2007

Speech enhancement by perceptual adaptive wavelet de-noising

Lan Xu
University of Windsor

Speech Enhancement by Perceptual Adaptive Wavelet De-noising

by

Lan Xu

A Thesis
Submitted to the Faculty of Graduate Studies
through the Department of Electrical and Computer Engineering
in Partial Fulfillment of the Requirements for
the Degree of Master of Applied Science at the
University of Windsor

Windsor, Ontario, Canada

2007

Your file Votre référence
ISBN: 978-0-494-34968-7
Our file Notre référence
ISBN: 978-0-494-34968-7

ABSTRACT

This thesis work summarizes and compares the existing wavelet de-noising methods. Most popular methods of wavelet transform, adaptive thresholding, and musical noise suppression have been analyzed theoretically and evaluated through Matlab simulation.

Based on the above work, a new speech enhancement system using adaptive wavelet de-noising is proposed. Each step of the standard wavelet thresholding is improved by optimized adaptive algorithms. The Quantile based adaptive noise estimate and the posteriori SNR based threshold adjuster are compensatory to each other. The combination of them integrates the advantages of these two approaches and balances the effects of noise removal and speech preservation. In order to improve the final perceptual quality, an innovative musical noise analysis and smoothing algorithm and a Teager Energy Operator based silent segment smoothing module are also introduced into the system. The experimental results have demonstrated the capability of the proposed system in both stationary and non-stationary noise environments.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor Dr. H. K. Kwan for his guidance and support during the course of my Master program study at University of Windsor; in particular, for introducing me the field of speech enhancement through perceptual adaptive wavelet de-noising, providing me with many papers for background reading, and advising me on the structure and content of the thesis. His advice has been proven very constructive and crucial for my study and research work. I would like to thank my internal reader, Dr. J. Wu, and my external examiner, Dr. R. Gaspar, for the very valuable comments on improving the clarity and readability of the thesis. I would like also thank Ms. A. Turner for her consistent support.

I am deeply indebted to my parents, Y.E.Yu and G.L.Xu, for their sacrifices and support throughout my life. I learned from them the value of knowledge and the passion of the world which are crucial for my life. I am very grateful to my husband X. B. Tang for his endless support. This thesis would not have been possible without his encouragement. I am grateful to my brother Y. Xu for providing much needed moral support.

My thanks also go to my fellow graduate students of the ISP lab, S. Arniker, T. Hila, R. Atwal, and J. Thevaril, for their genuine friendship. Moreover, I would like to sincerely thank all my dear friends, Q. He, Zh. Q. Ji, A. M. Jiang, Zh. Lan, B. X. Li, J. J. Li, J. S. Liao, Y. J. Mao, G. Q. Shu, J. L. Tang, A. Q. Wang, F. Wang, R. Q. Wang, N. Xing, J. Wiebe, X. Y. Xu, H. M. Zhang, Y. B. Zhou, and H. M. Zong, who have encouraged and supported me in the course of my graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbol	Definitions
t	Continuous-time index
$s(t)$	Continuous-time clean speech signal
$w(t)$	Continuous-time noise signal
$y(t)=s(t)+w(t)$	Continuous-time noisy speech signal
n	Discrete time index
$s(n)$	Discrete-time digital clean speech signal
$w(n)$	Discrete-time digital noise signal
$y(n)=s(n)+w(n)$	Discrete-time digital noisy signal
ω	Frequency
$F(\omega)$	Fourier coefficients
ψ	Wavelet function
a	Scale factor
k	Shift factor
$g[n]$	Low pass filter
$h[n]$	High pass filter
N_L	Maximum decomposition level
N	Length of the noisy signal in wavelet domain
Y	Wavelet domain noisy signal
T	Wavelet domain threshold
$\hat{\sigma}$	Standard deviation of the noise
c	Wavelet transform coefficient
λ	Soft threshold
$W(y)$	Forward wavelet transform operator
$W^{-1}(y)$	Inverse wavelet transform operators
$D(Y,\lambda)$	De-noising operator
$d(Y)$	Adaptive threshold operator
z	Critical band index (in Bark)

f	Frequency (in Hz)
$BW(f)$	Bandwidth of critical band
$Tq(f)$	Threshold in quiet at frequency f
$SF(z')$	Spreading function
z'	Separation between critical bands
Gm	Geometric mean
Am	Arithmetic mean
SFM	Spectral flatness measure
α'	Tonality factor
$O(i_c)$	Threshold offset
i'	Bark frequency of the masking signal
i	Coefficient index
j	Subband index
$c_{i,j}(n)$	i th coefficient in subband j, frame n
$Px_j(n)$	Energy of subband j
$PxS_j(n)$	Energy of subband j, convolved with spreading function
$Pn_j(n)$	Noise estimated power
$\overline{P}n_j(n-1)$	The average of the noise estimates of the previous m frames adjacent to frame
NMT	Simultaneous masking threshold
q	Quantile parameter
L	Frame number in a segment
$PSNR_j(n)$	Posteriori signal-to-noise ratio
$\alpha_j(n)$	Smoothing parameter
$\psi[y(n)]$	Discrete-time TEO
$Teo_{i,j}(n)$	TEO coefficients

LIST OF ACRONYMS

Acronym	Definition
AHT	Absolute Threshold of Hearing
CB	Critical Band
CBW	Critical Bandwidth
DFT	Discrete Fourier Transform
STFT	Short-time Fourier Transform
WT	Wavelet Transform
CWT	Continuous Wavelet Transform
DWT	Discrete Wavelet Transform
DWPT	Discrete Wavelet Packet Transform
WPD	Wavelet Packet Decomposition
PWT	Perceptual Wavelet Transform
SURE	Stein's Unbiased Estimate of Risk
SPL	Sound Pressure Level
NMT	Noise Masking Threshold
SNR	Signal-to-noise Ratio
QBNE	Quantile-based Time-frequency Noise Estimate
PSNR	Posteriori Signal-to-noise Ratio
segNR	Segmental Signal-to-ratio
VAD	Voice Activity Detector
TEO	Teager Energy Operator
WGN	White Gaussian Noise
MNSN	Mixed Non-stationary Noise
PSNRAT	Posteriori SNR Time-Adaptive Threshold
SHTAT	Smoothed Hard Thresholding with Aggravated Threshold Value
TEOAT	Teager Energy Operator based Adaptive Threshold

CHAPTER I

1. INTRODUCTION

1.1 Speech Enhancement

Speech enhancement is the term used to describe algorithms or devices whose function is to improve the perceptual quality or decrease the hearing fatigue of a noisy speech. The application of speech enhancement includes multimedia and wireless communications, air-ground communication systems in which the pilot's speech is corrupted by cockpit noise, teleconference systems and paging systems, etc. And it can also work as a front-end processing module to increase the robustness of speech processing applications.

In literature, a number of speech enhancement techniques have been proposed in the recent three decades [Ephraim2003] [Gustafsson2001] [Zhang2003]. According to the number of channels used in the noise suppression, these techniques can be classified into the single-channel systems or the multi-channel systems. Multi-channel systems use two or multiple channels in the speech noise suppression process, of which the dual-channel systems are most commonly seen. Although these systems are powerful, especially in suppressing noises corrupted by nonlinear models, they are complicated and expensive. Single-channel systems only use one channel in the speech noise reduction process. Its principle is illustrated in Figure1-1. Where, $s(n)$ is the digital representation of clean speech signal, $w(n)$ denotes the noise signal, and $y(n)$ represents the noise corrupted speech signal. The major task for a single-channel system is to design an effective and efficient noise suppressor module, which could precisely recover the original clean speech from a noisy input without excessive spectral distortions. Although a single-channel system's performance is highly limited by the noise conditions, it is used widely because it is easier and less costly to build. In this thesis, only a single-channel system is considered.

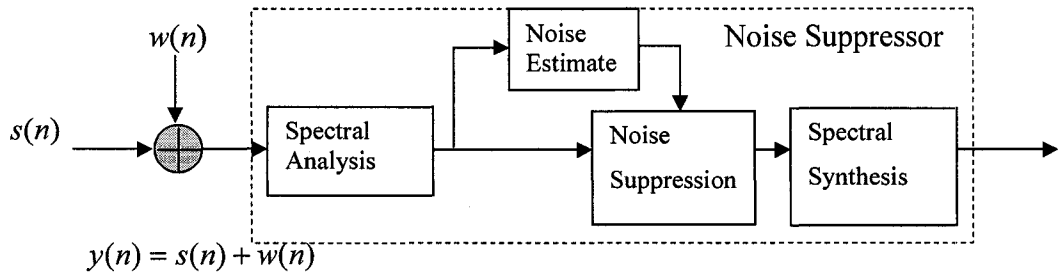


Figure 1-1 Single-channel speech enhancement system

As shown in Figure 1-1, the noise suppressor usually consists of four parts, spectral analysis, noise estimate, noise suppression, and spectral synthesis. The most popular spectral analysis method is Fourier transform. It provides the frequency response of signal that helps in differentiating signal and noise. However, the time-domain information is lost. To improve the performance of time-domain analysis, the wavelet transform has been studied widely in the recent twenty years. In this section, speech enhancements based on both Fourier transform and wavelet transform are discussed. Before that, the human auditory system is introduced.

1.1.1 Human Auditory System

The hearing system converts sound waves into mechanical energy and finally into electrical impulses perceived by the brain. It consists of the ear, auditory nerve fibers and a part of the brain. The ear contains three parts, i.e., the outer ear, the middle ear and the inner ear. The structure of it is shown in Figure 1-2 [Web1].

The outer part of the ear consists of the pinna (auricle), the ear canal (external auditory meatus) and the eardrum (tympanic membrane). The sound pressure in the air is collected by the pinna, amplified and conveyed by the ear canal, and then makes the ear drum vibrate. The sound energy is converted into the mechanical energy in this way.

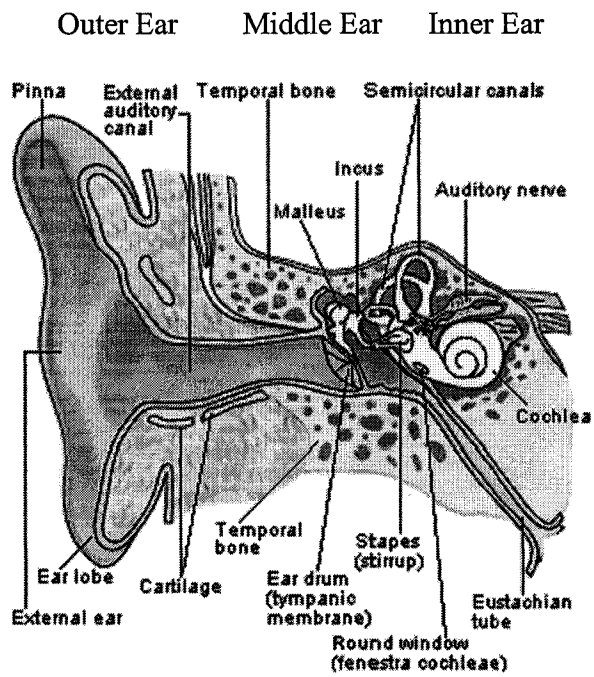


Figure 1-2 Simplified structure of the ear

[Web1]

The middle ear is an air-filled space containing the three smallest bones in the human body, including the hammer (malleus), anvil (incus) and stirrup (stapes). These bones form a system of levers which vibrate along with the eardrum. This vibration amplifies the sound and carries it to the inner ear via the oval window.

The inner ear has a great role in both hearing and the body balance. The hearing organ is a bony cone-shaped spiral called cochlea which is filled with fluids. The Cochlea is the part of the inner ear which converts incoming vibrations from the middle ear into the electrical impulses. The frequency-dependent response of the cochlea is an important feature for both speech enhancement and coding research. Especially, the frequency selectivity of masking effects, generally described in terms of Critical Bands (CB), can be used to lighten the over suppression, and increase the coding efficiency.

The frequency function of the cochlea can be best modeled as a set of continuous differential equations. However, for implementation purposes, it is normally modeled in

discrete sections as a bank of bandpass filters [Gui2005]. Although the modeling of the cochlea function has been an active research area for many years, there are still ambiguities in its mechanism such as the frequency selectivity of the auditory system and the nonlinear behavior of the cochlea.

1.1.2 Speech Enhancement Based on Fourier Transform

De-noising is a process of deriving an estimator of the original signal from the observed corrupted signal. However, it is always difficult to separate the speech signal and the noise signal in time domain. The difference between the original signal s and noise w may be more obvious in other domains. Noise is often considered to have more high frequency energy than the normal signal. Thus, in frequency domain, removing the high frequency contents of the corrupted signal y may reduce the influence of the noise w, as illustrated in Figure 1-3.

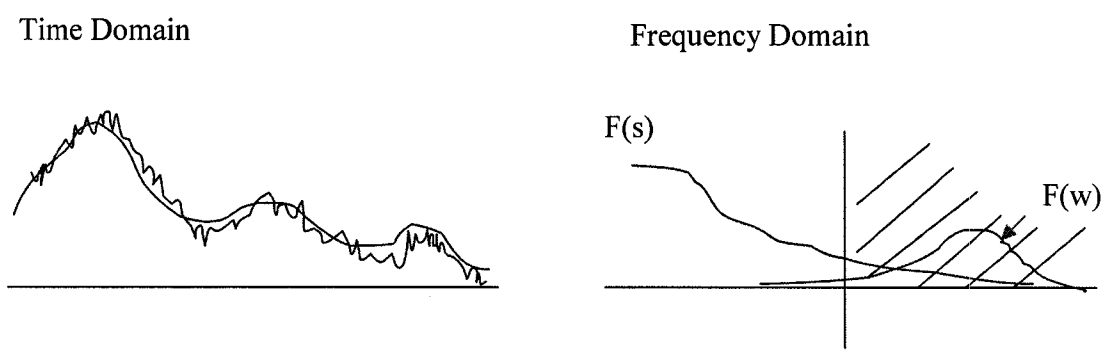


Figure 1-3 Time domain to frequency domain analysis

The Fourier transform is a traditional tool to convert the time domain signal into the frequency domain. Many speech enhancement methods have been developed based on the Fourier transform, such as Wiener Filtering, Iterative Wiener Filtering, Improved Iterative Wiener Filtering, Constrained Iterative Wiener Filtering, and the most popular Spectral Subtraction. The Fourier transform helps in differentiating signal and noise by giving the frequency response of the signal. However, the methods based on it tend to distort the signal since the high frequency component of the signal will also be removed.

Even if the high frequency components of some signal should not be removed, the Fourier approach will still remove it since de-noising cannot be localized. Conversely, for those parts whose high frequency components should be removed, the Fourier approach cannot particularly take care of it. In the mean time, time information is lost after the Fourier transformation, which can be observed in Figure 1-4 (b).

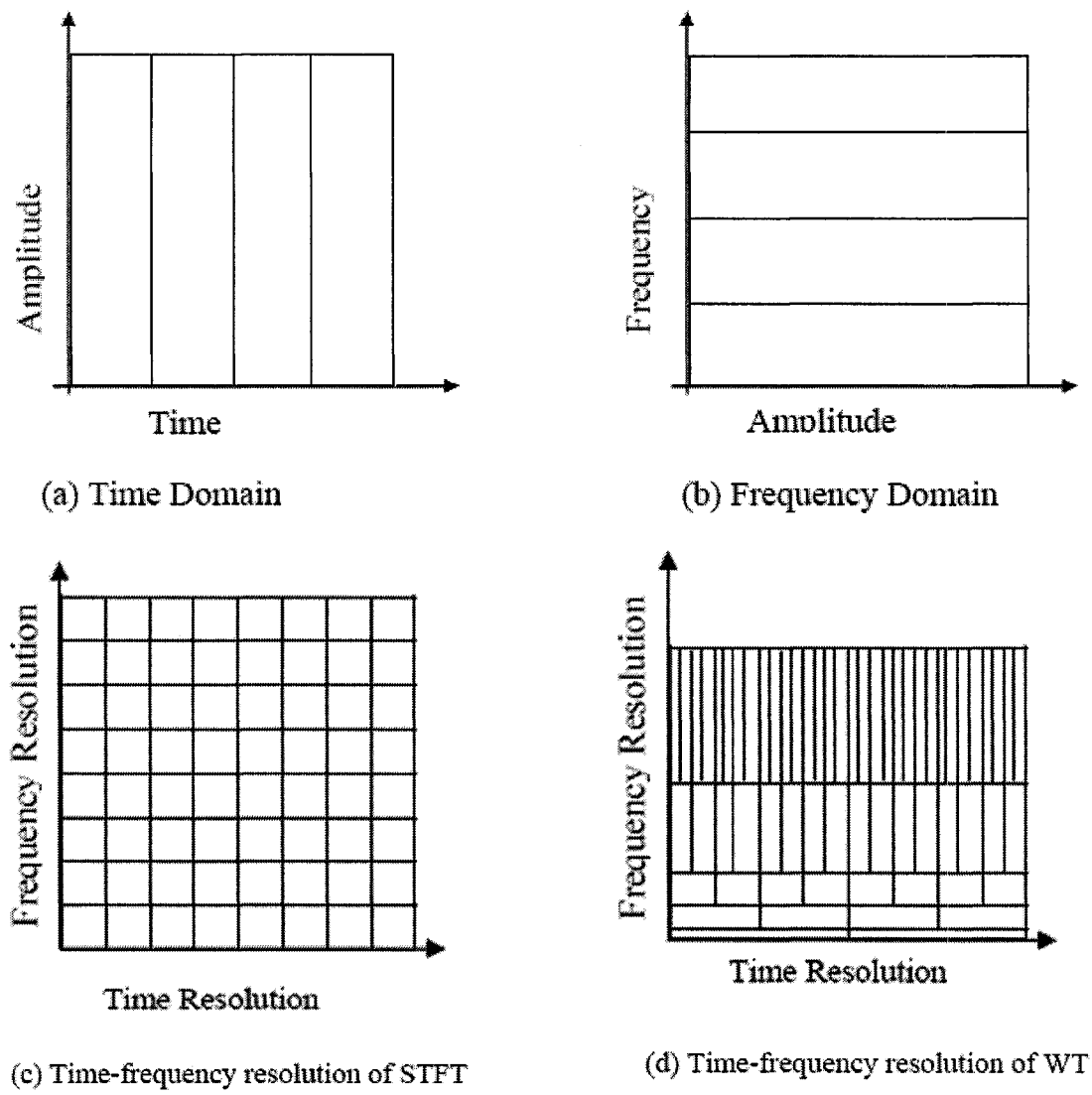


Figure 1-4 Different domain description

One main assumption in using DFT for calculation of the spectrum of a discrete signal is

that the observed signal is stationary during the observation time. In other words, the spectrum of the signal is assumed to remain the same during the observation time. For most practical signals, this assumption is not valid. For example, in speech signals, the spectrum of the signal may vary significantly from one point to another. This depends on the contents of the speech and the sampling period.

The short-time Fourier transform (STFT), is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. The Fourier transform is modified such that a two-dimensional time-frequency representation of the signal is obtained. This method depends on a window function as shown in Figure 1-4 (c).

The main purpose of the window in the time-dependent Fourier transform is to limit the extent of the transformed sequence so that the spectral characteristics are reasonably stationary over the duration of the window function. The more rapidly the signal characteristics change, the shorter the window should be. The resolution in frequency depends on the duration of the window function.

In discrete STFT (DSTFT), the fine resolution in the frequency domain is corresponding to the relative wide window in time domain which may not be proper since the signal is assumed short-time stationary. Based on this trade off, the window function is determined. In general, for DSTFT, after selecting the window function, the frequency and time resolutions are fixed for all frequencies and all times respectively. This approach does not allow any variation in resolutions in terms of time or frequency.

1.1.3 Speech Enhancement based on Wavelet transform

Wavelet transform can be defined for different class of functions. The intention in this transformation is to address some of the shortcomings of the STFT. Instead of fixing the

time and the frequency resolutions, one can let both resolutions vary in time-frequency plane in order to obtain a multi-resolution analysis.

In terms of the filter bank terminology, the analysis filter bank consists of band-pass filters with constant relative bandwidth (so-called .constant-Q. analysis). The way that the time-frequency plane is resolved in this approach is as shown in Figure1-4(d). In this case, the frequency responses of the analysis filters in the filter bank are regularly spaced in a logarithmic scale.

With this approach, the analysis is localized, and the time information is also reserved. The time resolution becomes quite good at high frequencies, while the frequency resolution is quite good at low frequencies. In 1995, Donoho and Johnstone proposed a new algorithm using wavelet thresholding for de-noising signals corrupted by Gaussian white noise [Donoho1995]. After that wavelet de-noising has become an extremely popular research topic and a new option for the development of speech enhancement methods as well.

