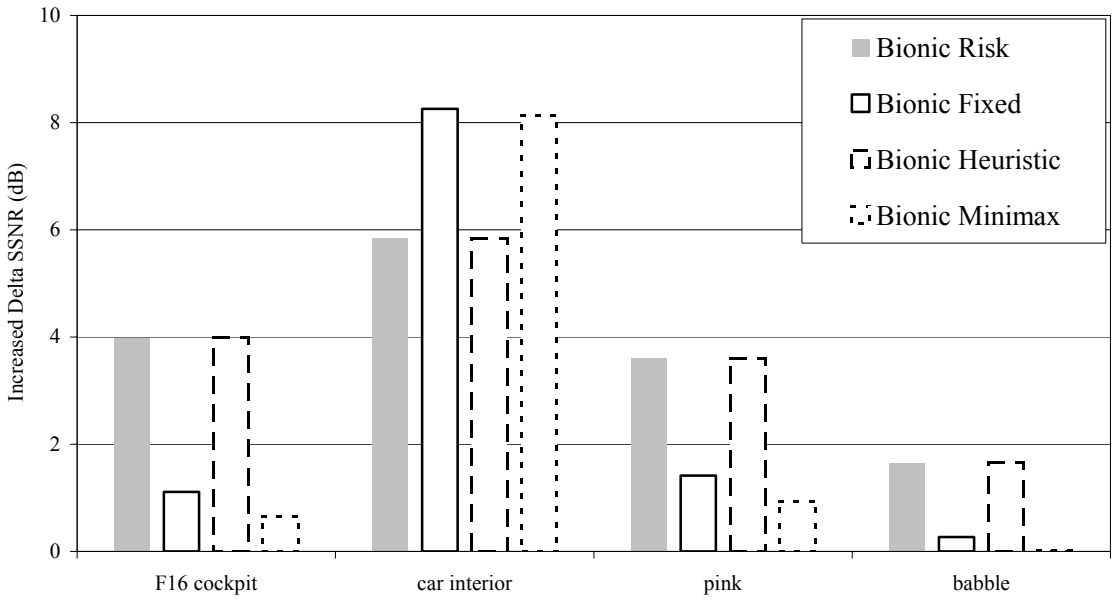

Figure 18 SSNR results for bionic wavelet thresholding in the real life noise case