The general procedure of wavelet de-noising can be illustrated as below, in Figure 1-5

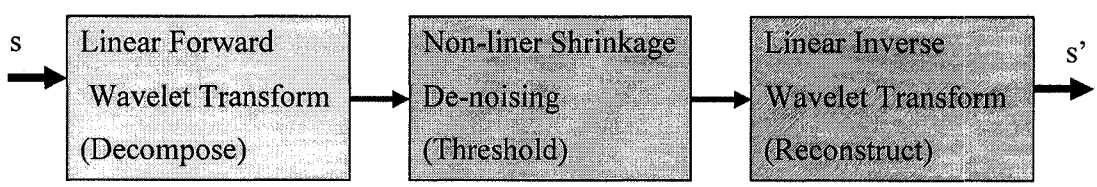


Figure 1-5 Principle of wavelet de-noising

1.2 Motivation

As discussed above, the wavelet transform has provided new opportunities to improve the perceptual effect of speech enhancement. In deed, a lot of research has been performed on high quality speech enhancement by wavelet de-noising over the past decade. Beyond the standard soft thresholding proposed by Donoho and Johnstone, new methods have

been developed to achieve outputs friendlier to human subjective perception. In literature, methods based on perceptual models which map the filter banks in the human inner ear, and other new adaptive technologies, are also presented. Different concepts and algorithms have been tried separately. However, not much work has been undertaken to analyze and compare them. Do they really improve the de-noising result? Which method is the most effective and efficient one? Which technology is worth further research? What other opportunities should be explored in the future work? With a strong interest in wavelet speech enhancement application and ambition to answer these questions, the perceptual adaptive wavelet de-noising has been selected as the topic of this thesis.

1.3 Objective

The objectives of this thesis work include summarizing and comparing the existing wavelet de-noising methods, so that an optimized perceptual adaptive wavelet de-noising algorithm which is effective in both stationary and non-stationary noise environments can be proposed.

CHAPTER II

2. SURVEY OF WAVELET DE-NOISING

Nowadays many computer software packages contain fast and efficient algorithms to perform wavelet transforms. Due to such easy accessibility, wavelets have quickly gained popularity among scientists and engineers, both in theoretical research and in applications. Wavelets have been widely applied in such research areas as image processing, computer vision, network management, data mining, and of course, speech processing. In 1995, Donoho and Johnstone proposed the famous method de-noising by wavelet soft thresholding [Donoho1995a] [Donoho1995b] [Johnstone1997]. This method has been used as a standard wavelet de-noising procedure for many years. Based on this fundamental procedure, various wavelet de-noising algorithms have been developed. The SureShrink is a relative mature one, followed by different effort trying to achieve a perceptual adaptive wavelet speech enhancement [Donoho1995c]. A brief survey of the wavelet de-noising technology is presented in this chapter to help understand the remainder of this thesis.

2.1 Wavelet Theory

2.1.1 Introduction to Wavelet

Wavelet theory is the mathematics associated with building a model for a signal, system, or process with a set of little waves or "wavelets". They must be oscillatory (waves) and have amplitudes, which quickly decay to zero in both the positive and negative directions (little). A wavelet is a waveform of effectively limited duration that has an average value of zero. Unlike sine waves (the basis of Fourier analysis), which are smooth and predictable, wavelets tend to be irregular and asymmetric as shown in Figure 2-1[Web 2]. The advantage of wavelet is that signals with sharp changes are better analyzed with an irregular wavelet than with a smooth sinusoid. Also, local features can be described better with wavelets that have local extent.

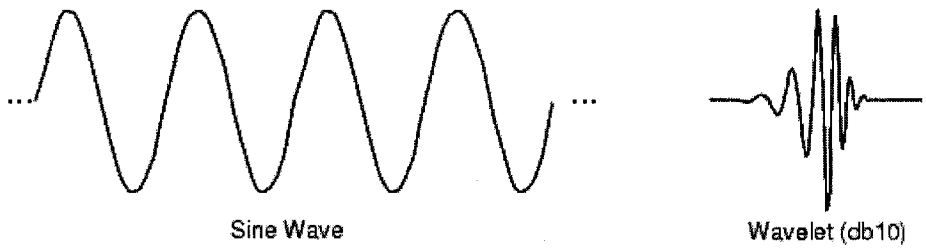


Figure 2-1 Shape of sinusoidal and Daubechies wavelet

[Web 2]

2.1.2 Continuous Wavelet Transform

The results of the Fourier transform are the Fourier coefficients $F(\omega)$, which when multiplied by a sinusoid of frequency ω yield the constituent sinusoidal components of the original signal. Similarly, the continuous wavelet transform (CWT) is defined as the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet function ψ. The results of the CWT are many wavelet coefficients C_ψ, which are a function of scale and position, as illustrated in Figure2-2[Web 3].

$$C_\psi = \int_{-\infty}^{+\infty} f(t)\psi(scale, position)dt \qquad \textbf{(2-1)}$$

Multiplying each coefficient by the appropriately scaled and shifted wavelet yields the constituent wavelets of the original signal.

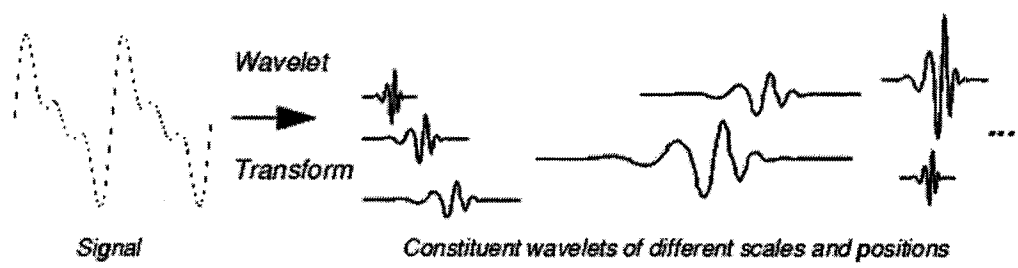


Figure 2-2 Continuous wavelet transform

[Web 3]

Scaling

Scaling a wavelet simply means stretching or compressing it. The scale factor a is related (inversely) to the frequency.

Low scale a => Compressed wavelet => Rapidly changing details => High frequency ω.

High scale a => Stretched wavelet => Slowly changing, features => Low frequency ω.

Shifting

Shifting a wavelet simply means delaying (or hastening) its onset, as shown in Figure 2-3. Mathematically, delaying a function $\psi(t)$ by k is represented by $\psi(t-k)$.

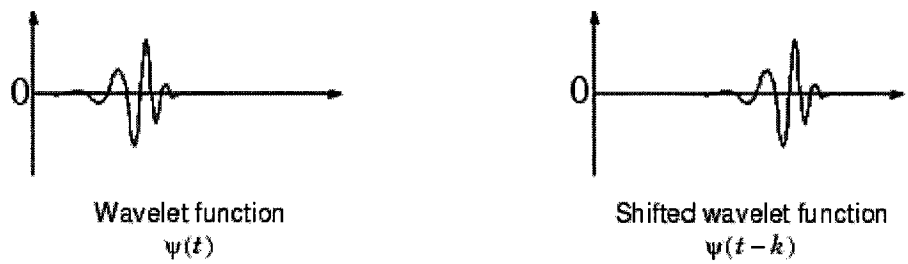


Figure 2-3 Wavelet shifting

[Web 3]

The continuous wavelet transform is the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet. This process produces wavelet coefficients that are a function of scale and position.

2.1.3 Discrete Wavelet Transform

It turns out that, if we choose scales and positions based on powers of two, called dyadic scales and positions, then, our analysis will be much more efficient and be as accurate as the CWT. This kind of analysis is called the discrete wavelet transform (DWT). Taking into account the non-stationary characteristic of real signals, the DWT provides high time resolution and low frequency resolution for high frequencies.

The DWT of a signal $x[n]$ is calculated by passing it through a series of filters. First the samples are passed through a low pass filter with impulse response $g[n]$. The signal is also decomposed simultaneously using a high-pass filter $h[n]$.The outputs giving the detailed coefficients (from the high-pass filter) and approximation coefficients (from the low-pass). Since half the frequencies of the signal have now been removed, half the samples can be discarding according to Nyquist's rule. The filter outputs are then down-sampled (or sub-sampled) by 2:

$$y_{low}[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot g[2n-k] \tag{2-2}$$

$$y_{high}[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot h[2n-k] \tag{2-3}$$

This decomposition is repeated to further increase the frequency resolution and the approximation coefficients decomposed with high and low pass filters and then down-sampled. This is represented as a binary tree with nodes representing a sub-space with different time-frequency localizations. The tree is known as a filter bank. Figure 2-4 [Web 4] represents a three level filter bank.

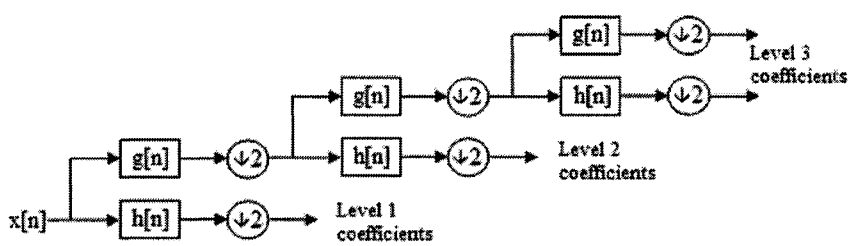


Figure 2-4 A three level filter bank

[Web 4]

At each level in the above diagram the signal is decomposed into low and high frequencies. Due to the decomposition process the input signal must be a multiple of 2^n where n is the number of levels.

2.1.3 Wavelet Packet Transform

The wavelet packet method, a wavelet transform where the signal is passed though more filters than the DWT, is a generalization of decomposition process that offers a richer range of capabilities for signal analysis. In the DWT, each level is calculated by passing the previous approximation coefficients though a high and low pass filters. However in the Wavelet packet decomposition (WPD), both the detail and approximation coefficients are decomposed, as represented in Figure 2-5 [Web 5]. The wavelet packet analysis offers much better frequency resolution than the simple wavelet analysis. In this way, subbands with smaller bandwidth across the whole spectrum can be achieved after the decomposition. The rough frequency analysis at the high frequency part becomes much more delicate.

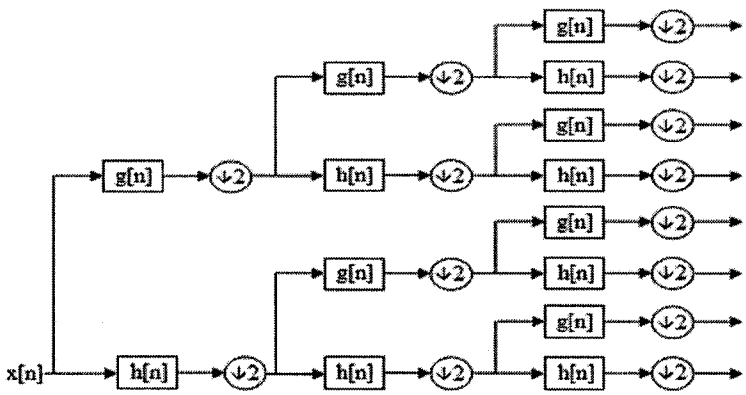


Figure 2-5 Multilevel wavelet packet decomposition

[Web5]

2.2 Standard De-noising Procedure Using Universal Threshold

As discussed above, some of the resulting wavelet coefficients correspond to details in the data set (high frequency sub-bands). According to Donoho and Johnstone's research, if the details are small, they might be omitted without substantially affecting the main features of the data set. The idea of thresholding is to set all high frequency sub-band

coefficients that are less than a particular threshold to zero. These coefficients are used in an inverse wavelet transformation to reconstruct the data set.

The general de-noising procedure involves three steps. The basic version of the procedure follows the steps described below.

1. Decompose - Choose a wavelet, choose a level N_L. Compute the wavelet decomposition of the signal s at level N_L.
2. Threshold detail coefficients - For each level from 1 to N_L, select a threshold and apply soft or hard thresholding to the detail coefficients.
3. Reconstruct - Compute wavelet reconstruction using the original approximation coefficients of level N_L and the modified detail coefficients of levels from 1 to N_L.

2.2.1 Soft Thresholding

Basically, wavelet de-noising methods involve either hard or soft thresholding. In the hard thresholding method, the coefficient is set to a specific value when its magnitude exceeds the threshold. On the other hand, soft thresholding shrinks or scales the coefficient that exceeds the threshold value. Hard thresholding is the simplest method. Soft thresholding has nice mathematical properties and the corresponding theoretical results are available.

Let Y denote the input, T denote the threshold.
The Hard Thresholding function is presented as

$$THR_H(Y,T) = \begin{cases} Y, |Y| \geq T \\ 0, |Y| < T \end{cases} \tag{2-4}$$

The Soft Thresholding function is presented as

$$THR_S(Y,T) = \begin{cases} \text{sgn}(Y)(|Y| - T), & |Y| \geq T \\ 0 & , |Y| < T \end{cases} \tag{2-5}$$

Their difference can be seen more clearly in Figure 2-6. Apparently, the hard procedure creates discontinuities at $x = \pm T$, while the soft procedure does not.

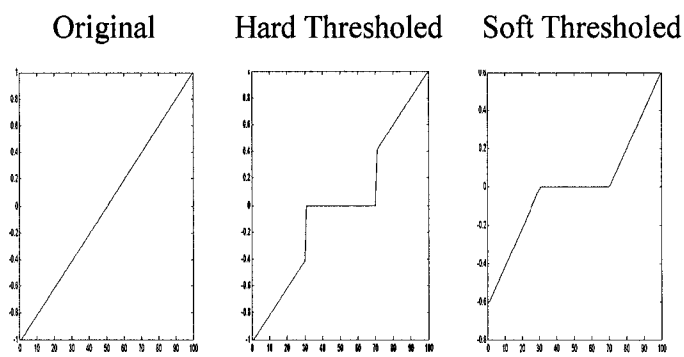


Figure 2-6 Hard thresholding and soft thresholding

[Web 7]

2.2.2 Noise Estimation

The adequate value for threshold can be determined in many ways. A universal threshold for discrete wavelet transform (DWT) has been introduced by Donoho [Donoho1995a] as:

$$T = \hat{\sigma}\sqrt{2\log(N)} \tag{2-6}$$

and for wavelet packet transform (WPT) case, the threshold value is determined as:

$$T = \hat{\sigma}\sqrt{2\log(N\log_2(N))} \tag{2-7}$$

where N is the length of noisy signal and $\hat{\sigma}$ is the standard deviation of the noise. $\hat{\sigma}$ is estimated by:

$$\hat{\sigma} = MAD/0.6745 = Median(|c|)/0.6745 \tag{2-8}$$

where c is the coefficient sequence from wavelet transform.

Donoho provide strong theoretical support to this classic wavelet de-noising algorithm. He proved that the theoretical advantages are really due to the wavelet basis [Donoho1995a] [Donoho1995b]. That is the foundation of various successful applications of wavelet de-noising.

This algorithm is simple and effective for removing Gaussian noise. However, the universal threshold is not effective for de-noising of colored and non-stationary noises in noisy speech signals. The universal method assumes that noise spectrum is white whereas normally it is colored in real life. So, the universal wavelet shrinkage does not result in good speech quality and cannot remove colored noises effectively. Another shortcoming of it is that the shrinkage of the unvoiced segments of speech which contain many noise-like speech components, leading to degraded speech quality. Also, the use of a universal threshold for all wavelet packet bands often results in poor correlation between the mean squared error criterion and the subjective quality in the presence of correlated noise and time-frequency discontinuities.

2.3 SureShrink

Donoho developed another wavelet shrinkage scheme (SureShrink) based on Stein's Unbiased Estimate of Risk (Sure) [Donoho1995c]. SureShrink is a procedure which suppresses noise by thresholding the empirical wavelet coefficients. The thresholding is adaptive: a threshold level is assigned to each dyadic resolution level by the principle of minimizing the Sure for threshold estimates.

SureShrink is smoothness-adaptive: if the unknown function contains jumps, the reconstruction (essentially) does also; if the unknown function has a smooth piece, the reconstruction is (essentially) as smooth as the mother wavelet will allow. The procedure is in a sense optimally smoothness-adaptive: it is near-minimax simultaneously over a whole interval of the Besov scale; the size of this interval depends on the choice of mother wavelet. Examples of SureShrink are given: the advantages of the method are

particularly evident when the underlying function has jump discontinuities on a smooth background.

In 1981, Stein [Stein1981] introduced a method for estimating the loss $\|\hat{w}-w\|^2$ in an unbiased fashion, where $\hat{w}$ is an estimator of w. If $\hat{w}$ can be written as $y+g(y)$, where $g=(g_i)_{i=1}^n$ is weakly differentiable, then

$$E\|\hat{w}-w\|^2 = E\left\{\|g(y)\|^2 + \sigma^2 n + 2\sigma^2 \nabla \cdot g(y)\right\} \tag{2-9}$$

where $\nabla \cdot g(y) = \sum_i \frac{\partial}{\partial y_i} g_i$, σ^2 is noise variance.

Applying Sure to wavelet shrinkage, we have

$$\begin{aligned} E\|\hat{w}-w\|^2 &= E\left\{\|g(y)\|^2 + \sigma^2 n + 2\sigma^2 \nabla \cdot g(y)\right\} \\ &= E\left\{\sum_{i=1}^{n} (|y_i| \wedge \lambda)^2 + \sigma^2 n - 2\sigma^2 \cdot \#\{i : |y_i| \le \lambda\}\right\} \\ &= E\{SURE(\lambda, y)\} \end{aligned} \tag{2-10}$$

The best λ is the one that can minimize $SURE(\lambda, y)$ and $E\|\hat{w}-w\|^2$.

$$\lambda_{best} = \arg\min_{\lambda \ge 0} SURE(\lambda, y) \tag{2-11}$$

Therefore the SureShrink Algorithm can be summarized as

1. Wavelet Decomposition

   $$y_{i,j} = W\{x_i\}, \quad i = 1,\ldots,n;\ j = 0,1\ldots,J \tag{2-12}$$

2. Sure Shrinkage: For each level evaluate λ_{best} based on SURE. Then apply wavelet shrinkage

   $$\hat{w}_{i,j} = \eta_S\{y_{i,j}, \lambda_{best}, j\} \tag{2-13}$$

3. Inverse transform

   $$\hat{x} = W^T\{\hat{w}\} \tag{2-14}$$

From the above discussion, it can be seen that SureShrink is adaptive to signal because λ_{best} is directly evaluated from the observed data. λ_{best} can be level dependent which means that different scale of wavelet coefficients may have different λ_{best}. Compared with the standard wavelet shrinkage, the SURE threshold selection rules are more conservative, that is proved later by the simulation results in this thesis.

2.4 Latest Development

As discussed above, there are some problems with the universal wavelet thresholding method when it is applied to the noisy speech corrupted by real-life noise. Although SureShrink has the contribution to make the shrinkage adaptive to signal, it tends to be too conservative. A lot of background noise is left while the distortion of the speech part is reduced. Therefore, in recent ten years, authors have been working on the development of adaptive de-noising schemes with better trade-off between noise suppression and speech distortion control. A more precise definition of the problem and analysis of the possibility of relevant improvement will be provided in this section.

2.4.1 A More Precise Definition

As assumed in chapter I, the observed data consists of the clean signal $s(t)$ and additive noise $w(t)$

$$y(t) = s(t) + w(t) \tag{2-15}$$

Then, for the standard universal thresholding there are three steps as shown below [Taswell2000]

$$Y = W(y) \longrightarrow Z = D(Y,\lambda) \longrightarrow S = W^{-1}(Z)$$

where $W(\cdot)$ and $W^{-1}(\cdot)$ denote the forward and inverse wavelet transform operators; $D(\cdot,\lambda)$ denote the de-noising operator with soft threshold λ. And the nonlinear thresholding can be illustrated as

$$D(U,\lambda) \equiv sgn(U)\,max(0,\left||U|-\lambda\right|) \tag{2-16}$$

The operator $D(U,\lambda)$ nulls all values of U for which $|U| = \lambda$ and shrinks toward the origin by an amount λ all values of U for which $|U| > \lambda$. It is the latter aspect that has led to $D(U,\lambda)$ being called the shrinkage operator in addition to the soft thresholding operator.

In the case of adaptive thresholding, the threshold λ does not only depend on sample size n but also on U. Then de-nosing procedure is extended to four steps as shown below

$$Y = W(y) \longrightarrow \lambda = d(Y) \longrightarrow Z = D(Y,\lambda) \longrightarrow S = W^{-1}(Z)$$

where $d(Y)$ is the adaptive thresholding operator.

Apparently, we can generate many different kinds of wavelet shrinkage de-noising procedures by combining different choices for $W(\cdot)$, $d(\cdot)$ and $D(\cdot,\cdot)$.

2.4.2 Research Direction

The definition above has made the further work clearer. In other words, to improve the wavelet de-noising, there are three directions to go. The first one is to build a more effective decomposition structure to model the human auditory filter banks. The decomposition should have a proper subband width to make sure the time-frequency analysis delicate enough. Besides, the ideal state is that it could accurately map the critical bands, so that the auditory masking rules could be used to avoid the over suppression and consequently improve the perceptual performance of the speech enhancement.

The accuracy of noise estimate has always been crucial for the right thresholding. Therefore, one task is to develop a more accurate noise estimate algorithm under real-life noise situation. The adaptive noise estimation algorithm is a noise estimation technique that is updated adaptively and continuously from the nearest previous speech frames without explicit speech pause detection. An effective adaptive noise estimation algorithm should have the ability to track the change of the SNR rapidly.

Furthermore, the thresholding algorithm is another handle for us to adjust the noise suppression. It could also be adaptive in order to yield a smoother output. In addition, musical noise is a typical problem with blind source separation using a time-frequency mask. Musical noise has been widely considered in the field of single channel speech enhancement with spectral subtraction. Thus, it is also necessary to add a musical noise control module after the general thresholding.

In literature, the authors have actually followed these three directions of investigation to pursue the further improvement in this area. These new methods will be discussed one by one in the following chapter III and IV.

CHAPTER III

3. PERCEPTUAL WAVELET THRESHOLDING

3.1 Critical Bands

Auditory perception is based on a critical band analysis in the inner ear. A critical band is the bandwidth around a center frequency beyond which subjective responses of the hearing system abruptly change. The notion was first introduced by Fletcher (1940) and has played an important role when constructing the perceptual wavelets [Virag1999] [Carnero1999]. Later in this chapter, the relationship between critical bands and the simultaneous masking property of the human auditory system will be discussed.

Generally, the human auditory frequency range is divided into 25 critical bands which spread from 20Hz to 20 kHz [Virag1999] [Carnero1999], as shown in Table 3-1. These critical bands can be thought as a bunch of filters with non-uniform temporal and spectral response, working as a central analysis mechanism in the inner ear, illustrated in Figure 3-1. The critical bandwidth (CBW) of these filters is of approximately 100 Hz below 500 Hz. Above 500Hz, the bandwidth corresponds to about 20% of the center frequency value.

According to Fletcher's experiment, in order to measure the bandwidth of a critical band centered at any frequency, a tonal signal inaudible is made by a narrowband noise centered at that frequency. If the bandwidth of the noise increases, the level of the inaudible sinusoid increases. When the bandwidth of the noise exceeds a certain value, i.e., the critical bandwidth, the level of the sinusoid input remains almost constant. Figure 3-2 [Moore1996] shows how the threshold changes as a function of the noise bandwidth. When the noise bandwidth becomes wider than the critical bandwidth, here which is around 300Hz for 2 kHz signal, the threshold level tends to be flat.

Table 3-1 Critical bands in the range of 0~22 kHz

[Virag1999] [Carnero1999]

Critical Band Number(Bark)	Frequency(Hz)			
	Lower Cutoff Frequency	upper Cutoff Frequency	Critical Band Width	Center Frequency
0	0	100	100	50
1	100	200	100	150
2	200	300	100	250
3	300	400	100	350
4	400	510	110	450
5	510	630	120	570
6	630	770	140	700
7	770	920	150	840
8	920	1080	160	1000
9	1080	1270	190	1170
10	1270	1480	210	1370
11	1480	1720	240	1600
12	1720	2000	280	1850
13	2000	2320	320	2150
14	2320	2700	380	2500
15	2700	3150	450	2900
16	3150	3700	550	3400
17	3700	4400	700	4000
18	4400	5300	900	4800
19	5300	6400	1100	5800
20	6400	7700	1300	7000
21	7700	9500	1800	8600
22	9500	12000	2500	10750
23	12000	15500	3500	13750
24	15500	22050	6550	18775

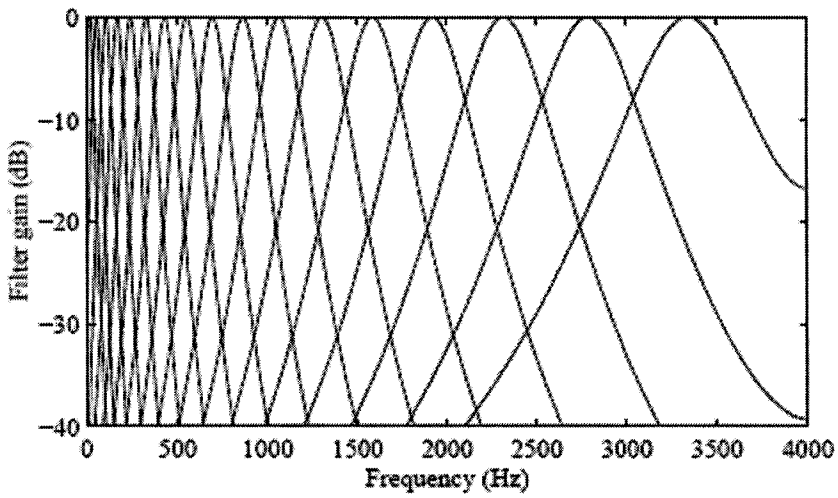


Figure 3-1 Example of non-uniform filter banks in the inner ear

[Moore1996]

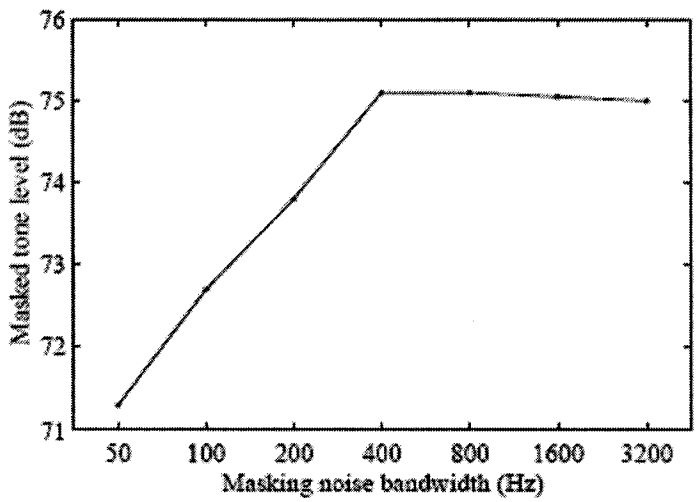


Figure 3-2 Threshold of a just audible 2 kHz test tone

[Moore1996]

The critical band scale, or Bark scale, is indispensable for the study of auditory masking since it represents the natural scale of the inner ear, and all models of masking require some kind of critical band analysis. The distance from one critical band center to the center of the next band is 1 Bark. Thus, the human auditory frequency range covers approximately 25 Barks. The center frequency location of these subbands is known as the

critical band rate and approximately follows the expression [Zwicker1999]:

$$z = 13\arctan(7.6\times10^{-4} f) + 3.5\arctan(1.33\times10^{-4} f)^2 \quad \textbf{(3-1)}$$

where z denotes the critical ban number (in Bark), f is the frequency (in Hz) which can be calculated by the following formula

$$f = 650\times\sinh(z/7) \quad \textbf{(3-2)}$$

The corresponding bandwidth is also a function of the frequency, shown as [Zwicker1999]

$$BW(f) = 25 + 75\times\left[1 + 1.4\times(f/1000)^2\right]^{0.69} \quad \textbf{(3-3)}$$

3.2 Absolute Threshold of Hearing

Not all the sounds can be heard by human ear. Whether the human ear responses to a sound depends on its frequency arrange and intensity. Normally, the frequency response scope of a young people is 20 Hz ~ 20 kHz. When the sound pressure is above 0 dB, it can be heard by human auditory system. A sound with magnitude over 120 dB can make our ear uncomfortable.

The absolute threshold of hearing (AHT), or threshold in quiet, is the minimum average sound pressure level (SPL) for the human ear to detect any stimulus. This threshold is frequency dependent and can be closely modeled by a non-linear function of frequency, as shown in Figure 3-3[Zwicker1999][Web 6]. The following formula expresses the threshold in quiet at frequency f (in Hz) [Terhardt1982]

$$Tq = 3.64(f/1000)^{-0.8} - 6.5\exp(-0.6(f/1000-3.3)^2) + 10^{-3}(f/1000)^4 \quad dB \quad \textbf{(3-4)}$$

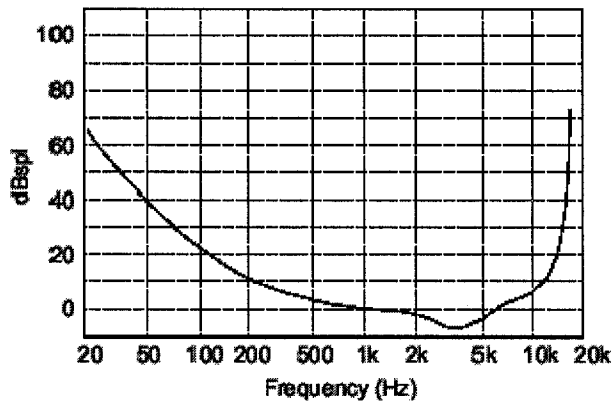


Figure 3-3 Absolute threshold of hearing

[Zwicker 1999][Web 6]

3.3 Auditory Masking

Auditory masking is a phenomenon connected to the hearing perception of neighbouring signal components. It indicates that a weaker audio signal becomes inaudible (masked) by a louder signal occurring simultaneously or close in time. This explains why people need to raise their voice to make them understood in a very noisy environment. In speech enhancement, the masker is the original input signal, and the maskee is background noise. The masking phenomena can be exploited to reduce the mis-suppression in a situation with high signal-to-noise radio.

Two main categories of masking, depending on the time and frequency location of the masker and maskee, may be considered. When both signals occur at the same time, masking is considered *simultaneous* and is modeled in the frequency domain. On the other hand, if B either precedes or succeeds A, masking is termed *temporal or non-simultaneous* [Carnero1999].

3.3.1 Simultaneous Masking

Simultaneous masking indicates the masking phenomenon among the different frequency

components of sounds occurring at the same time.

As shown in literature, the nature of the masker being noise-like or tonal has an impact on the masking curve. For instance, the maximum of the masking curve due to a single tone is sharper (peaky) [Zwicker1999]. Additionally the distance between the masker level and the masking threshold is greater for tonal signals.

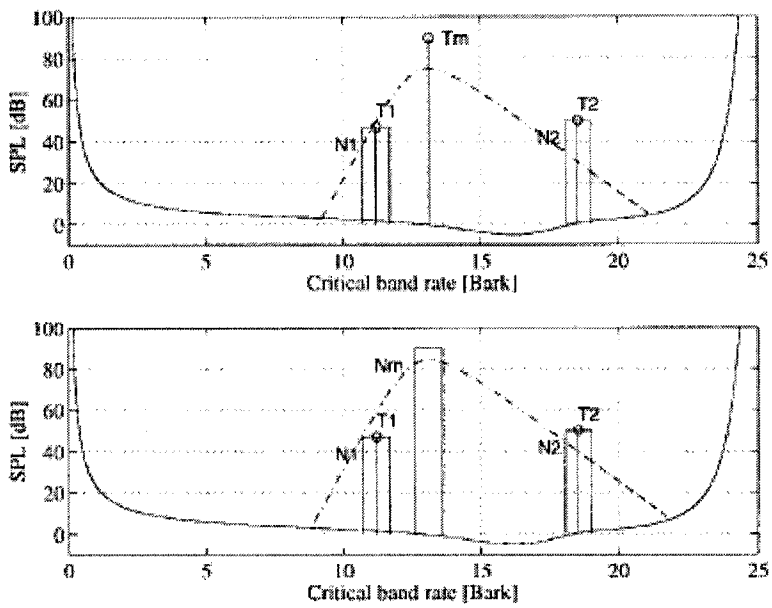


Figure 3-4 Simultaneous masking

Top: Tone masker. Bottom: Noise-like masker (one Bark wide)

[Carnero1999, p6]

This masking threshold has been modeled by a spreading function centered on the masker, which illustrating the shape of the energy distribution (excitation pattern) along the basilar membrane. Based on the psychoacoustic findings, the spreading function is a function of the frequency and the level of the masker. In almost all masking models a triangular shape (on a critical band scale) is assumed for the spreading function, as shown in Figure 3-4[Carnero1999]. The patterns lying completely below the masking threshold are totally masked, whereas those lying only partially below it are partially masked. Additionally, the masking threshold offset of tone-like signal and noise-like signal is

different, as can be seen from the figure.

Different slopes of the function on both sides have been reported in the literature. In this thesis Johnston's Masking Model [J.D.Johnston1988] was adopted. In order to calculate the masking threshold, the power in each critical band is found; then the Bark power spectrum will be spread over all critical bands through convolving the Bark spectrum with the following spreading function

$$SF(z') = 15.81 + 7.5(z' + 0.474) - 17.5(1 + (z' + 0.474)^2)^{0.5} \qquad \textbf{(3-5)}$$

where z' is the separation between critical bands. As can be seen, this spreading function is independent of the level and frequency of the masker.

For the noise-masking-tone, the masking threshold is 5.5 dB below the spread spectrum. For the tone-masking-noise the masking threshold is $(14.5 + i')$ dB below the spread spectrum, where i' is the bark frequency of the masking signal [J.D.Johnston1988]. In order to determine the nature of the signal as being tone-like or noise-like, the spectral flatness measure which is defined as follows is used

$$SFM = 10\log_{10}\left(\frac{Gm}{Am}\right) \quad dB \qquad \textbf{(3-6)}$$

where Gm and Am are the geometric mean and arithmetic mean respectively. The tonality factor is then defined as

$$\alpha' = min\left(\frac{SFM}{SFM_{max}}, 1\right) \qquad \textbf{(3-7)}$$

where SFM_{max} corresponds to a signal which is assumed to be a pure tone and is set to -60 dB; a zero value for SFM represents noise. To find the masking threshold the

following offset is subtracted from the spread spectrum (in dB)

$$O(i') = \alpha'(14.5 + i') + 5.5(1 - \alpha') \quad \textbf{(3-8)}$$

Finally the masking threshold is compared with the threshold of hearing to make sure that it is not below the threshold of hearing.

3.3.2 Temporal Masking

Besides the frequency domain masking phenomena, two main time domain masking phenomena have been observed in human audition: pre-masking, which is also called backward masking, and post-masking, which is also termed forward masking. They are both depicted in Figure 3-5[Carnero1999]. Maskees lying below the two decaying curves are inaudible. Post-masking has a more important effect than pre-masking since it has a longer duration. Pre-masking appears approximately 20 ms before the masker, whereas post-masking lasts for about 100 to 200 ms. Temporal masking is maximum for signals close in frequency and within the same critical band. The full effect of temporal masking is closely related to the duration of the masker. Maximum masking is produced by maskers lasting about 200ms. Below that value, the masking threshold shows faster decay slopes and, hence, a shorter duration. This clearly suggests that temporal masking is a highly nonlinear effect. According to several researches,

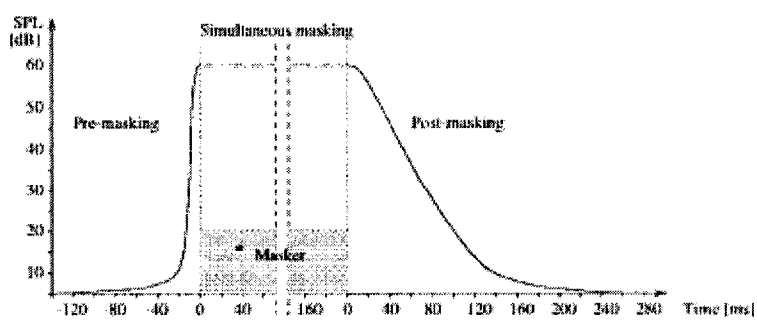


Figure 3-5 The temporal masking

[Carnero1999, p6]

These two kinds of masking are widely accepted as separate mechanisms for the purpose of modeling. However, they are closely interconnected. In this thesis, only simultaneous masking threshold is considered.

3.4 Critical Bands Analysis in the Time-Frequency Domain

3.4.1 Critical Band Modeling

Within the 0~8 kHz frequency scope, there are 21 critical bands as shown in Table 3-1. Compared with other conventional transform tools, the discrete wavelet packet transform (DWPT) provides a much more accurate mapping of the critical bands, as observed from Figure 3-6 [Black1995, p1]. By using a six level DWPT, the minimum frequency bandwidth of 125Hz can be achieved, that is close to the 100 Hz bandwidth of low frequency critical bands.

In literature, different models have been proposed [Shao2005] [Pinte´r1996] [Black1995]. For use here, the Daubechies wavelet was selected as the mother wavelet, since it has the best preservation frequency selectivity as the number of stages of the DWPT increases. It has been proven by simulation that db8 and db10 are the best choice to describe speech signals. However, there is a limit of decomposition level at a particular frame length and particular wavelet. For example, if the frame length is 256, the maximal decomposition level achieved with db8 is four. Thus, only db1 or db2 is available for six level mapping in this case. Figure 3-7 shows the DWPT mapped critical bands.

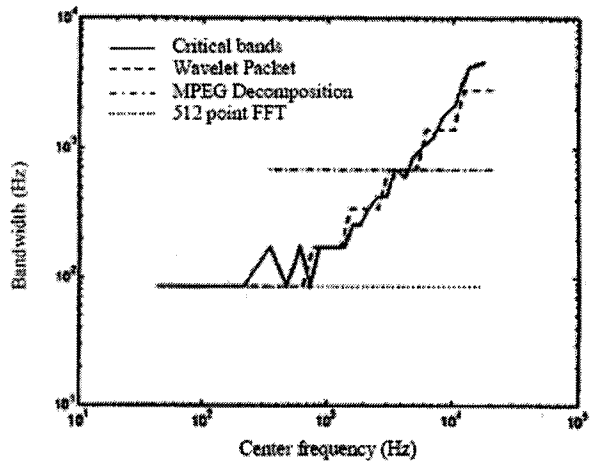


Figure 3-6 Approximation to critical bands

[Black1995, p1]

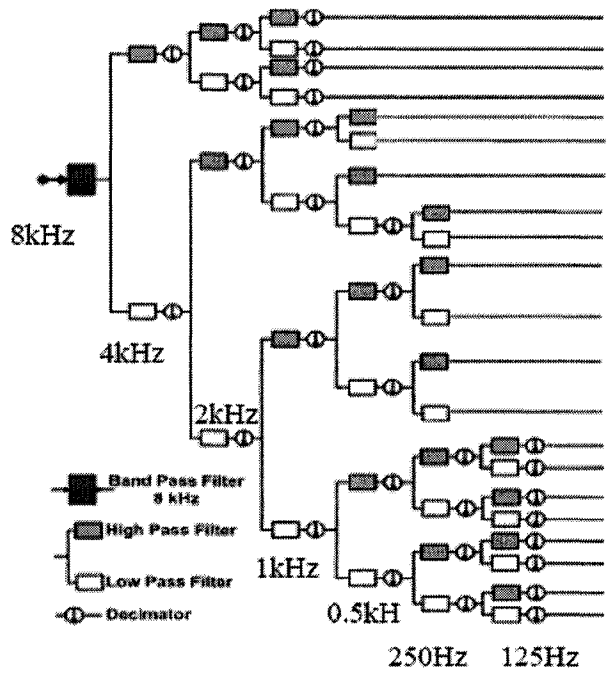


Figure 3-7 DWPT mapped 21 critical bands for 8 kHz speech signal

The resulting critical bands rate and bandwidth are plotted in Figure 3-8 [Carnero1999], along with the corresponding model of real critical bands. As can be seen, the DWPT

mapping results are very close to the factors of real critical bands and then provides a delicate time-frequency analysis.

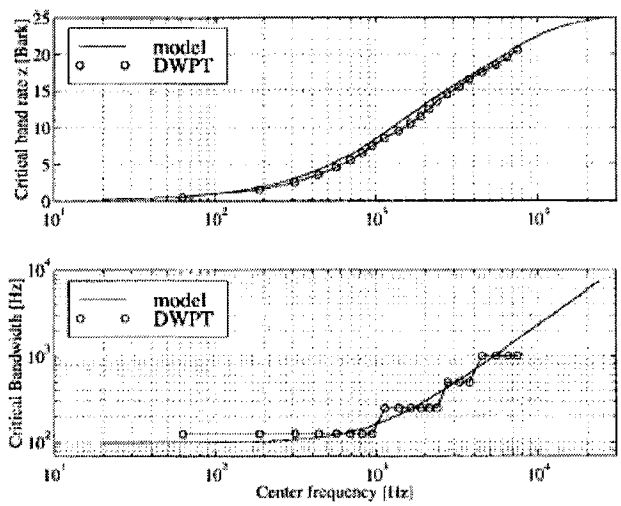


Figure 3-8 DWPT modeled critical band

Top: critical band rate Bottom: critical bandwidth

[Carnero1999, p5]

3.4.2 Noise Masking Threshold

In time-frequency domain, steps for calculating simultaneous masking threshold can be summarized as:

1) Modeling critical bands using 6-level wavelet packet decomposition and compute the energy of each subband (Bark power spectrum).

$$Px_j(n) = \sum_{i=1}^{N_j} c_{i,j}(n)^2, j = 1,2...21 \qquad \textbf{(3-9)}$$

where n is the frame index, j is the subband (Bark) index, $c_{i,j}(n)$ denotes the ith coefficient in subband j.