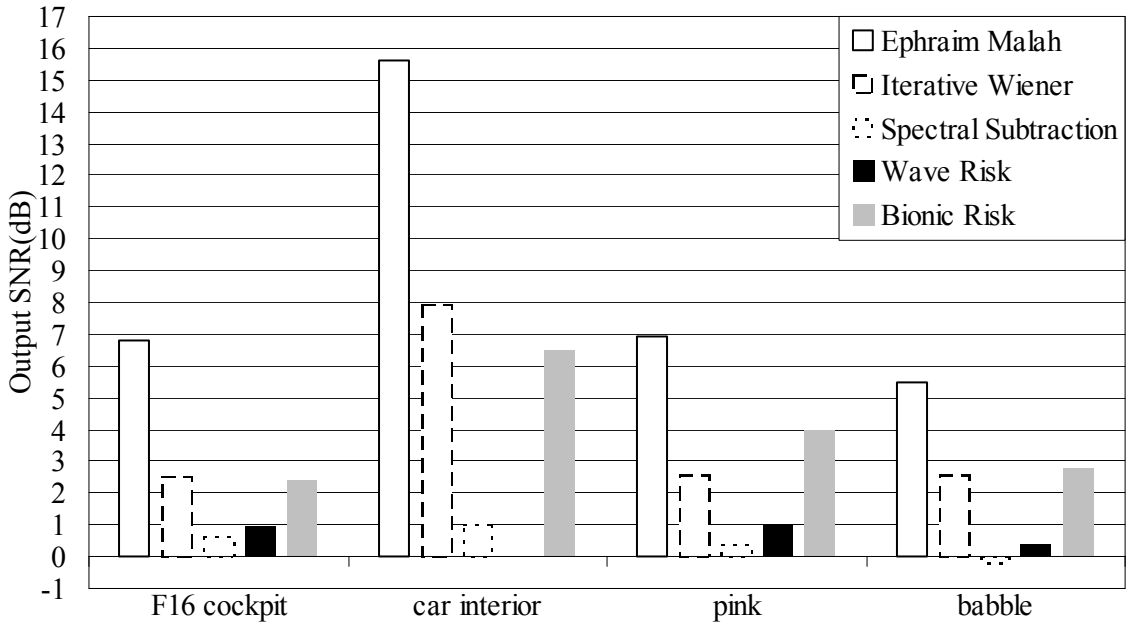

Figure 19 SNR comparisons of five methods for the real life noise

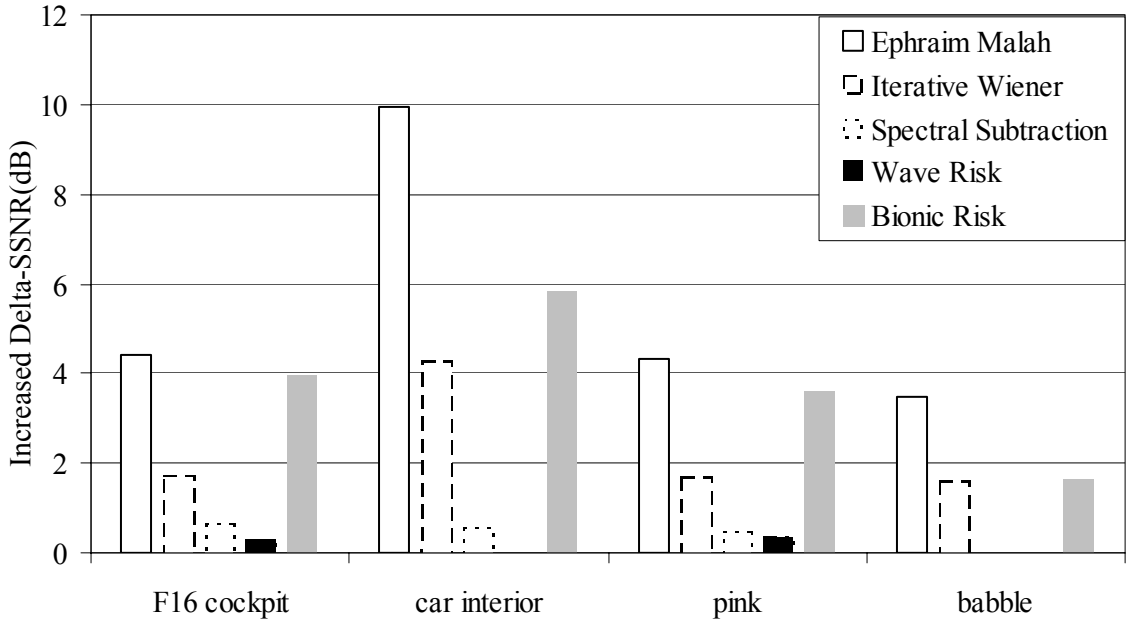

Figure 20 SSNR comparisons of five methods for the real life noise

4.3 Subjective measure

The best speech quality measurement requires a subjective judgment by a listener as to how "good" speech material sounds. Subjective speech quality measures have varied forms. One method is to ask a group of listeners to rank the quality of speech along a predetermined scale after the comparisons of original and processed speech data (Deller 1994). The mean opinion score (MOS) test is the most widely used subjective quality measure (Deller 1994). In this method, listeners rate the speech on a five-point scale where a listener's subjective impressions are assigned a numerical value.

Speech intelligibility is normally measured by having human subjects listen to recorded or spoken speech and subsequently write or otherwise specify what they understood to have

been spoken. However this testing is quite time-consuming and tedious and involves a consideration of different factors, which can make tests not directly comparable under differing conditions. (Niederjohn 1995). One commonly used test for intelligibility is the diagnostic rhyme test (DRT) that requires listeners to circle the word spoken among a pair of rhyming words.

4.3.1.Experimental Results

Our tests were performed by a group of 15 listeners with no previous familiarity with the test materials. Each subject participated in listening to four different sentences corrupted by four types of real-life noise with each sentence being processed by 5 different methods, totaling 20 test samples. Each listener's scores are tabulated. An overall quality score is then calculated after averaging the 15 listening results and illustrated in Figure 20.