2) Convolving the Bark spectrum with the following spreading function

$$PxS_j(n) = \sum_{j'=1}^{21} SF\ (j,j')Px_{j'}(n) \qquad \textbf{(3-10)}$$

$$SF(j,j') = 15.81 + 7.5((j-j')+0.474) - 17.5(1+((j-j')+0.474)^2)^{0.5} \qquad \textbf{(3-11)}$$

3) Subtraction of a relative threshold offset depending on the noise-like or tone-like nature of the masker. For the noise-masking-tone, the masking threshold is 5.5 dB below the spread spectrum. For the tone-masking-noise the masking threshold is (14.5+ j) dB below the spread spectrum, where j is the critical band index. In order to determine whether the nature of the signal is tone-like or noise-like, the spectral flatness measure which is defined as follows is used

$$SFM = 10\log_{10}\left(\frac{(\prod_{i=1}^{N_j} c_{i,j}^2(n))^{1/N_j}}{\bar{c}_{i,j}(n)\Big|_{i=1}^{N_j}}\right) \quad dB \qquad \textbf{(3-12)}$$

The tonality factor is given by equation (3-4) as $\alpha' = min(\frac{SFM}{SFM_{max}},1)$

where $SFM_{\max}$ corresponds to a signal which is assumed to be a pure tone and set to -60 dB; a zero value for SFM represents noise. To find the masking threshold the following offset is subtracted from the spread spectrum (in dB)

$$O_j = \alpha'(14.5+j)+5.5(1-\alpha') \qquad \textbf{(3-13)}$$

Then, the simultaneous masking threshold is obtained from the following formula

$$NMT_j = PxS_j - 10^{O_j/10} \tag{3-14}$$

4) Finally, NMT_j is compared in each critical band with the maximal threshold in quiet, and the maximum of each value retained, giving $FNMT_j$. The absolute threshold of hearing (threshold in quiet) is computed by

$$Tq\max = \max(Tq(f')) \quad dB \tag{3-15}$$

where f' denotes the frequency with the jth subband (approximate Bark), step by one.

3.5 Perceptual Wavelet Subtraction

To achieve a perceptual speech enhancement, authors have used auditory masking properties to perform adaptive subtractive technique in Fourier domain [Virag1999] and wavelet domain [Carnero1999]. Figure 3-9 [Carnero1999] shows the system structure of the perceptual wavelet subtraction. In wavelet domain, after the perceptual transform, a rough subtraction is performed first using the following formula

$$\tilde{X}_j^2 = Y_j^2 - \overline{D_j^2} \tag{3-16}$$

where $\overline{D_j^2}$ is the averaged noise estimate calculated with a speech pause detector. Then noise masking threshold values are extracted from $\tilde{X}_j^2$. To reduce the effect of residual noise, a parametric-type approach using an over-subtraction factor α and a spectral flooring factor η was introduced into the algorithm. With the DWPT, this approach can be expressed as [Carnero1999]

$$\tilde{X}_j^2 = \begin{cases} Y_j^2 - \alpha_j \overline{D_j^2}, & if \quad Y_j^2 - \alpha_j \overline{D_j^2} > \eta_j \overline{D_j^2} \\ \eta_j \overline{D_j^2}, & otherwise \end{cases} \tag{3-17}$$

The parameters α_j and η_j are dependent on the time-frequency masking threshold in the jth subband. The adaptation rule follows sigmoid curves with $\alpha_{\min} = 1$, $\alpha_{\max} = 3$, $\eta_{\min} = 0$, $\eta_{\max} = 0.01$.

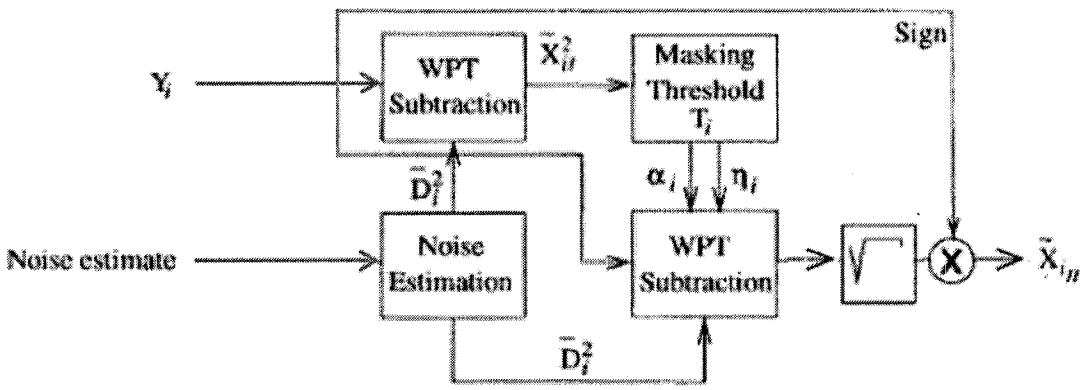


Figure 3-9 System block of perceptual wavelet subtraction

[Carnero1999, p9]

Since the noise masking threshold (NMT) has a smoother evolution than the SNR and the adaptation based on NMT is better correlated with perception than using the SNR, using it rather than the SNR to track the noise change takes some advantages. However, this method is involved in much more complex computation, which is very disadvantageous for the real-life application.

CHAPTER IV

4. ADAPTIVE THRESHOLD

Although standard wavelet soft thresholding has been proven to be effective for removing Gaussian white noise, it is obvious that the function of this simple method is restricted due to the time invariable algorithm and rough frequency-domain division. In real world, the background noise generally shows uneven power spectral intensity, which may also be time-variant. In other words, the speech signal may be polluted by a non-stationary noise with different local SNR at different time segments or frequency sub-bands. Thus, using a time-frequency invariable threshold results in over suppression at high SNR parts and deficient restraining at low SNR parts.

Among recent literature, two basic adaptive approaches have been studied to improve the accuracy of the standard thresholding. One of them is adaptive noise estimate, and the other is using a thresholding adjuster to track the changing of local SNR. In this chapter, these two methods will be introduced. In the mean time, methods for removing the musical noise are discussed in this chapter too. Finally, a new adaptive wavelet speech enhancement system is proposed.

4.1 Adaptive Noise Estimate

Instantaneous noise spectrum estimation is a critical component of single channel speech enhancement. Adaptive noise estimation algorithm is a noise estimation technique that is updated adaptively and continuously from the nearest previous speech frames without explicit speech pause detection.

4.1.1 Quantile-based Time-frequency Noise Estimate

4.1.1.1 Quantile

Quantiles are essentially points taken at regular intervals from the cumulative distribution function of a random variable. Dividing ordered data into q essentially equal-sized data subsets is the motivation for q-quantiles; the quantiles are the data values marking the boundaries between consecutive subsets. Put another way, the k th q-quantile is the value x such that the probability that a random variables will be less than x is at most k/q and the probability that a random variable will be less than or equal to x is at least k/q. There are $q-1$ quantiles, with k an integer satisfying $0 < k < q$.

If instead of using integers k and q, the p-quantile is based on a real number p with $0 < p < 1$ then this becomes: The p-quantile of the distribution of a random variable X can be defined as the value(s) x such that,

$$\begin{cases} P(X \le x) \ge p \\ P(X \ge x) \ge 1-p \end{cases} \tag{4-1}$$

4.1.1.2 Quantile-based Time-frequency Noise Estimate

The Quantile-based Time-frequency Noise Estimate (QBNE) method was originally proposed by V. Stahl and A. Fischer in 2000 [Stahl2000]. The principle idea derives from a minimum statistic algorithm by Martin in [Martin1993] [Martin1994]. The main idea of QBNE is to use the quantile value of a set of noisy signal energy as the noise estimate, so that to balance the current noise estimate using the data of previous frames.

Given a noisy speech $x(t)$, a buffer is used to store the power of the signal $Px_j(n)$ over a pre-defined duration L. The buffer contents are sorted and the q-th quantile is taken as the noise estimated power $Pn_j(n)$. Where, n denotes the frame index and j is the index of subbands. $c_{i,j}(n')$ is the wavelet coefficient. The process can be summarized as

follows [Fu2003] [Bai2003] [Lee2004] and shown in Figure 4-1[Fu2003].

1. Take the Wavelet Packet Transform and obtain $c_{i,j}(n')$

$$Px_j(n') = \sum_{i=1} \left|c_{i,j}(n')\right|^2, n' = n-L+1,\ldots,n-1,n \quad \textbf{(4-2)}$$

2. Sort $Px_j(n')$ in ascending order and re-index

$$Px_j(1) \le Px_j(2) \le Px_j(3) \le \ldots \le Px_j(L) \quad \textbf{(4-3)}$$

3. select the q-th quantile $Px_j(qL)$

4. Assign noise estimate

$$Pn_j(n) = Px_j(qL) \quad \textbf{(4-4)}$$

$$\hat{\sigma}_j(n) = \sqrt{Pn_j(n)} \quad \textbf{(4-5)}$$

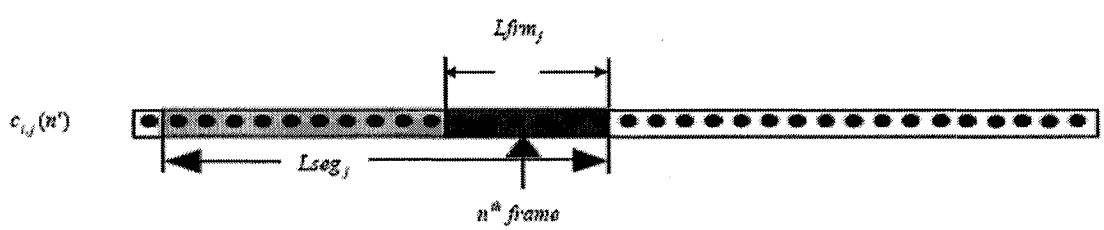


Figure 4-1 Quantile-based time-frequency noise estimate

[Fu2003, p4]

The parameter q normally takes a value of 0.5 in [Fu2003], which represents the median. However, some experimental results show the probability of having more than 20% duration being silence for various segment lengths [Ris2001]. For example, when the time segment length is 600ms, the probability of having more than 20% silence is greater than 85%. This indicates that the median assumption is too aggressive, leading to the increased likelihood of overestimating the noise level. Thus some authors chose a level associated with q=0.2[Lee2004] [Ris2001] [Stahl2000]. In addition, instead of using the quantile itself, the arithmetic mean of the lower 20% (i.e., q < 0.2) of the noisy speech power spectrum was used as the noise estimation. According to their experiments, this "low energy envelope" tracking method generally obtains better estimates compared to

other published quantile methods [Stahl2000]. Then the forth steps above can be modified to

$$\hat{\sigma}_j(n) = \sqrt{\sum_{n'=1}^{int(q \cdot L)} Px_j(n') \ / int(q \cdot L)} \tag{4-6}$$

For speech enhancement application, the threshold for j -th subband at the n-th frame, $\lambda_j(n)$ is estimated as

$$\lambda_j(n) = \hat{\sigma}_j(n) \cdot \sqrt{2 \log(N \ \cdot \log_2 N)} \tag{4-7}$$

QBNE is a statistics based adaptive time-frequency dependent noise estimation method. It is effective for tracking the slowly varying non-stationary noises and then improves the accuracy of noise deduction. However, the QBNE is inaccurate at frequencies where the speech components are consistently dominant [Lee2004]. Thus, it will over suppress the speech components when the local SNR is high, while it works well in the case with low SNR.

4.1.2 Exponential Smoothing and Sigmoid Tracking with PSNR

4.1.2.1 Exponential Smoothing

In statistics, smoothing method refers to calculating a weighted average among the latest data and the previous statistic, so that the estimate is closer to the real data. Exponential smoothing is a particular type of moving average technique, a smoothing method applied to time series data. The simplest form of exponential smoothing is given by the formulas

$$s_0 = x_0 \tag{4-8}$$

$$s_t = \alpha x_t + (1-\alpha) s_{t-1} \tag{4-9}$$

where α is the smoothing parameter, and $0<\alpha<1$, s_0 is the first statistic, s_t is the latest smoothed statistic, x_t is the real data.

Values of α close to unity have less of a smoothing effect and give greater weight to recent changes in the data, while values of α closer to zero have a greater smoothing effect, and are less responsive to recent changes. The term "Exponential" means , as time passes, the smoothed statistic s_t becomes the weighted average of a number of the past observations x_{t-n}, and the weights assigned to previous observations are in general proportional to the terms of the geometric progression $\{1, (1-\alpha), (1-\alpha)^2, (1-\alpha)^3, \ldots\}$. There is no formally correct procedure for choosing α. Sometimes the statistician's judgment is used to choose an appropriate factor. Alternatively, a statistical technique may be used to optimize the value of α.

Recently, some researchers have tried to use exponential smoothing to achieve adaptive noise estimate. The application is given as

$$Pn_j(n) = \alpha_j(n) \cdot Pn_j(n-1) + (1-\alpha_j(n)) \cdot Px_j(n) \qquad \textbf{(4-10)}$$

In [Lei2005], the sigmoid function is used to update the smooth parameter $\alpha_j(n)$ by a posteriori signal-to-noise ratio ($PSNR$). The definition of $PSNR$ is given as

$$PSNR_j(n) = Px_j(n) / \overline{P}n_j(n-1) \qquad \textbf{(4-11)}$$

where, $\overline{P}n_j(n-1)$ is the average of the noise estimates of the previous m frames adjacent to frame $n-1$ and given in form

$$\overline{P}n_j(n-1) = \frac{1}{m} \sum_{i=1}^{m} Pn_j(n-i) \qquad \textbf{(4-12)}$$

4.1.2.2 Sigmoid Function

The smoothing parameter of j -th subband at the n-th frame $\alpha_j(n)$ is then adaptively changed as a sigmoid function of the $PSNR$.

$$\alpha_j(n) = \frac{1}{1+e^{-a(PSNR_j(n)-T)}} \tag{4-13}$$

where, a and T are the slope and center-offset of the sigmoid function respectively. Sigmoid functions with different slopes are shown in Figure 4-2 [Lin 2003]. As can be seen, the slope becomes sharper when the value of a raises.

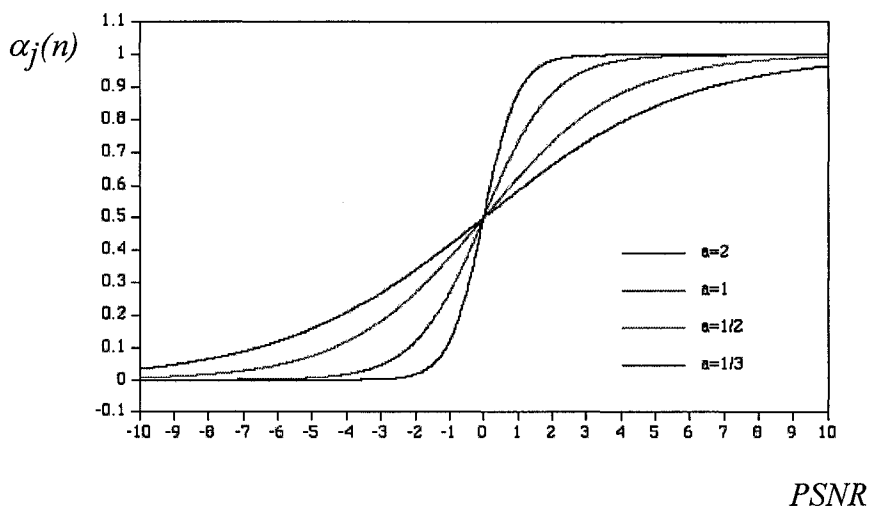


Figure 4-2 Sigmoid function

[Lin2003, p1]

Thus the smoothing parameter $\alpha_j(n)$ is closed to 0 when the speech is absent in frame n, that is, the estimate of noise power in frame n rapidly follows the power of the noisy signal in the absence of speech. On the other hand, if the speech is present, the new noisy signal power is much larger than the previous noise estimate. Then the value of the smoothing parameter $\alpha_j(n)$ increases rapidly with increasing $PSNR$. So the noise update is slower or almost stops because of the large value of smoothing parameter.

4.2 Adaptive Threshold Adjuster

In addition to the adaptive noise estimate, setting an adjuster to modulate the standard threshold value is another way to lighten the inaccurate noise suppression [Lei2005] [Lin2003] [Hu2004].

4.2.1 Posteriori SNR Time-Adaptive Threshold

As discussed in 4.1.2, the smoothing parameter $\alpha_j(n)$ rapidly follows the change of $PSNR$, and its value is among the range of 0 to 1. Therefore, it is an idea adaptive threshold adjuster. The time-adapted wavelet threshold is then defined as

$$\lambda_j(n) = \lambda_{0j}(n)(1-\alpha_j(n)) \tag{4-14}$$

where, the standard level-dependent threshold λ_{0j} is calculated by

$$\lambda_{0j}(n) = \hat{\sigma}_j(n)\cdot\sqrt{2\log(N_j)} \tag{4-15}$$

In this way, the threshold values are adapted to the SNR values across speech frames. For a speech-dominated frame, the increased SNR value results in lower threshold. The wavelet threshold of the corresponding frame should be adapted to smaller value so that the speech distortion can be reduced. On the contrary, the wavelet coefficients are almost determined by the noise component in a noise-dominated frame. More background noise can be removed by having larger wavelet threshold.

Finally, the noise components are suppressed by the soft thresholding to the decomposed noisy wavelet packet coefficients. The processing steps are shown in Figure 4-3.

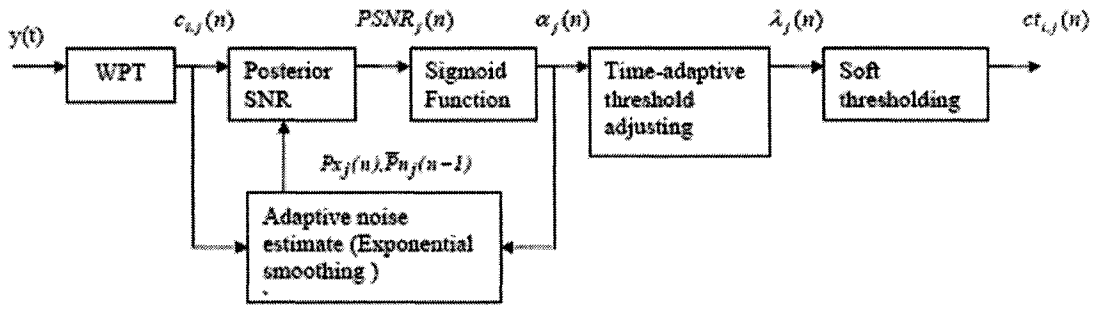


Figure 4-3 Posteriori SNR time-adaptive thresholding

4.2.2 Smoothed Hard Thresholding with Aggravated Threshold Value

As speech signal is a non-stationary signal, the signal-to-noise ratio of speech segments fluctuates across time. And this information could be used to adapt the threshold values. An aggravated threshold algorithm is proposed in [Ghanbari2005] to track a VAD based segmental signal-to-ratio ($SSNR$).

Here, the $SSNR$ is defined as

$$SSNR_j(n) = 10\log_{10}\frac{Px_j(n)}{Pn_j(n)} \tag{4-16}$$

where, $Px_j(n)$ denotes the energy of the noisy signal at j-th subband and frame n, and $Pn_j(n)$ is the noise estimate at j-th subband and frame n, which is defined as the signal energy of the latest silence segments.

The tracking function is:

$$T_j(n) = \begin{cases} \lambda_{0j}(n)(1+e^{-\frac{SSNR_j(n)}{\tau}}), SSNR_j(n) \geq 0 \\ 2\lambda_{0j}(n), SSNR_j(n) < 0 \end{cases} \tag{4-17}$$

where, $2<\tau<3$, and $\lambda_{0_j}(n)=(\frac{MAD_j(n)}{0.6745})\sqrt{2\log(N_j)}$ is the standard threshold.

As shown in Figure 4-4 [Ghanbari2005], the adaptive threshold value is an exponential function of the VAD (voice activity detector) based $SSNR$. When the $SSNR$ is smaller than 0, which means the estimated noise energy is stronger than that of the clean signal, the threshold value is doubled, and much more noise will be removed. On the other hand, as the $SSNR$ rises from 0, the threshold decreases exponentially to the standard threshold value.

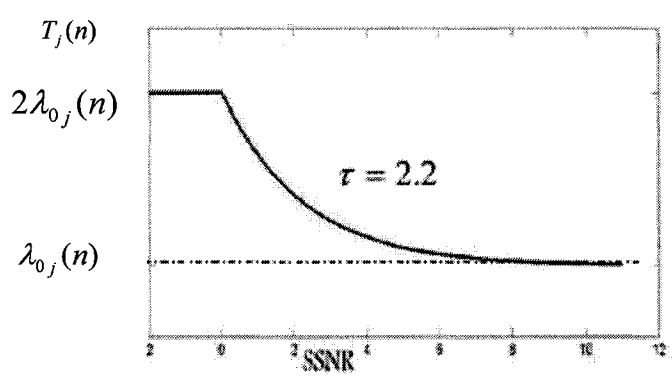


Figure 4-4 Aggravated threshold

[Ghanbari2005, p3]

Compared to Donoho's universal algorithm, this algorithm aggravates the threshold value dramatically. It removes the noise more completely, but in the mean time, it also results in more serious speech distortion.

To resolve these problems, the authors used a modified hard threshold to smooth the thresholding results. This function can be described as:

$$c_j(n)=\begin{cases} c_j(n), & |c_j(n)|\geq T_j(n) \\ sign(c_j(n))\cdot\frac{|c_j(n)|^{\gamma}}{T_j(n)^{\gamma-1}}, & |c_j(n)|<T_j(n) \end{cases} \quad \textbf{(4-18)}$$

As shown in Figure 4-5[Ghanbari2005], when the coefficients are under the threshold, they are non-lineally shrinked instead of being set to zero. It partly avoids the over-threshold of speech components of the signal and reduces the musical noise as well.

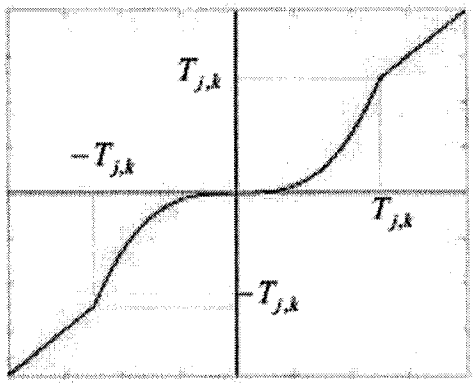


Figure 4-5 Smoothed hard thresholding

[Ghanbari2005, p3]

However, since the noise estimate is derived from the VAD, the inaccurate factors of VAD will be passed to the threshold calculation, and finally lead to inaccurate de-noising. Besides, the minimum value of the threshold is the universal threshold value. It means when the *SSNR* rises to a high value, the threshold keeps higher than necessary.

4.2.3 Teager Energy Operator

The Teager energy operator (TEO) is a powerful nonlinear operator proposed by H.M.Teager and S.M.Teager [Teager1990]. It is defined in both the continuous and discrete domains and is very useful for analyzing single component signals from an energy point-of –view. It has been successfully used in various speech applications [Bahoura2001].

For a given band-limited discrete speech signal $y(n)$, the discrete-time TEO can be approximated by

$$\psi[y(n)] = y^2(n) - y(n+1)y(n-1) \tag{4-19}$$

where, the energy operator spans three adjacent samples of the signal and is still a very local property of the signal.

This operator is able to effectively track the change in both amplitude and frequency of the signal [Teager1990] [Cairns1996] [Kaiser1990] [Kaiser1993] [Jabloun1999] [Chen2004]. Particularly, it is applied to the wavelet coefficients to enhance the discriminability of speech and non-speech frames in each subband generated from PWPD [Jabloun1999], as shown in Figure 4-6. When the TEO is high, it indicates that the current frame tends to speech segment, while the current frame will be judged to be pure noise frame when the TEO is close to 0. Thus the threshold algorithm can be designed to track the change of the TEO, so that to achieve the adaptive effect.

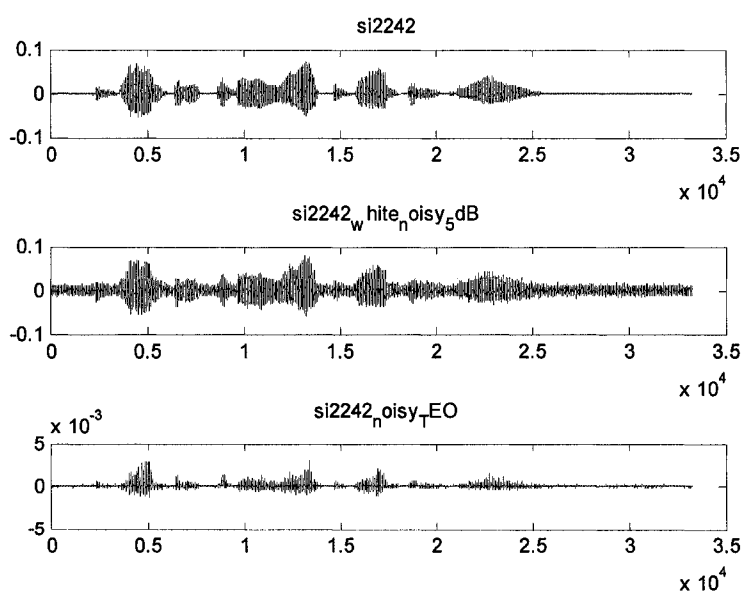


Figure 4-6 Teager's energy operator

Top: clean signal; middle: noisy signal; bottom: TEOs of the noisy signal

Applying the TEO algorithm to the wavelet domain, the TEO coefficients is obtained by

$$Teo_{i,j}(n) = \psi[c_{i,j}(n)] \tag{4-20}$$

where, i is the coefficient index, j is the subband index from the Wavelet Package Decomposition.

Then, an initial mask is obtained by smoothing the TEO coefficients using an IIR low-pass filter,

$$M_{i,j}(n) = Teo_{i,j}(n) * H_j(n) \tag{4-21}$$

A threshold adjuster is defined as [Jabloun1999]

$$\alpha_{i,j}(n) = \begin{cases} 0, & M_{i,j}(n)\Big|_{i=0}^{N_j} = 0 \\ \dfrac{M_{i,j}(n)}{\max(M_{i,j}(n))}, & otherwise \end{cases} \tag{4-22}$$

Therefore, the time adaptive threshold is defined as

$$\lambda_{i,j}(n) = \lambda_{0j}(n)(1 - \alpha_{i,j}(n)) \tag{4-23}$$

Finally, this threshold is applied into the soft thresholding. Figure 4-7 shows the complete algorithm.

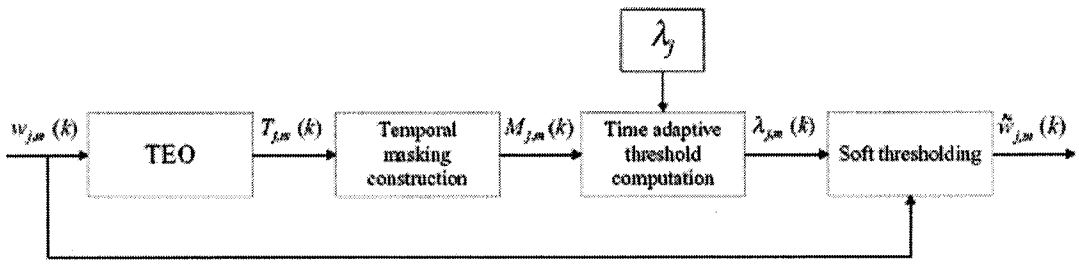


Figure 4-7 Adaptive thresholding tracking the change of Teager's energy operator
[Chen2004, p132]

4.3 Musical Noise Suppression

Musical noise is a typical problem in the field of single channel speech enhancement with spectral subtraction and wavelet thresholding. Musical noise is heard when an output has isolated peaks and/or short ridges in its spectrogram. It sounds metallic or tin-like. Generally, different frame length and overlapping rate result in different de-noising effect and different intensity of residual musical noise. Thus the frame length can not be too short, and the overlapped part should not be less than 50%. Other than optimizing the frame length and overlapping rate, several methods have been provided in literature to remove or smooth these isolated peaks and short ridges. In this thesis, a time-frequency adaptive smoothing method is also proposed to improve the musical noise suppression.

4.3.1 Floor Construction

Floor construction is the simplest method to smooth the residual noise. Using this method, the processed coefficients are set to a relatively low value instead of zero, such as setting them to1/10 of the original magnitude. Thus, a spectral floor is built up to reduce the difference from peak to peak. Normally, some light background noise will be introduced into the signal again, but the uncomfortable feeling of musical noise is lightened.

4.3.2 Adaptive Minimizing

An effective musical noise suppression method used with spectral subtraction is derived

from Boll's research in [Boll1979]. It replays each spectral coefficient after subtraction by the corresponding minimum spectral intensity value among the adjacent frames. Assume the maximal value of the residual noise measured during non-speech segment is $\max|W(\omega)|$. Then, the smoothing algorithm is

$$|S_m(\omega)| = \begin{cases} |S_m(\omega)|, & |S_m(\omega)| \geq max|W(\omega)| \\ min|S_j(\omega)|, & |S_m(\omega)| < max|W(\omega)| \end{cases}, j = m-1, m, m+1 \quad \textbf{(4-24)}$$

where, m is the frame index, and ω is frequency.

If we transfer this method into the wavelet domain, it can be defined as

$$|c_{i,j}(n)| = \begin{cases} |c_{i,j}(n)|, & |c_{i,j}(n)| \geq \max|\sigma_{i,j}(s)| \\ \min|c_{i,j}(n')|, & |c_{i,j}(n)| < \max|\sigma_{i,j}(s)| \end{cases}, n' = n-1, n, n+1 \quad \textbf{(4-25)}$$

where, n is the frame index, j is subband index, and i is coefficient index in a subband. $\max|\sigma_{i,j}(s)|$ denotes the maximal coefficient value of the residual noise measured during silent segment.

A drawback of this method is that it evidently increases the computational complexity.

4.3.3 Silent Segment Musical Noise Suppression

Profiting from the auditory masking phenomenon, the musical noise in speech frames is not as noticeable as in non-speech segment. Therefore, if the musical noise residual in silent segment could be mostly removed, the final de-noising result will be improved. In order to smooth the suppression result, some white noise at a proper intensity is added to the silent segment.

Similar to the algorithm in 4.3.2, this method depends on the result from VAD.

4.3.4 Adaptive Smoothing Based on Energy Analysis

Observing the musical noise through a spectrogram, it can be seen that this kind of noise has two peculiarities. One is they are isolate, and the other is they are scattered. It indicates that within an appropriate number of frames surrounding the noise, the local energies of these frames are noticeably uneven. In other word, within a speech or low-noise background segment at same length, the local energy of each frame is close to the average energy of this segment. Therefore, an adaptive algorithm based on local energy analysis is proposed in this research work.

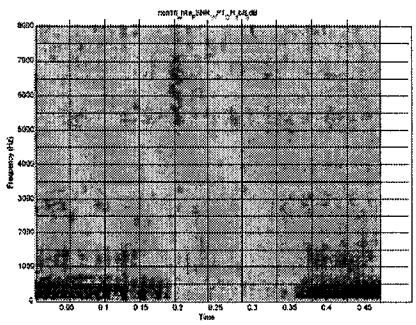


Figure 4-8 Time-frequency energy analysis

For a 16 kHz sampled signal, if we set the frame length equal to 256, then the time duration of each frame is 16ms. Assume the clean speech rich16 of 0.45s is polluted by white noise, and the signal-to-noise ratio is 5dB. Its spectrogram after adaptive thresholding is shown in Figure 4-8. Here 3 frames are selected as the processing segment mapping to 48ms. As can be seen, if the center frame is speech frame, the total energy of the segment is bigger than 2.5 times of the maximal energy in this segment, otherwise, the total energy is much smaller.

If the frame is musical noise, the coefficients will be set to the corresponding mean value of the segment. The mathematic model of this algorithm is illustrated as

$$c_{i,j}(n)=\begin{cases} c_{i,j}(n), & sum(PxT(n'))\geq max\,PxT(n')*2.5 \\ sign(c_{i,j}(n))mean\left|c_{i,j}(n')\right|, & sum(PxT(n'))<max\,PxT(n')*2.5 \end{cases}, n'=n-1,n,n+1 \quad \textbf{(4-26)}$$

where, PxT is the frame energy after adaptive thresholding.

Furthermore, since the frequency of human speech concentrates under 4 kHz, if the suppression at frequencies higher than 4 kHz is intensified, it will not have obvious damage to the speech parts. Thus, the above formula can be modified to

$$c_{i,j}(n)=\begin{cases} c_{i,j}(n), & sum(PxT(n'))\geq max\,PxT(n')*\varphi \\ sign(c_{i,j}(n))mean\left|c_{i,j}(n')\right|, & otherwise \end{cases}, n'=n-1,n,n+1 \quad \textbf{(4-27)}$$

when $f_j \leq 4kHz$, $\varphi=2.5$

when $f_j > 4kHz$, $\varphi=1.8$

Compared to the conventional methods, the proposed algorithm does not depend on voice activity detector, so that reducing the computation complexity. Moreover, using the mean value instead of the minimum value of the coefficients yields a smoother output.

4.4 Optimized Perceptual Adaptive Wavelet De-noising

In this thesis, a new speech enhancement system using adaptive wavelet de-noising is proposed. This algorithm uses wavelet packet transform to map the filter banks in the human inner ear. Adaptive noise estimate and threshold adjuster are adopted to track the local signal-to-noise ratio.

Compared to other adaptive threshold methods, the Quantile based noise estimate (QBNE) works well when the input SNR is low, reducing the residual noise evidently. That has been proven by the simulation results in chapter V. The problem of this method is over suppressing speech components due to universal statistic algorithm. It means the

thresholding is too aggressive when the local SNR is high. Thus, if a parameter is used to track the variety of the local SNR and adjust the final threshold, the thresholding could be more accurate.

On the other hand, the posteriori SNR time-adaptive threshold is a good adaptive thresholding adjuster. According to the simulation results, it works well on tracking the transformation of local signal-to-noise ratio under a non-stationary noise situation. The exponential smoothing and sigmoid function provide proper and continual adaptive performance. Additionally, this algorithm extracts the noise power from the noisy speech signal alone, avoiding the voice activity detection. But, compared to the QBNE based method, there is more residual noise left. This shortcoming mostly comes from the inaccurate noise estimate.

It is interesting that these two methods are compensatory to each other. The QBNE provides a good adaptive noise estimate, while the PSNR contributes the ability of tracking the local SNR. Therefore, an optimized thresholding method may be obtained through combining them together.

To build such a system, the QBNE based threshold has been used as the basic adaptive threshold, replacing the original standard threshold. And the PSNR based smoothing parameter has been used as the thresholding adjustor. The system blocks are shown in Figure 4-9. Here, the posterior SNR is modified into standard SNR format instead of a ratio, so that the sigmoid function can reflect the local SNR intuitively. The modified PSNR is presented as

$$PSNR_j(n) = 10\log(Px_j(n)/\overline{P}n_j(n-1)) \tag{4-28}$$

The fourth part of the optimized system is the new algorithm for musical noise suppression discussed in 4.3.4. The isolated and scattered musical noise components are extracted from the first-step de-noising result, based on the adaptive local energy

analysis. And then, these musical noise coefficients are set to coefficient average of the adjoining frontward and backward frames. The contribution of this new method is that the analysis model describes the properties of the musical noise, so that the corresponding noise suppression becomes more accurate.

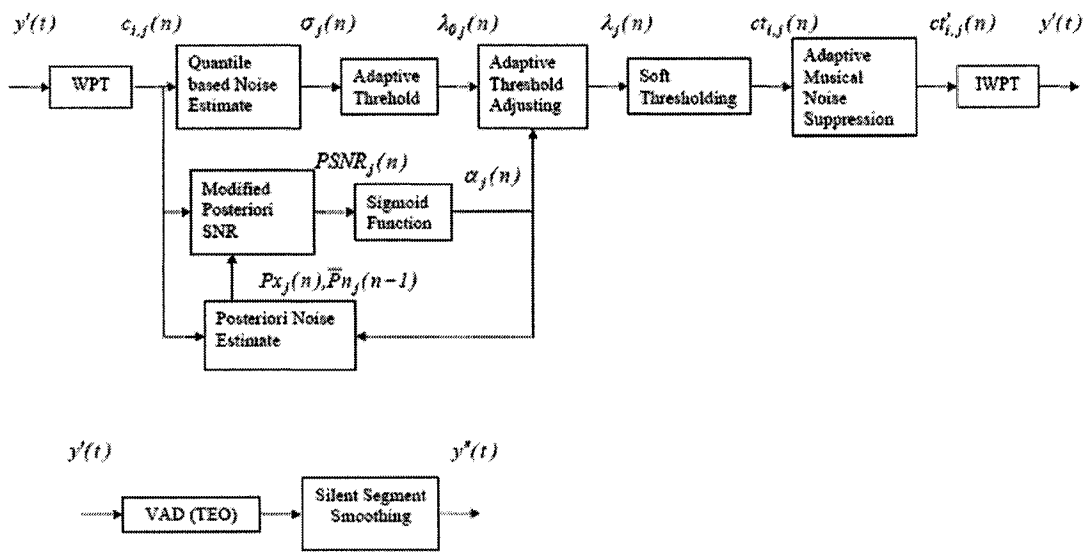


Figure 4-9 Proposed adaptive wavelet speech enhancement system

After the wavelet domain processing, a novel time domain silent segment smoothing module was also added into the system. The purpose of this module is to smooth the residual noise left in silent segment, in order to improve the final perceptual effect. Multi-frame TEO analysis is adopted to perform the voice activity detection. The processing can be decomposed into four steps:

1) To calculate the TEO value of each frame

$$\psi[y(n)] = y^2(n) - y(n+1)y(n-1) \qquad (4\text{-}29)$$

where, n is the frame index.

2) Take the TEO absolute values of current frame and three previous frames for calculating the ratio of minimum and maximal TEO absolute value within these four frames.

$$TEO_{min} = min(abs(TEO(1:L_{TEO})) \quad (4\text{-}30)$$

$$TEO_{max} = max(abs(TEO(1:L_{TEO})) \quad (4\text{-}31)$$

where, L_{TEO} is the number of frame in the processing segment, equal to 4. Then, the expected ratio is presented as

$$M = TEO_{min} / TEO\,max \quad (4\text{-}32)$$

3) Perform judgment. When M is smaller than a particular value a, it means the TEO values in the processing segment are even. When TEO_{max} is smaller than another particular small value b, it means the TEO values in the processing segment are quite small. According to the previous waveform analysis in 4.2.3, they are just the two properties of the silent segment waveform. Thus, if these two conditions are satisfied, we consider the current segment as silent segment.

$$F = \begin{cases} SilentSegment, & and(M < a, TEO_{max} < b) \\ SpeechSegment, & otherwise \end{cases} \quad (4\text{-}33)$$

where, F works as the flag of the judgment.

4) Smooth the samples of the silent segment. Here we use a mean value filter to perform the smoothing to each sample within the current segment.

CHAPTER V

5. SIMULATION AND RESULTS ANALYSIS

In chapter III and IV, the latest algorithms of perceptual adaptive wavelet speech enhancement have been introduced. An innovative combination of QBNE and PSNR, a new method of musical noise suppression, a new TEO based time-domain silent segment processing module, and the optimized wavelet speech enhancement system are proposed. It is an important part of this thesis work to comparing and evaluating these methods through Matlab simulation. This chapter will discuss the details of Matlab simulation followed by the analysis of the results.

5.1 Matlab Simulation Setup

5.1.1 Speech

The original speeches used for simulation and test are taken from the famous TIMIT speech databases. TIMIT is a corpus of phonetically labelled transcribed speech of American English speakers of different sexes and dialects. It has been widely used for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. This database was commissioned by DARPA and worked on by many sites, including Texas Instruments (TI) and Massachusetts Institute of Technology (MIT), hence the corpus was named.

TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States, including New England, Northern, North Midland, South Midland, Southern, New York City, Western and Army Brat. These sentences are assorted into three types. The dialect sentences (the SA sentences) were meant to expose the dialectal variants of the speakers and were read by all 630 speakers. The phonetically-compact sentences (the SX sentences) were designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. The phonetically-diverse sentences

(the SI sentences) were selected from existing text sources, so as to add diversity in sentence types and phonetic contexts.

In this thesis, 16 sentences spoken by 8 female and 8 male from 8 dialect regions are selected from TIMIT as the benchmark speech signals. All of them are sampled at 16 kHz, and quantized into 16 bits.

5.1.2 Noise

Both stationary and non-stationary noises are considered in this thesis. The stationary artificial noises, such as White Gaussian Noise (WGN), are generated at desired intensity using Matlab function directly. And the real-life noises are selected from the NOISEX-92 database. In the area of speech processing researches, NOISEX-92 is a well known standard noise database, recording various real-life noises. Fifteen stationary and non-stationary noise samples are involved in the simulation, including WGN, pink noise, voice babble, HF radio channel noise, factory floor noise, jet cockpit noise, destroyer engine room noise, F-16 cockpit noise, military vehicle noises, tank noise, machine gun noise, and car interior noise. All of the noises are down-sampled from 19.98 kHz to 16 kHz, equal to the sampling rate of the speech signals.

To simulate a typical non-stationary noise, several stationary noises are randomly mixed together segment by segment in this thesis. For example, it is assumed that the white Gaussian noise is the basic background noise, and the first one second of the signal is also polluted by speech babble, while the remaining parts are polluted by pink or car interior noise. The process is illustrated in Figure 5-1. Since the power level of each noise is unequal, a slowly varying input signal-to-noise ratio can be achieved within the whole noisy signal.

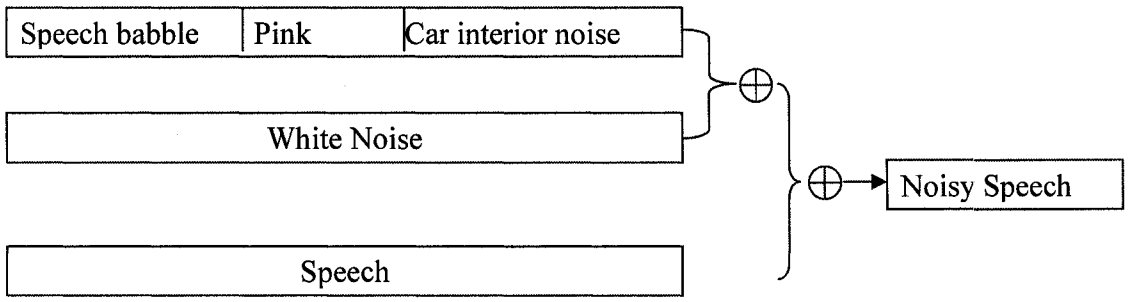


Figure 5-1 Creation of non-stationary noise

5.1.3 Noisy Signals

The noisy signal is generated by adding a noise signal $w(n)$ to a clean speech signal $s(n)$. Thus the noisy signal is given by

$$y(n) = s(n) + \lambda \cdot w(n) \tag{5-1}$$

where, parameter λ decides the intensity of the noise signal added to the clean speech, thus decides the signal-to-noise ratio.