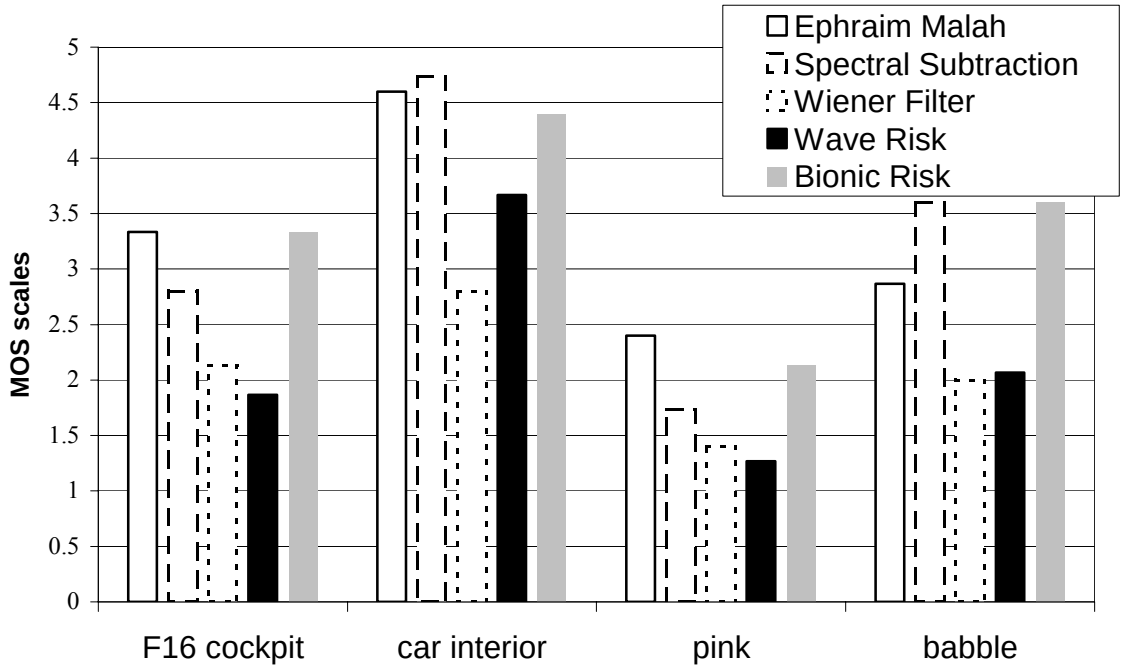

Table 2 Mean opinion score result

Chapter 5 Summary and Conclusions

5.1 Analysis and Conclusions

In our experiments, the thresholding based on the risk rule has shown consistently the best performance, which is also reflected by the fact that the risk rule is the most commonly used for basic wavelet thresholding due to its sound theoretical underpinning. In dealing with white Gaussian noise case, five methods were compared and the results have shown obviously that bionic risk thresholding performs far better than its wavelet counterpart. My guess of the cause to this phenomenon is due to more practical consideration of the human cochlear mechanism in the bionic wavelet transform than in the wavelet transform. Meanwhile, the bionic risk method shows a sudden performance drop when dealing with white Gaussian noise at 10 dB SNR level. The possible cause of this result could be that the thresholding techniques we employed were derived originally for the discrete wavelet transform and therefore they don't fit with the nonlinear nature of the bionic wavelet transform. When facing the situation where there is strong speech signal presence and weak noise presence, the threshold doesn't distinguish well between the target signal and the noise and therefore most likely, the speech contents could have been thresheld out mistakenly as noise, which finally leads to the sudden dropping of its performance. As to the real-life noise, the nature of the different noises is very irregular and non-stationary, and pink and car interior noises appear some low-pass characteristics. Applying a simple threshold rule for all four noises with diverse nature is not a good idea and this is one issue we should further address in the future research. In terms of the testing results, the bionic risk method didn't perform the best in dealing with car interior noise that has strong low-pass nature. This

unsatisfactory performance is, I suppose, due to the over-thresholding effect because car interior noise is low-pass noise with little energy and the bionic risk estimator again behaves poor in adjusting itself according to noise characteristics. In the final comparison of five methods, bionic thresholding again did significantly outperform its wavelet counterpart as we had expected.

For both white Gaussian and real-life noise cases, the performance of the bionic risk method is consistently second to that of Ephraim-Malah filtering. The two algorithms are on the same computational scale and Ephraim Malah obviously stands out as the best method so far. If we could come up with a more reasonable threshold estimator exactly based on the bionic wavelet transform, will this improved bionic thresholding method outperform Ephraim-Malah filtering? We cannot answer this question based on our research efforts for now and it is another important issue to be addressed in the future.

5.2 Suggestions for Future Research

The bionic wavelet transform was originally derived from using Morlet wavelet due to the fact that the auditory model had employed the Morlet wavelet to achieve the best compromise of time and frequency resolution [Yao, 2001]. Because the Morlet wavelet does not support the fast discrete wavelet transform, this method can only be implemented through numerical integration although using 22 analyzing scales ends up at the same computation intensity as Ephraim Malah filtering. If we can replace the Morlet wavelet with other mother wavelets that have scaling functions and therefore allow a fast discrete algorithm, this method could be put

into applications such as cochlear implants for which this new time-frequency analysis method was originally intended (Yao 2002).

Secondly, as for the marriage of discrete wavelet based thresholding and the bionic wavelet transform for speech enhancement, there are many things to be further investigated. The traditional wavelet thresholding developed by Dr. Donoho was based on the discrete wavelet transform to come up with an asymptotically optimal threshold estimator (Donoho 1995). If we can migrate from the continuous bionic wavelet transform into the discrete bionic wavelet transform, we would expect to derive an appropriate threshold estimator that benefits from the strength of the auditory model.

Thirdly, the bionic wavelet transform is a subset of adaptive wavelet transform and it stands out particularly due to its consideration of human perception for the adaptation of the speech signal, while other adaptive wavelet transforms, say wavelet packets (Wickerhauser 1994; Strang.G 1996), are all based on some mathematical or statistical entropy criterion. Wavelet packets and other adaptive wavelet transforms have already proven to have far better performance than the standard wavelet transform (Mallat 1998) and hence the development of the bionic wavelet transform indicates a promising path in advanced signal processing.

References

Bahoura, M. a. J. R. (2001). "Wavelet speech enhancement based on the Teager engergy operator." IEEE Signal Processing Letters **8**(1): 10-12.

Berouti, M. (1979). Enhancement of Speech Corrupted by Acoustic Noise. International conference on audio,speech and signal processing of IEEE.

Bertrand, J. (1990). Discrete mellin transform for signal analysis. International conference on audio, speech and signal processing, IEEE.

Boll, S. (1979). "Suppression of acoustic noise in speech using spectral subtraction." IEEE Trans. on Acoustic,Speech Signal Processing **27**(Apr.): 113-120.

Branson, L. K. (1997). Ph.D dissertation:The application of moment and cumulant spectra to formant tracking of speech embedded in noise. Dept. of Electrical and Computer Engineering, Marquette University.

Chang, S., Y.Kwon (2002). Speech enhancement for non-stationary noise enviroment by adaptive wavelet packet. International conference on audio,speech and signal processing of IEEE.

Conway, R. (1994). Feature based speech intelligibility enhancement in high noise level. Ph.D dissertation: Dept. of Electrical and Computer and Engineering. Milwaukee, Marquette University.

Conway, R., Heinen,J.,Niederjohn,R. (1990). "Evaluation of a technique involving adaptive processing with feature extraction to enhance the intelligibility of noise corrupted speech." Proc. of IEEE,International conference on industrial electrical, control and instrumentation: pp.28-33.

Daubechies, I. (1992). Ten lectures on wavelets. Philadelphia, USA, Society for Industrial and Applied Mathematics.

Deller, J. R., Proakis,J.G.,Hansen,J.H.L. (1994). Discrete-time processing of speech signals. New York, Macmillan Publishing Company.

Deng, L. a. G., C.D. (1988). "A composite model of the auditory periphery for the processing of speech." J. Phonet. **16**(1): 93.

Donoho, D. L. (1995). "Denosing by soft thresholding." IEEE Trans. on Information Theory **41**(3): 613-627.