Assuming a noisy signal at SNR_{input} is tested, then

$$SNR_{input} = 10\log_{10}\frac{\sum s^2(n)}{\sum \lambda^2 w^2(n)} \tag{5-2}$$

$$\lambda = \sqrt{\frac{\sum s^2(n)}{\sum w^2(n)} \cdot 10^{\frac{-SNR_{input}}{10}}} \tag{5-3}$$

The value of 256 is taken as the length of each frame here, mapping to 16 ms of 16 kHz signals. These frames are overlapped by each other at the overlapping rate 50%.

The Matlab simulation setup is summarized in Table 5-1

Table 5-1 Matlab simulation setup

	Type	Source	Sampling rate	Length
speech	8 femal,8 male	TIMIT	16 kHz	2s~4s
noise	WGN	Matlab	16 kHz	/
	15 Real-life	Noisex-92	19.98 kHz	/
	Artificial non-stationary	proposed	16 kHz	/
	Value			
SNR_{input}	0dB~15dB			
Frame length	256 samples, 16 ms			
Overlapping	50%			

5.2 Evaluation Methods

Quality measure of speech enhancement is generally classified into subjective evaluation and objective evaluation. Subjective measures evaluate the perceptual quality of a speech based on the subjective rating by human listeners. Currently the most accurate and preferable method of speech enhancement rating is subjective evaluation [Hu2006]. This method, however, is time consuming and costly. Comparing to subjective evaluation, quantized objective measures are faster and more economical, but shows low correlation with the subjective speech quality [Quackenbush1988]. Spectrogram accurately reflects the dissimilarity between the original clean speech and processed signal by making the speech visible. It is sorted into objective measures in this thesis.

5.2.1 Subjective Evaluation

The subjective evaluation in this thesis research is derived from ITU-T recommendation P.835 and was conducted by Dynastat, Inc [Hu2006]. In order to reduce the listener's uncertainty in a subjective test, three components of a noisy speech signal, the speech signal, the background noise and the overall effect, are considered. The process of rating the enhanced speech is:

1. Rating the speech signal alone using a five-point scale of signal distortion (SIG) (Table 5-2).
2. Rating the background noise alone using a five-point scale of background intrusiveness (BAK) (Table 5-3).
3. Rating the overall effect using a five-point scale of the Mean Opinion Score (Table 5-4).

Five male and female listeners attended the subjective test. Sixteen sentences from TIMIT, polluted by sixteen different types of noise at different input signal-to-noise ratio, are evaluated.

Table 5-2 Scale of signal distortion (SIG)

[Hu2006, p2]

5- Very natural, no degradation
4- Fairly natural, little degradation
3- Somewhat natural, somewhat degraded
2- Fairly unnatural, fairly degraded
1- Very unnatural, very degraded

Table 5-3 Scale of background intrusiveness (BAK)

[Hu2006, p2]

5- not noticeable
4- somewhat noticeable
3- Noticeable but not intrusive
2- Fairly conspicuous, somewhat intrusive
1- Very conspicuous, very intrusive

Table 5-4 Scale of overall effect

5- excellent
4- good
3- fair
2- poor
1- bad

Previous researches proved that the overall subjective evaluation is influenced more by speech distortion. A regression analysis was designed to substantiate this phenomenon [Hu2006]. As shown in equation (5-4), the predicted overall score was considered as the function of the rating score of the speech and noise distortion.

$$R_{OVRL} = -0.0783 + 0.571 \cdot R_{SIG} + 0.366 \cdot R_{BAK} \qquad \textbf{(5-4)}$$

where R_{OVRL} is the predicted overall rating, R_{SIG} and R_{BAK} denote the SIG and BAK rating respectively. According to Y. Hu and P.C. Loizou's test results, the predicted overall rating scores are quite close to the real overall rating. It confirms that listeners integrate the effects of both speech signal and background distortion when making their ratings. And, these two types of distortion contribute differently to the overall evaluation.

Listeners seem to pay more attention to the speech distortion rather than to the background noise.

5.2.2 Objective Evaluation

Objective measures are the methods using mathematical models to evaluate the processing quality. Different several objective speech quality measures have been widely used including global SNR, segmental SNR(segSNR), weighted-slope spectral (WSS) distance, perceptual evaluation of speech quality, log likelihood ratio (LLR) and Itakura-Saito (IS) distance measure, etc[Hu2006]. As the most popular evaluation indexes, SNR and segSNR are recruited in the simulation work of this thesis.

5.2.2.1 Global SNR and Segmental SNR

Global SNR (SNR) is defined as the ratio of the clean speech power to the noise power, obtained globally from the time domain. The calculation of input SNR and output SNR uses the unit of decibels (dB) and is defined as

$$SNR_{input} = 10 \cdot \log_{10}\left[\frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} w^2(n)}\right] \tag{5-5}$$

$$SNR_{outputt} = 10 \cdot \log_{10}\left[\frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} (s(n) - \hat{s}(n))^2}\right] \tag{5-6}$$

where $s(n)$ is the clean speech, $w(n)$ is the additive background noise, and $\hat{s}(n)$ is the processed speech signal.

Segmental SNR (segSNR) is used more widely for its higher correlation degree to the subjective results. Instead of taking the global data for the calculation, this method takes over short segments of the speech signal and then recruits the mean SNR value of the overall segments as the evaluation result. It can be denoted as

$$segSNR = \frac{1}{L} 10 \cdot \sum_{l=0}^{L-1} log_{10} \left[\frac{\sum_{n=Nm}^{Nm+N-1} s^2(n)}{\sum_{n=Nm}^{Nm+N-1} (s(n) - \hat{s}(n))^2} \right] \tag{5-7}$$

where L is the total number of the frames, Nm represents the number of samples in each frame.

Both of global SNR and segmental SNR result low correlation with overall subjective evaluation. Thus they only work as accessory evaluation measures in this thesis.

5.2.2.2 Spectrogram Analysis

The spectrogram is color-based visualizations of the evolution of the power spectrum of a speech signal as it changes over time. It is generally created by calculating the frequency spectrum of windowed frames (STFT) of a compound signal. In a spectrogram, the horizontal dimension represents time and the vertical dimension represents frequency. Each thin vertical slice of the spectrogram shows the spectrum during a short period of time, using darkness to stand for amplitude. Darker areas show those frequencies where the simple component waves have high amplitude. An example of spectrogram is shown in Figure 5-2. The content of sentence si2242 from TIMIT database is "twenty two and twenty three".

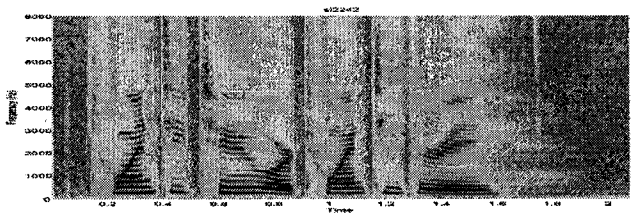


Figure 5-2 Spectrogram of speech signal si2242 from TIMIT

Spectrograms are widely used for speech and audio analysis. As we can see from above figure, spectrogram reflects all the information of frequency, signal intensity, and time period. For the clean signal, the background is pure and smooth, without abrupt change. While the signal is polluted by a noise, its spectrogram shows a noisy background as displayed in Figure 5-3. According to the experiment results, spectrograms are highly correlated to the subject evaluation. In the mean time, it remains the advantage of objective measures, low time consuming and low cost. The shortcoming of this method is that it is not quantized, thus not as convenient as a quality indicator for researchers.

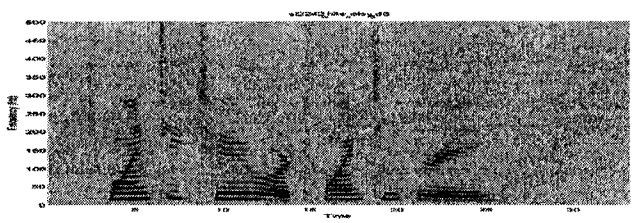


Figure 5-3 Spectrogram of white noisy si2242 at SNR=5dB

5.3 Matlab Simulation and Results

In this section, Matlab simulation steps and results of each method will be presented. Algorithms are grouped according to the methodology they use. Most comparisons in this section depend on the subjective measures, spectrogram and time-domain plot, which are highly correlative with human subjective response. SNR and segSNR are recruited to prove the rough tendency of enhancement results. As shown in Table 5-1, 16 sentences, 17 noise situations, and 4 SNR are involved in the whole experiment. Since the experiment results for different speech and noises tend to be consistent, only 2 sentences

(one female and one male), 2 noises (WGN and artificial non-stationary noise), and 3 SNR are explained in this section.

Here Daubechies wavelets are selected as the mother wavelet, since they best preserve the frequency selectivity as the number of stages of the DWPT increases. It has been proven that db8 or db10 is the best to describe speech signals. However, there is a limit of decomposition level at a particular frame length and particular wavelet. For example, if the frame length is 256, the maximal decomposition level achieved with db8 is four. Thus, only db1 or db2 is available for six level mapping for the perceptual wavelet thresholding. From this point, db2 is of benefit to more delicate frequency analysis. To select a proper mother wavelet for DWPT and PWT decomposition, all of db2, db4 and db8 were tried in this thesis work. Results of these three wavelets, however, are approximately the same. Therefore, 4-level DWPT decomposition with wavelet db8 for and 6-level PWT decomposition with wavelet db2 are adopted.

5.3.1 Standard Wavelet De-noising and SureShrink

The purpose of the comparison within this group is to set the better algorithm and the corresponding results as the benchmark of the following simulation and analysis. Figure 5-4, 5-5, and 5-6 illustrate the time-domain waveform of the signals corrupted by Gaussian White Noise, with different $SNR_{input,}$ before and after the processing. Obviously, the standard wavelet de-noising with universal threshold removes more background noise than the SureShrink does. However, it evidently distorts the speech components in the mean time.

a)

b)

c)

d)

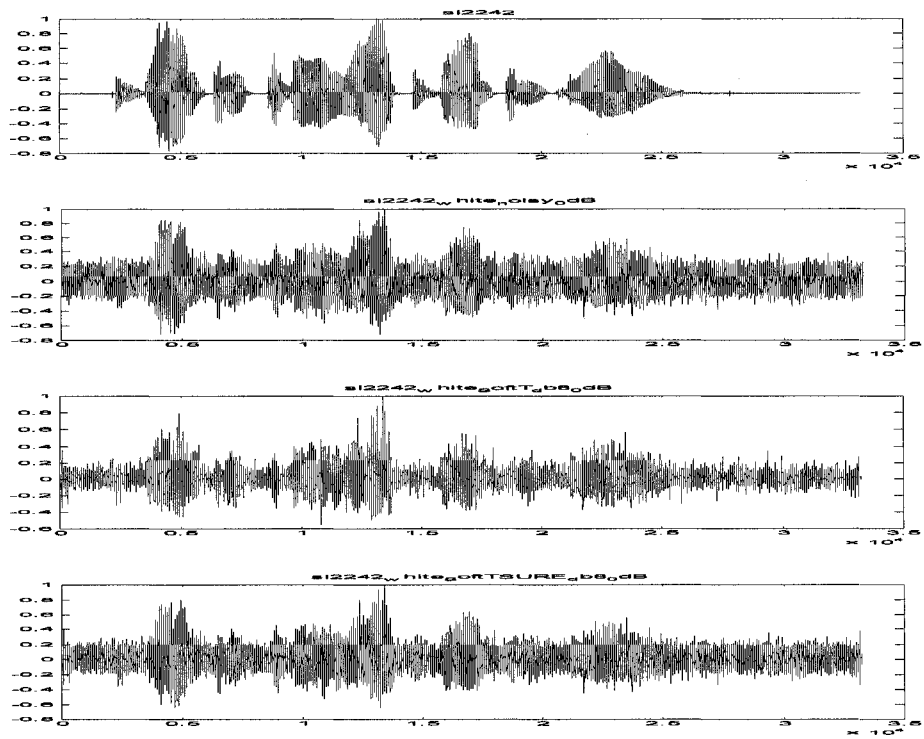


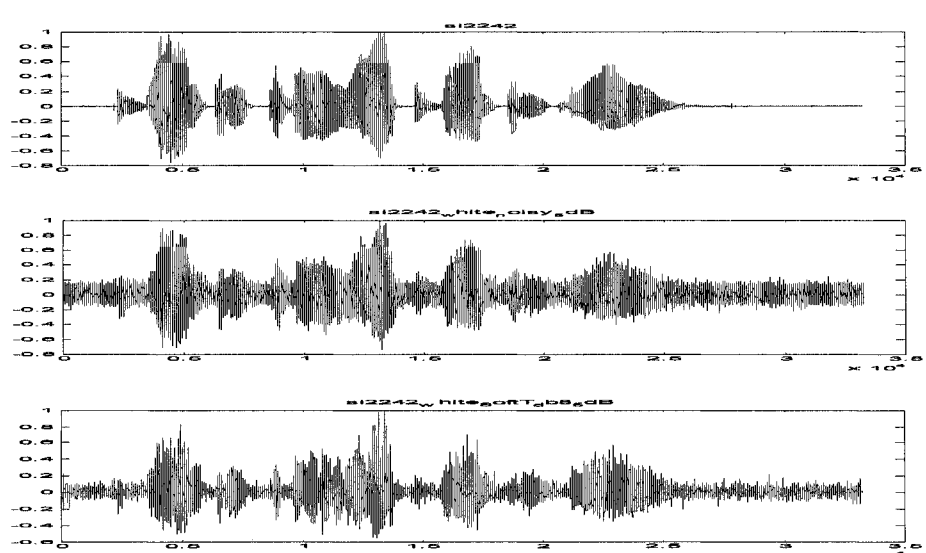
Figure 5-4 Time-domain waveforms with SNR_{input} =0dB

(a) Clean Signal (b) WGN noisy signal(c) Enhanced signal by standard wavelet de-nosing (d) Enhanced signal by Sureshrink

a)

b)

c)



d)

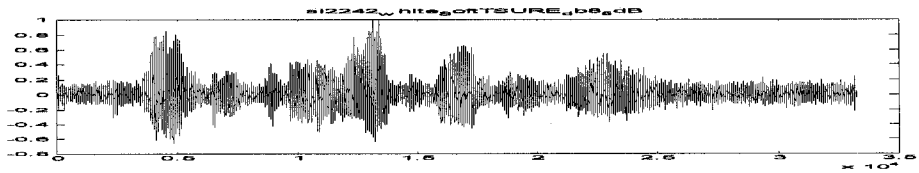


Figure 5-5 Time-domain waveforms with SNR_{input} =5dB

(a) Clean Signal (b) WGN noisy signal(c) Enhanced signal by standard wavelet de-nosing (d) Enhanced signal by Sureshrink

a)

b)

c)

d)

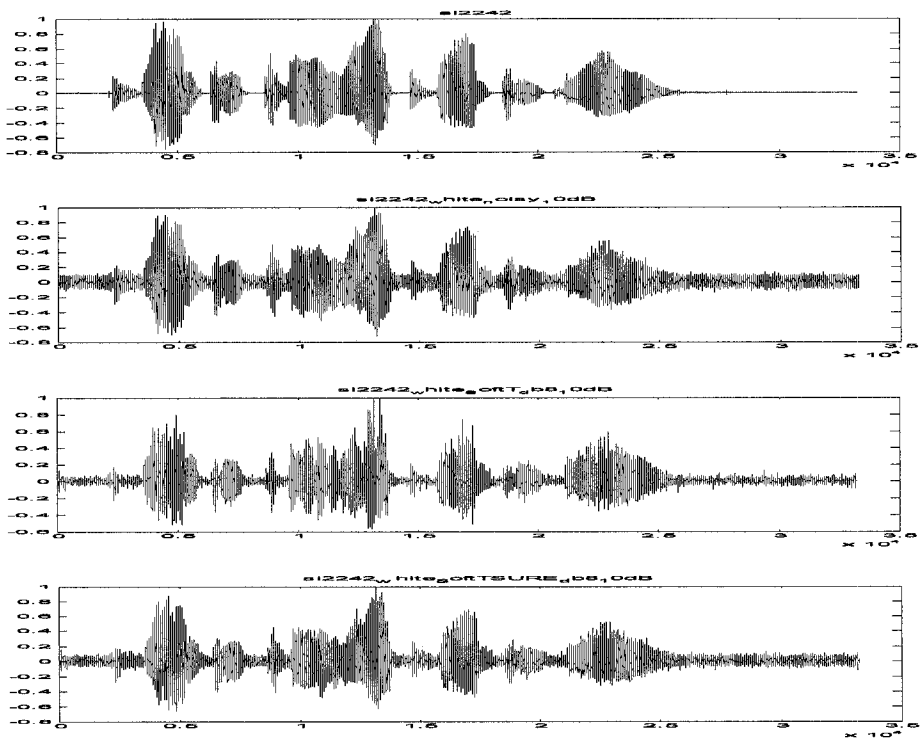


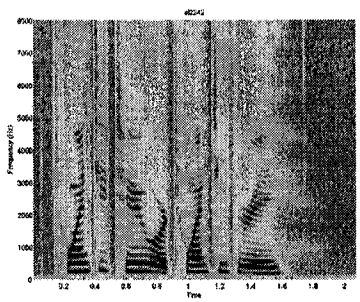
Figure 5-6 Time-domain waveforms with SNR_{input} =10dB

(a) Clean Signal (b) WGN noisy signal (c) Enhanced signal by standard wavelet de-nosing (d) Enhanced signal by Sureshrink

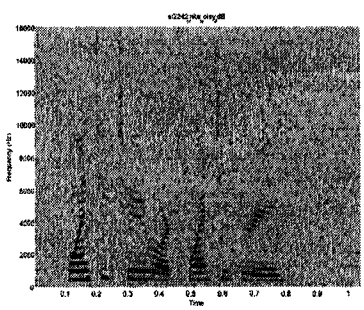
The same tendency is reflected by the spectrograms, as shown in Figure 5-7. The standard method, shown in (c0~2), sacrifices the speech fidelity, while performs well in removing the background noise. Furthermore, since the standard soft threshold does not process the approximate part of the decomposed signal, evident residual noise in the corresponding frequency subband has been left. On the contrary, the SureShrink, shown

in (d0~2), leaves more background noises, but gets ahead of the standard method by remaining the speech components well. As discussed before, the overall subjective evaluation is influenced more by speech distortion, thus SureShrink is supposed to yield better subjective effects.

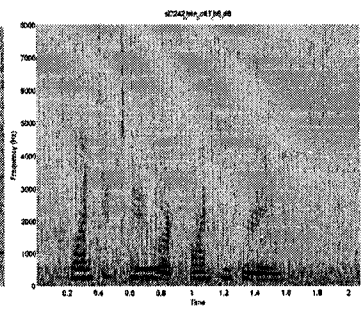
This hypothesize has been proved in Table 5-5. The SureShrink provides better subjective evaluation scores than the standard method does, and the standard method even gives worse subjective scores than the noisy signal. Under the real-life noise and the mixed non-stationary noise (MNSN) situation, the simulations yielded similar results. Thus, SureShrink was selected as the benchmark of the following comparison.



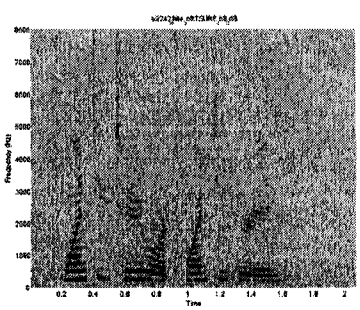
(a) Clean Signal



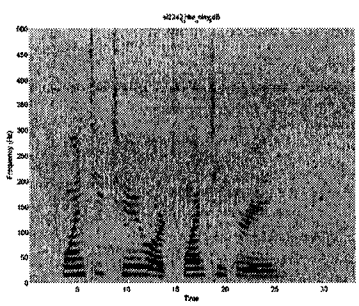
(b0)WGN noisy signal 0dB



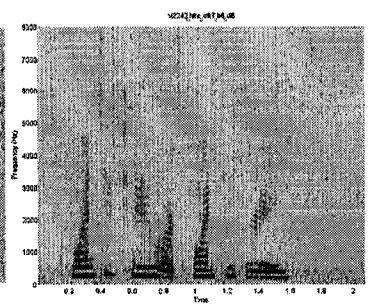
(c0) Enhanced by standard ST



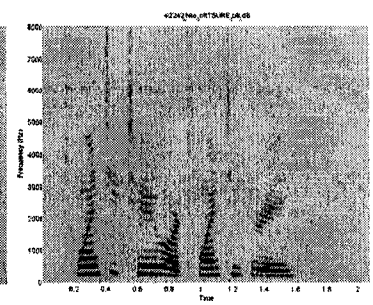
(d0) Enhanced by SureShrink



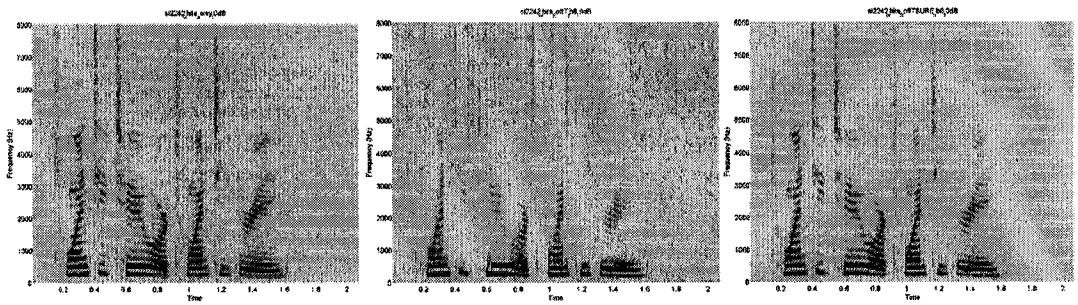
(b1)WGN noisy signal 5dB



(c1) Enhanced by standard ST



(d1) Enhanced by SureShrink



(b2)WGN noisy signal 10dB (c2) Enhanced by standard ST (d2) Enhanced by SureShrink

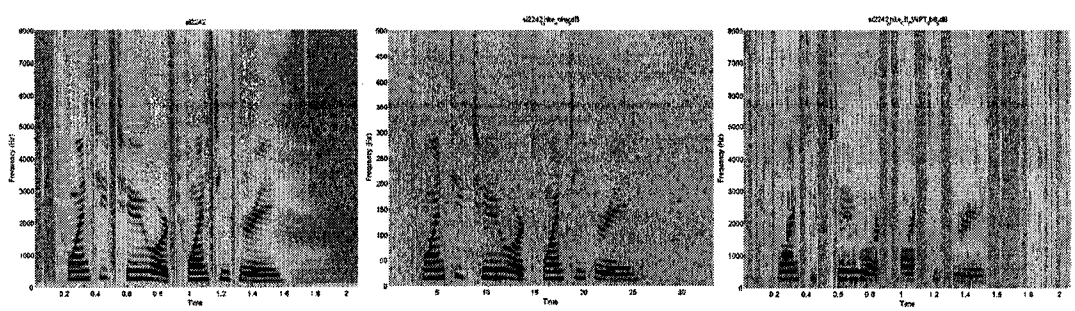
Figure 5-7 Spectrograms of si2242

Table 5-5 Subjective evaluation of standard threshold and SureShrink

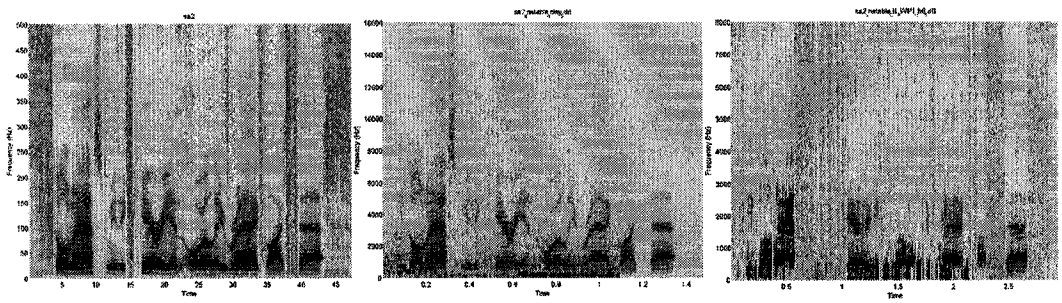
Speech	SNR_{input}	Noise Type	Noisy Speech			Standard T			SureShrink		
			SIG	BAK	ORL	SIG	BAK	ORL	SIG	BAK	ORL
Si2242	0dB	WGN	4	1	2.572	2	2	1.796	4	2	2.938
		MNSN	4	1	2.572	2	2	1.796	4	2	2.938
	5dB	WGN	4	2	2.938	2	3	2.162	4	3	3.304
		MNSN	4	2	2.938	2	3	2.162	4	3	3.304
	10 dB	WGN	4	3	3.304	3	3	2.733	4	4	3.67
		MNSN	4	3	3.304	3	3	2.733	4	4	3.67

5.3.2 Perceptual Wavelet Thresholding

The simulation of this method was performed with 6-level PWT wavelet decomposition. Minimum 125 kHz bandwidth was achieved. Noise masking threshold was used to adjust the noise suppression. The results are shown in Figure 5-8 as below.



(a) Clean Signal (b) WGN noisy signal at 5dB (c) Enhanced by PWT



(a1) Clean Signal (b1) non-stationary noisy signal at 5dB (c1) Enhanced by PWT

Figure 5-8 Spectrograms of perceptual wavelet thresholding

It can be seen from the spectrogram (c) and (c1) that most noises during the silent segments were removed. However, the speech components were distorted evidently. Especially, those abrupt cuts at the edge of speech parts result in unexpected and sharp noise, which make the subjective effect quite bad as shown in Table5-6.

Another disadvantage of this method is high computation complex. Complicated decomposition and processing steps make real-time speech enhancement even harder.

Table 5-6 Subjective evaluation of SureShrink and perceptual wavelet thresholding

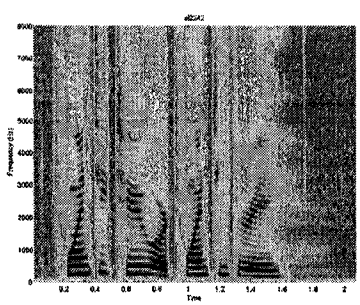
Speech	SNR_{input}	Noise Type	Noisy Speech			SureShrink			PWT		
			SIG	BAK	ORL	SIG	BAK	ORL	SIG	BAK	ORL
Si2242	0dB	WGN	4	1	2.572	4	2	2.94	2	2	1.79
		MNSN	4	1	2.572	4	2	2.94	2	2	1.79
	5dB	WGN	4	2	2.938	4	3	3.304	3	3	2.73
		MNSN	4	2	2.938	4	3	3.304	3	3	2.73
	10 dB	WGN	4	3	3.304	4	3	3.304	3	4	3.10
		MNSN	4	3	3.304	4	3	3.304	3	4	3.10

5.3.3 Comparison of Adaptive Threshold Algorithms

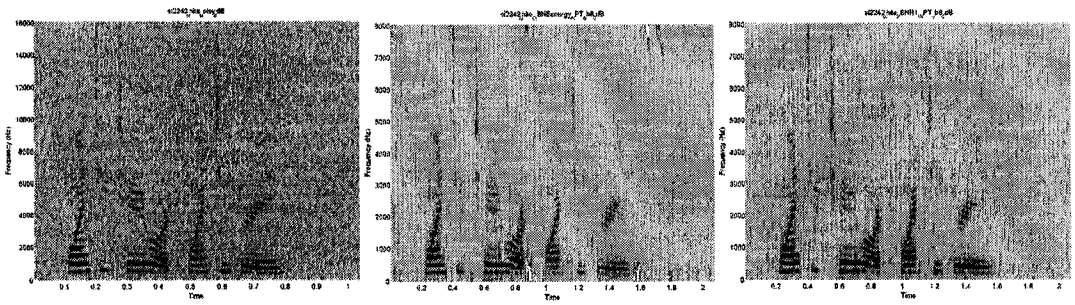
In chapter IV, two types of adaptive algorithms are introduced, the adaptive estimation and the adaptive adjuster. Four adaptive speech enhancement methods were discussed in detail. They are

1. Quantile-based Time-frequency Noise Estimate combined with standard soft threshold (QBNE)
2. Posteriori SNR Time-Adaptive Threshold (PSNRAT)
3. Smoothed Hard Thresholding with Aggravated Threshold Value (SHTAT)
4. Teager Energy Operator based Adaptive Threshold (TEOAT)

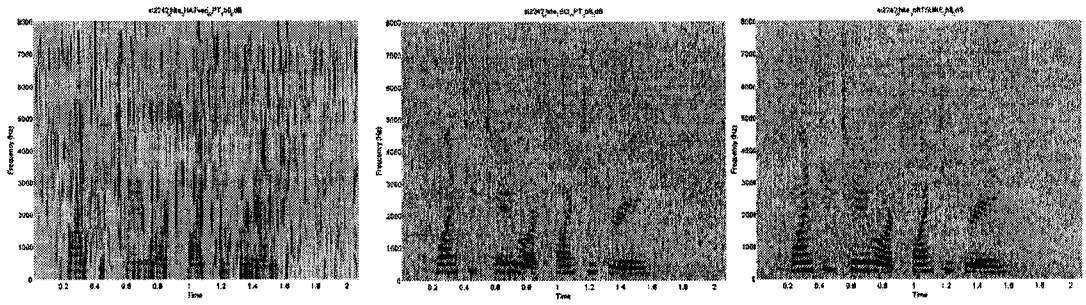
Since the spectrogram is highly correlated to human subjective evaluation, the comparison in this group will primarily depend on it. Figure 5-9 illustrates the de-noising of 5dB WGN corrupted si2242, from a female speaker. It can be seen that the method of SHTAT (a) damage the speech signal too much. Although TEOAT (f) works gently, it is not satisfying with some evident distortion and weak de-noising result. Correspondingly, QBNE (c) and PSNRAT (d) show better performance with low-level distortion and effective de-noising. Comparing to the traditional SureShrink (g) algorithm, QBNE (c) removed much more noises, however, introduced noticeable residual musical noise and slight speech distortion. Some speech edge components had been cut as shown in spectrogram (c). PSNRAT (d) get an advantage over QBNE (c) of remaining the details of speech components though it resulted in slightly heavier residual noise.



(a) Clean Signal



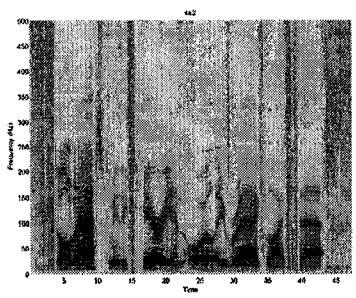
(b) 5dB WGN corrupted si2242 (c) Enhanced by QBNE (d) Enhanced by PSNRAT



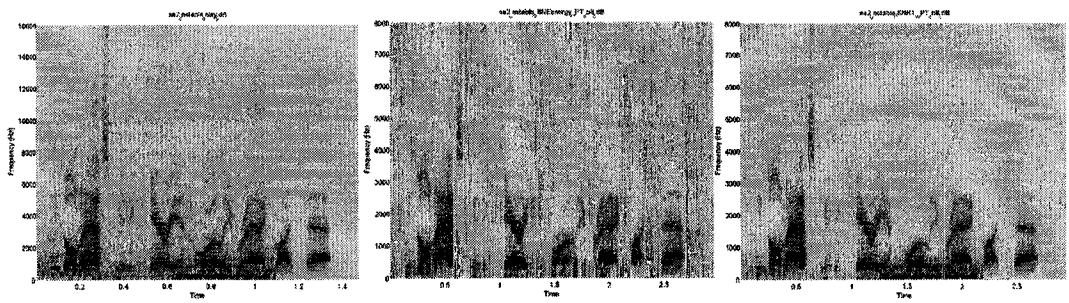
(e) Enhanced by SHTAT (f) Enhanced by TEOAT (g) Enhanced by SureShrink

Figure 5-9 De-noising of 5dB WGN corrupted si2242

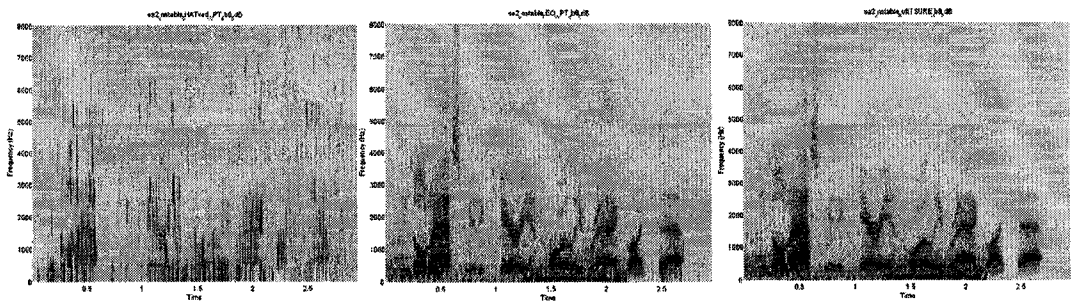
In Figure 5-10, the de-noising of 5dB non-stationary noise corrupted sa2, from a male speaker, is shown. The similar results as those under WGN situation were obtained. Among all the four algorithms, QBNE (c) and PSNRAT (d) provided distinctly better de-noising result, over the SureShrink algorithm too. The former gives a clearer background but worse distortion, while the latter shows a better trade-off between the concern of de-noising and reducing the speech distortion.



(a) Clean Signal



(b) 5dB NSN corrupted sa2 (c) Enhanced by QBNE (d) Enhanced by PSNRAT



(e) Enhanced by SHTAT (f) Enhanced by TEOAT (g) Enhanced by SureShrink

Figure 5-10 De-noising of 5dB non-stationary noise corrupted sa2

5.3.4 Comparison of Musical Noise Suppression Methods

Although QBNE and PSNRAT have better de-noising effect than the traditional standard method, noticeable residual noises are left. In this section, the four musical noise suppression methods introduced in chapter 4 are compared within a group. These four methods include Floor Construction, Adaptive Minimizing, proposed Adaptive Smoothing, Silent Segment Suppression combined with Adaptive Smoothing.
Both QBNE and PSNRAT are recruited.

For the QBNE based musical noise suppression, each simulation result from each method is shown in Figure 5-11. Obviously, musical noise suppression by proposed Adaptive Smoothing (e) is much more effective than floor construction and adaptive minimizing, maintaining the speech well and smoothed most musical noise. The musical noise suppression by silent segment suppression combined with Adaptive Smoothing (f) is

obtained from suppressing the silent segment from the whole signal already processed by adaptive smoothing. There is an abrupt change between the speech and silent segment, though a certain amount of Gaussian White noise has been added to the signal. This steep edge yields an annoying sound embed in the whole signal. In worse cases, the inaccurate voice activity detection may result in incorrect suppression of speech parts.

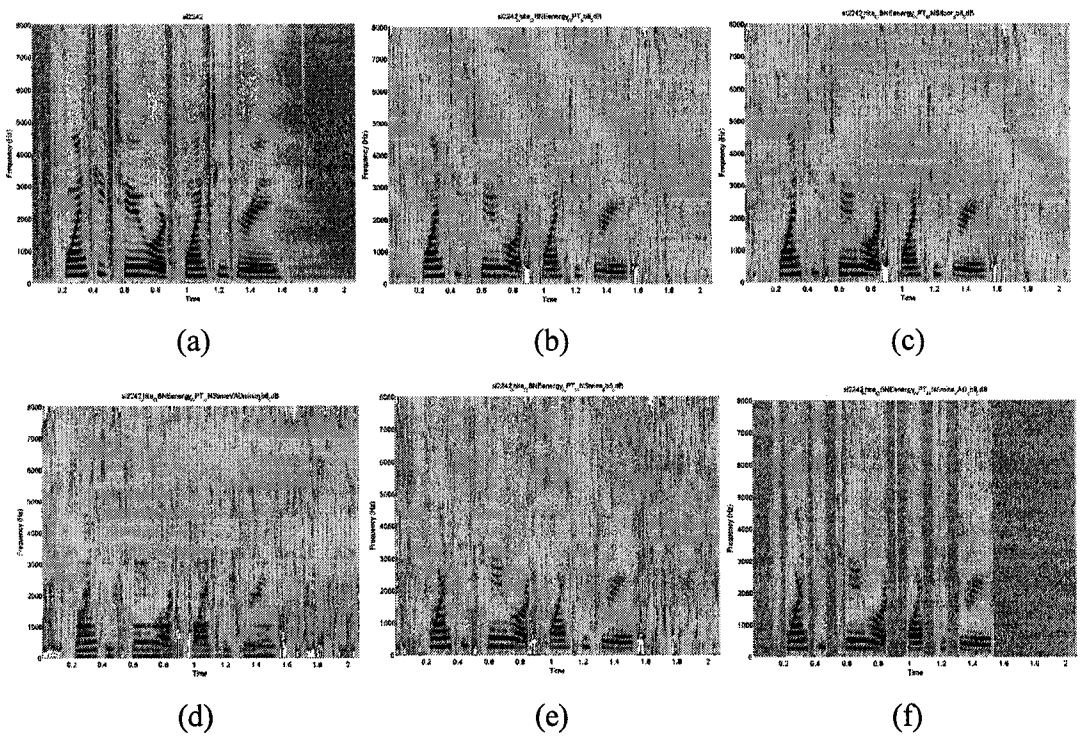


Figure 5-11 Musical noise suppression of 5dB GWN corrupted si2242 based on QBNE

(a) Clean Signal (b) Enhanced signal by PSNRAT (c) Musical noise suppression by Floor Construction (d) by Adaptive Minimizing (e) by proposed Adaptive Smoothing (f) by Silent Segment Suppression combined with Adaptive Smoothing

For the PSNRAT based musical noise suppression, the simulation similar results shown in Figure 5-12. Among the four methods, the proposed Adaptive Smoothing (e) yields best musical noise suppression result. There is more residual noise left because the PSNRAT brings heavier residual noise to this processing part. However, since the human auditory system is more sensitive to the speech distortion than to the background noise, PSNRAT still gets a little bit higher score than QBNE does, as shown in Table 5-7.

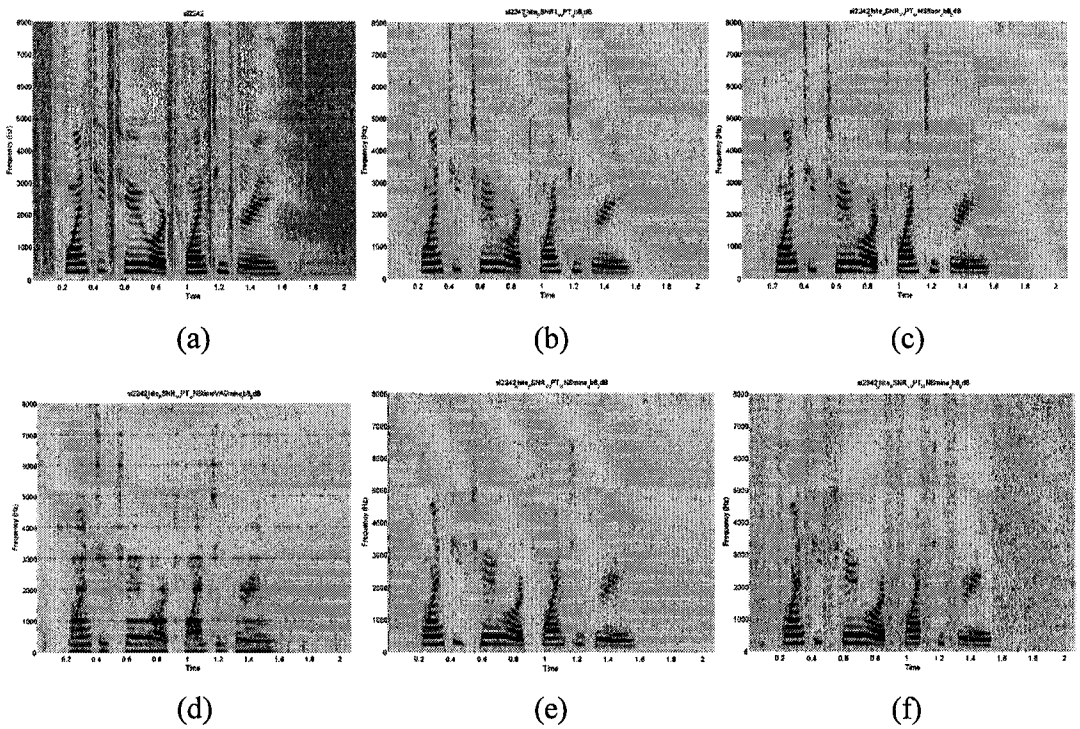


Figure 5-12 Musical noise suppression of 5dB GWN corrupted si2242 based on PSNRAT

(a) Clean Signal (b) Enhanced signal by PSNRAT (c) Musical noise suppression by Floor Construction (d) by Adaptive Minimizing (e) by proposed Adaptive Smoothing (f) by Silent Segment Suppression combined with Adaptive Smoothing

Table 5-7 Subjective evaluation of musical noise suppression by proposed adaptive smoothing

Speech	SNR_{input}	Noise Type	Noisy Speech			QBNE based			PSNRAT based		
			SIG	BAK	ORL	SIG	BAK	ORL	SIG	BAK	ORL
Si2242	5dB	WGN	4	2	2.938	3	4	3.0987	4	3	3.304

5.3.5 Proposed Adaptive Wavelet Speech Enhancement System

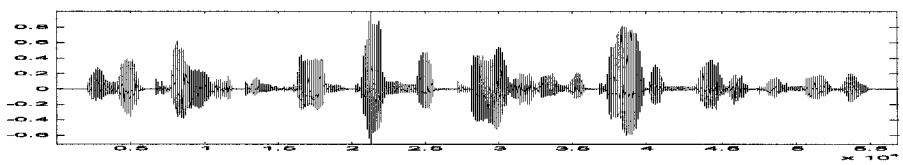
The principle and structure of the proposed optimized adaptive wavelet speech enhancement system was introduced in chapter IV. This section will focus on the comparison among the traditional standard soft thresholding, SureShrink, QBNE

thresholding, PSNR thresholding, and the proposed system. Two typical noise cases are recruited, WGN and Babble (non-stationary).