Ephraim, Y., D. Malah (1984). "Speech enhancement using a minimum mean square error short time spectral amplitude estimator." IEEE Trans. on Acoustic,Speech Signal Processing **32**(6): 1109-1121.

Ephraim, Y., Van Trees,H.L. (1995). "A signal subspace approach for speech enhancement." IEEE Trans. Speech and Audio Processing **3**(4): 251.

Furst, M. a. L., M. (1988). "A cochlear model for acoustic emissions." J. Acoustic Society of America **84**(1): 222-229.

Gelfand, S. A. (1998). Hearing: an introduction to psychological and physiological acoustics. New York, Marcel Pekker Inc.**:** p 128-134.

Giguere, C. a. W., P.C. (1994). "A computational model of the auditory periphery for speech and hearing research, I ascending path." J. Acoustic Society of America **95**(1): 331-342.

Guinan, J. J. J., Peake,W.T. (1967). "Middle ear characteristics of anesthetized cats." Journal of the Acoustical Society of America **41**: 1237-1261.

Guo, D. (2000). A study of wavelet thresholding denoising. IEEE International conference on signal processing.

Hansen, J. H. L., Clements,M.A. (1985). Enhancement of speech degraded by non-white additive noise, Techical report submitted to Lockheed Corp.,DSPL-85-6,Georgia Institute of Technology.

Haykin, S. (1986). Adaptive Filter Theory. Englewood Cliffs, NJ, Prentice-Hall, Inc.

Holschneider, M. (1989). Wavelet, time frequency methods and phase space. First international conference on wavelet, Springer-Verlag.

Hubbard, A. E., Mountain,D.C. (1995). Models of the cochlea. New York, Springer.

Johnstone, I. M. a. S., B.W. (1997). "Wavelet threshold estimators for data with correlated noise." J. Royal Statistical Society **59**(2): 319-351.

Jones, D. (1991). "Efficient approximation of continuous wavelet transform." Electronic letters, IEEE **27**(9): 748-750.

Kemp, D. T. (1978). "Simulated acoustic emissions from within the human auditory system." J. Acoustic Society of America **64**(5): 1386-1391.

Lim, J. S., Oppenheim, A.V. (1979). "All-Pole modeling of degraded speech." IEEE Trans. on Acoustic,Speech Signal Processing **26**(3): 197-210.

Lim, J. S., Oppenheim, A.V. (1979). "Enhancement and bandwidth compression of noisy speech." Proc.of the IEEE **67**(12): 1586-1604.

Mallat, S. (1989). "A theory for multiresolution signal decomposition." IEEE TRans. on Pattern Anal. Machine Intell.(11): 674-693.

Mallat, S. (1998). A wavelet tour of signal processing. New York, Academic Press.

Mathworks. (1998). Matlab Wavelet Toolbox User's Guide, V5.12.

Meyer, Y. (1992). Wavelets and Operators, Cambridge University.

Monzon, L. A. (May,1994). Ph.D dissertation: Constructive Multiresolution Analysis and the Structure of Quadrature Mirror Filters, Yale University.

Niederjohn, R. J. (1995). "Speech intelligibility enhancement in high levels of wideband noise." 1994-1995 Annual Review of Communications.

Probst, R., Lonsbury-Martin,B.L. (1991). "A review of otoacoustic emissions." J. Acoustic Society of America **89**(5): 2027-2067.

Quatieri (2001). Discrete-Time Speech Signal Processing,Principles and Practice, Prentice Hall PTR.

Rubin, P., Vatikiotis-Bateson,E. (1998). Measuring and modeling speech production. Animal Acoustic Communication. M. J. O. S.L. Hopp, and C.S. Evans (Eds). Springer-Verlag, 1998.

Schroeder, M. (1975). "Models of hearing." Proc. IEEE, **63**(9): 1332-1350.

Seok, J. W. (1997). Speech enhancement with reduction of noise components in the wavelet domain. International conference on audio,speech and signal processing of IEEE.

Sheikhzadeh, H. (2001). An improved wavelet-based speech enhancement system. Eurospeech.

Shields, U. C. (1970). Sepration of added speech signals by digital comb filtering. Dept. of Elec. Eng. Cambridge, USA, MIT.

Stein, S. M. (1981). "Estimation of the mean of a multivariate normal distribution." Ann. Statist. **9**(6): 1135-1151.

Strang.G, N., T. (1996). Wavelets and filter banks, Wellesley-Cambridge Press.

Van Compernolle, D. (1993). "Speech Enhancement for Applications in Communication and Recognition." Revue HF Tijdschrift (Special Issue: Speech Processing for Telecommunications) **1993**(1-2-3): 99-108.

Whitehead, M. L., Stagner,B.B. (1994). "Measurement of otoacoustic emissions for hearing assessment." IEEE Eng. Med. Biol., Mag. **13**: 210-226.

Wickerhauser, M. V. (1994). Adapted wavelet analysis from theory to software algorithms, A.K.Peters.

Yang, X. a. W., K (1992). "Auditory representations of acoustic signals." IEEE Trans. on Information Theory **38**(2): 824-839.

Yao, J., Y.T.Zhang (1999). Cochlear is an inhomogeneous,active and non-linear model. Proc. IEEE EMBS/BMES, Atlanta, GA.

Yao, J., Y.T.Zhang (2001). "Bionic wavelet transform: a new time-frequency method based on an auditory model." IEEE Trans. on Biomedical Engineering **48**(8): 856-863.

Yao, J., Y.T.Zhang (2002). "The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations." IEEE Trans. on Biomedical Engineering **49**(11): 1299-1309.

Zheng, L., Y.T.Zhang (1999). "Synthesis and decomposition of transient-envoked otoacoustic emissions based on an active auditory model." IEEE Trans. on Biomedical Engineering **46**(9): 1098-1106.

Appendix: Matlab code for the bionic wavelet transform

```
function wt = mycwt3(x,fs)
% function wt = mycwt3(x,fs)
```

```
% continuous wavelet transform
% x sampled at fs
% frequencies & wavelets made to match those
% used for Bionic Wavelet Transform

x=x(:);

N=22;
f0=5/(2*pi);          % Base frequency for Morlet in Hz

bwtf0=15165.4;
bwta=(exp(log(bwtf0)/64).^(7:N+6))';   % per paper
bwtf=bwtf0./bwta;                  % gives central frequencies 221 - 5291 Hz
a=f0./bwtf;                        % Equivalent

dt=1/fs;
len=length(x);

for n=1:N
   winlen=floor(4*a(n)/dt);    % Makes sure wavelet is exactly zero-centered
   t=dt*(-winlen:winlen);
   wav=(1/sqrt(a(n)))*exp(-(t/a(n)).^2./2).*cos(5*t/a(n));
   wtn=conv(wav,x);
   wt(n,:)=wtn(((length(t)+1)/2):end-(length(t)-1)/2)';
end

function x = myicwt3(wt,fs)
% function x = myicwt3(wt,fs)
% inverse continuous wavelet transform
% x sampled at fs
% frequencies & wavelets made to match those
% used for Bionic Wavelet Transform

[N,len]=size(wt);

f0=5/(2*pi);          % Base frequency for Morlet in Hz
bwtf0=15165.4;
bwta=(exp(log(bwtf0)/64).^(7:N+6))';   % per paper
bwtf=bwtf0./bwta;                  % gives central frequencies 221 - 5291 Hz
a=f0./bwtf;                        % Equivalent
dt=1/fs;

bwtp1=(exp(log(bwtf0)/64).^(7:N+7))';
bwtp1f=bwtf0./bwtp1;
fwidths=diff(bwtp1f);
apct=fwidths/(bwtp1f(N+1)-bwtp1f(1));     % Percent of frequency range used at each scale

for n=1:N
   winlen=floor(4*a(n)/dt);
```

```
    t=dt*(-winlen:winlen);
    wav=(1/sqrt(a(n)))*exp(-(t/a(n)).^2./2).*cos(5*t/a(n));
    wtn=conv(wav,wt(n,:));
    xa(n,:)=wtn(((length(t)+1)/2):end-(length(t)-1)/2);
end

%Won't be quite exact scaling, but nearly
x=dt*sum(xa.*repmat(apct,1,len))';
```