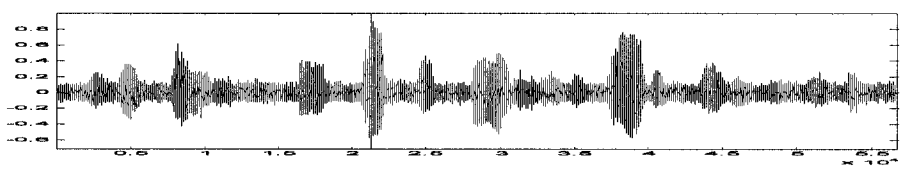
i) WGN Environment

Figure 5-13 illustrates the waveforms of the signals before and after the enhancement. It is obvious that (g) Enhanced signal by proposed system has a stronger effect of removing background noises than the other methods do.

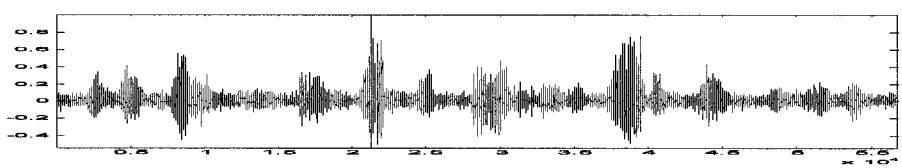
(a) Clean signal of sx366



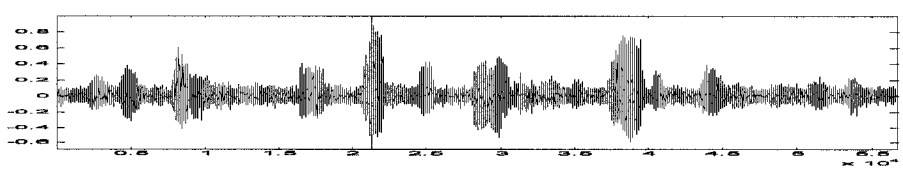
(b) Corrupted signal



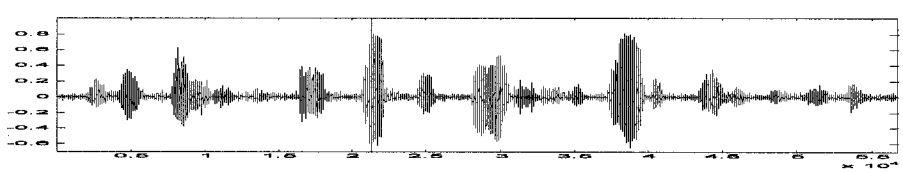
(c) Enhanced signal by standard ST



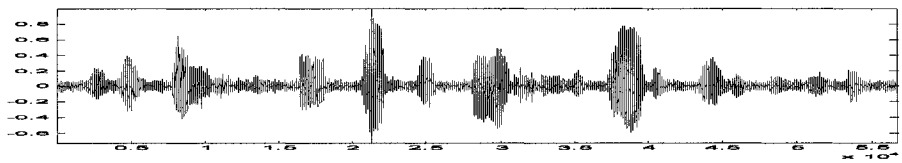
(d) Enhanced signal by SureShrink



(e) Enhanced signal by QBNE thresholding



(f) Enhanced signal by PSNR thresholding



(g) Enhanced signal by proposed system

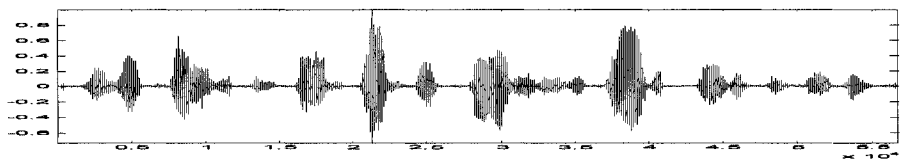
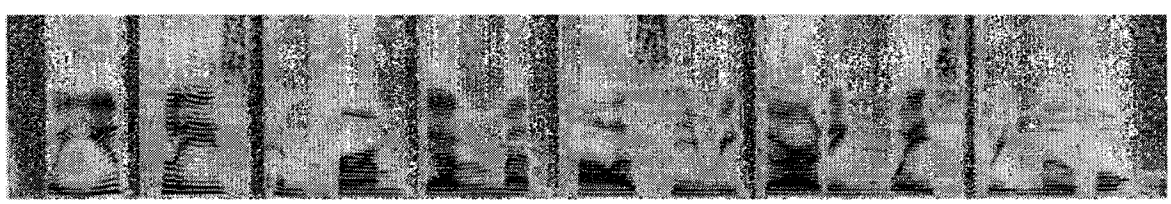


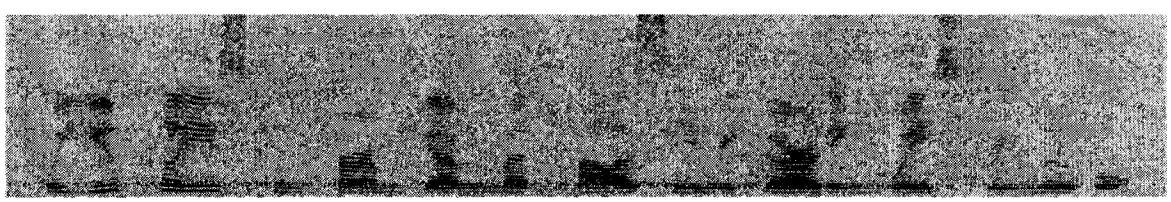
Figure 5-13 Waveform comparison with GWN (SNRinput=5dB)

The spectrogram comparison is shown in Figure 5-14. Compared with (b) Corrupted signal, the proposed system (g) yielded a good de-noising result, with cleaner background than others. Although some speech distortion was introduced into the output signal, most major components were saved. Thus, the subjective evaluation of the speech distortion, which will be discussed later, is close to that from the SureShrink processing.

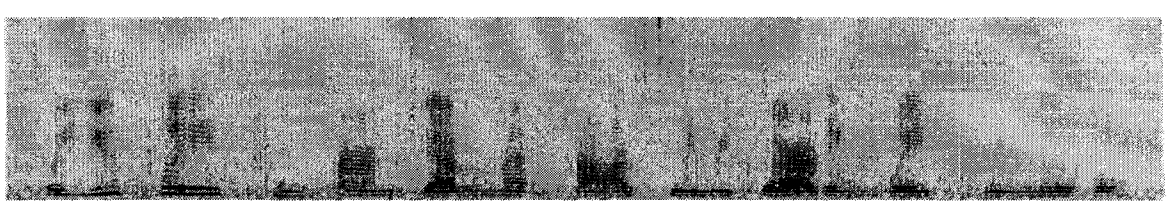
(a) Clean Speech (sx366)



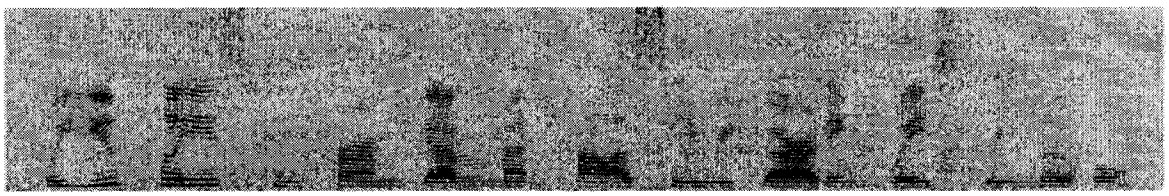
(b) Noisy Speech (White Noise, SNR=5dB)



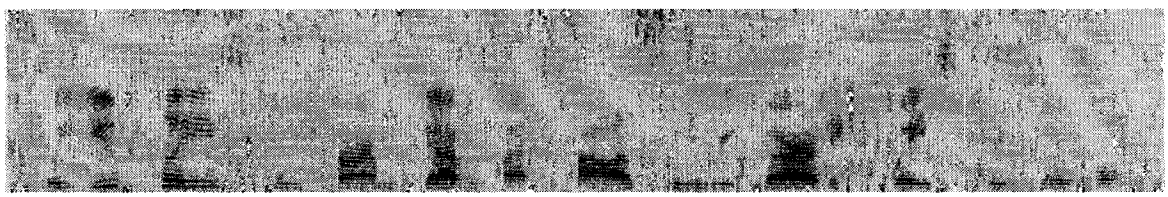
(c) Enhanced Speech (Standard Soft T)



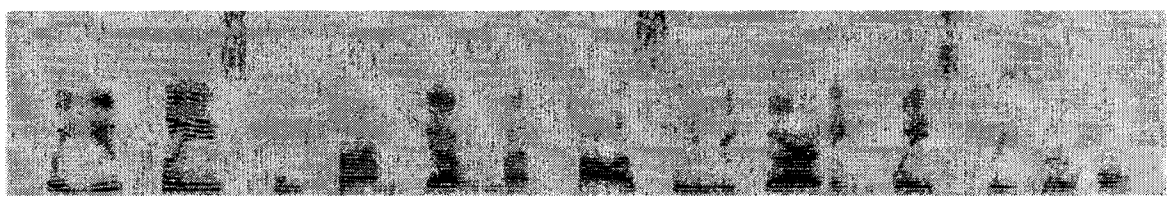
(d) Enhanced Speech (SURE)



(e) Enhanced Speech (QBNE)



(f) Enhanced Speech (PSNR)



(g) Enhanced Speech (Proposed System)

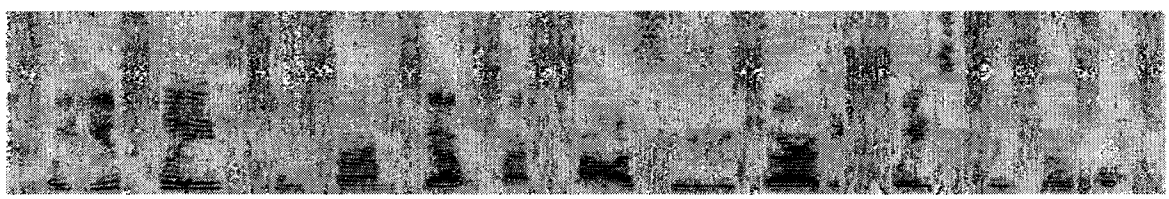


Figure 5-14 Spectrogram comparison with GWN (SNRinput=5dB)

The global SNR values for sentence sx366, corrupted by the WGN at a variety of SNRinput conditions from 0 dB to 15 dB, enhanced by these five methods, are illustrated in Figure 5-15. Curve yielded by optimized proposed system is staying higher than other curves when the SNRinput is at the range from 0dB to 15dB. One exception is the point of 0dB, but the output SNR of the proposed system is still very close to the best one, QBNE thresholding. When the SNRinput becomes better, the curves of PSNR thresholding and SureShrink run closer. The simulation results indicate that the proposed system is more effective than the other methods when the input SNR is lower than 15dB. It can be observed that when the input SNR is low, like lower than 5dB, the optimized

QBNE works effectively. But when the input SNR goes higher, especially when higher than 8dB, this method may result in a quite low output SNR.

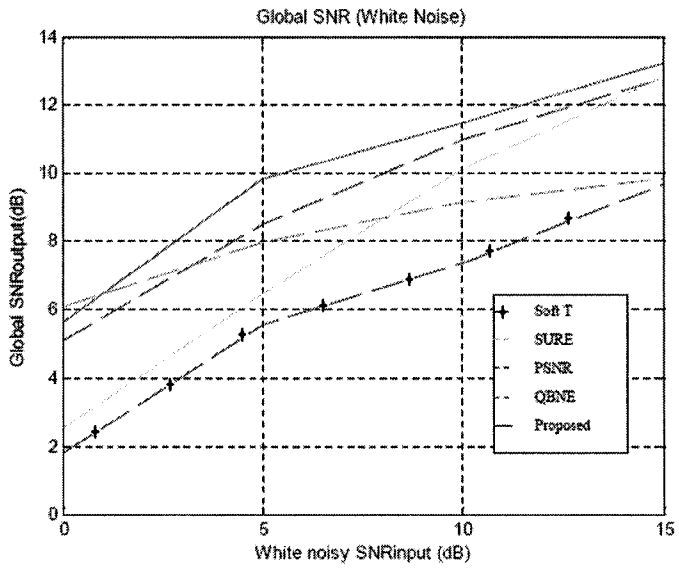


Figure 5-15 Global SNR output with GWN

Figure 5-16 illustrates the segmental SNR of the GWN noisy signal with SNRinput from 0dB to 15dB. Obviously, the proposed system works best and yields the highest line.

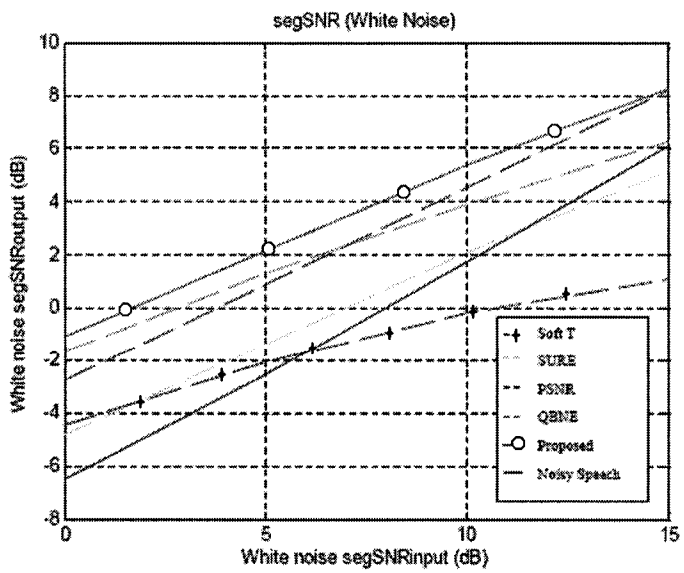
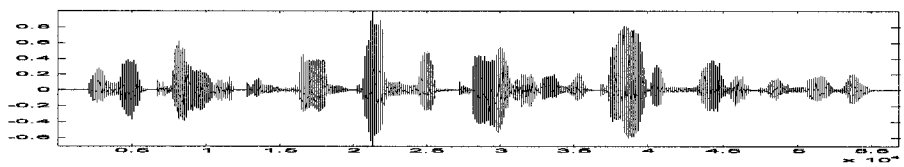


Figure 5-16 segSNR output with GWN

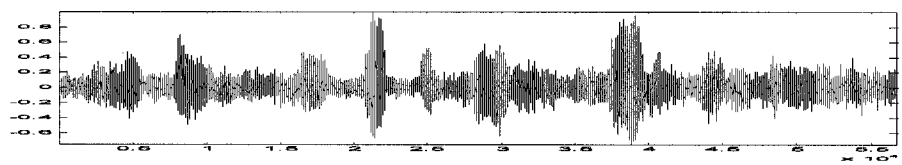
ii) Non-stationary noise environment

As introduced above, non-stationary noise shows a slow varying local SNR and frequency composition. The non-stationary noise presented in this section is Babble. Figure 5-17 illustrates the waveforms of the signals before and after the enhancement. Similar to the results of WGN de-noising, (g) Enhanced signal by the proposed system has a much stronger effect of removing background noises than other method yield. In the mean time it performances well in preserving the speech components.

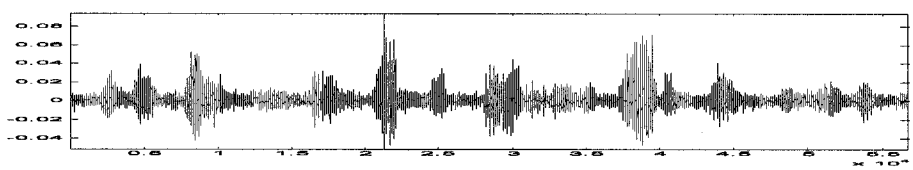
(a) Clean signal of sx366



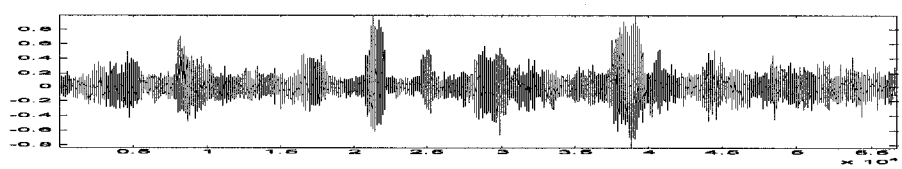
(b) Noisy Speech (Babble, SNR=0dB)



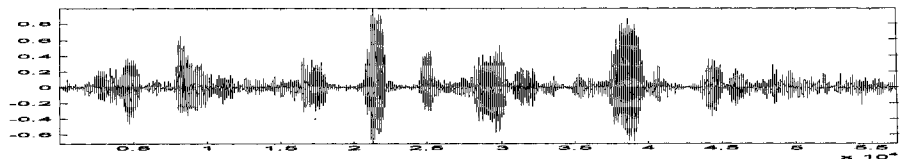
(c) Enhanced Speech (Standard Soft T)



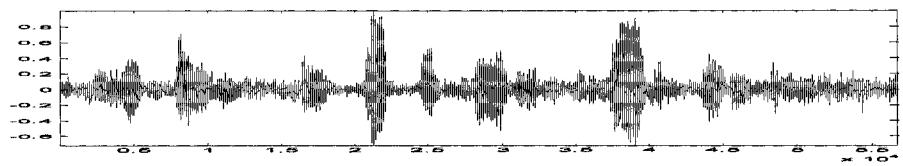
(d) Enhanced Speech (SURE)



(e) Enhanced Speech (QBNE)



(f) Enhanced Speech (PSNR)



(g) Enhanced Speech (Proposed System)

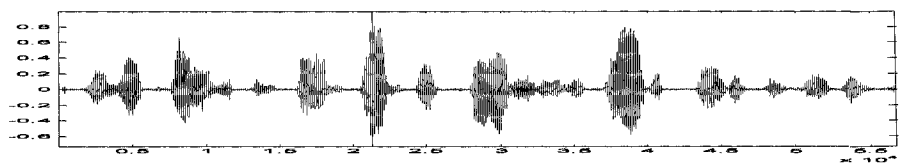
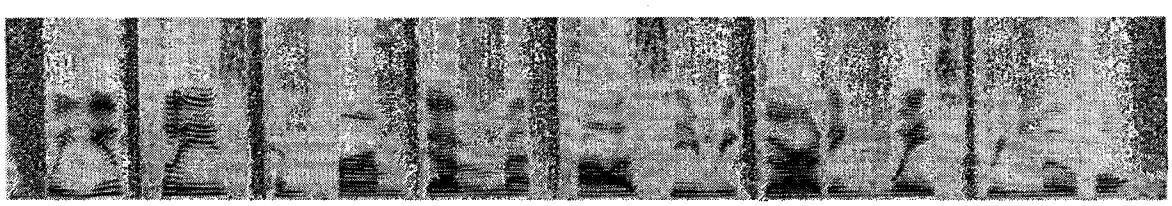


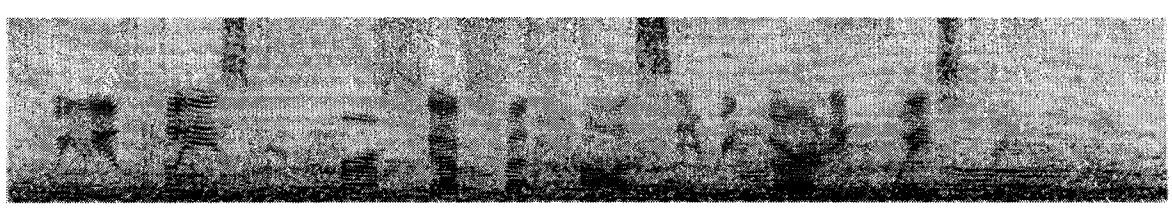
Figure 5-17 Waveform comparison with non-stationary noise (SNRinput=0dB)

The spectrograms in Figure 5-18 indicate that the proposed system works well in the non-stationary environment, suppressing the background noise effectively, and recovering most important speech components which were not overwhelmed by the noise. The low frequency noise is left in (c). It proves the shortcoming of the standard soft thresholding. The result yielded from SureShrink algorithm has a blur speech part surrounded by the noises, which is not a satisfied enhancement effect.

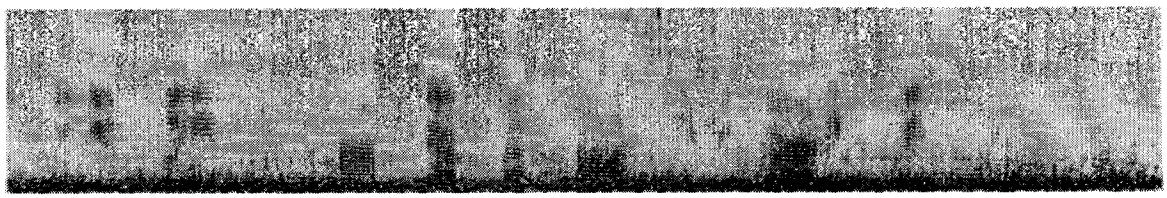
(a) Clean signal of sx366



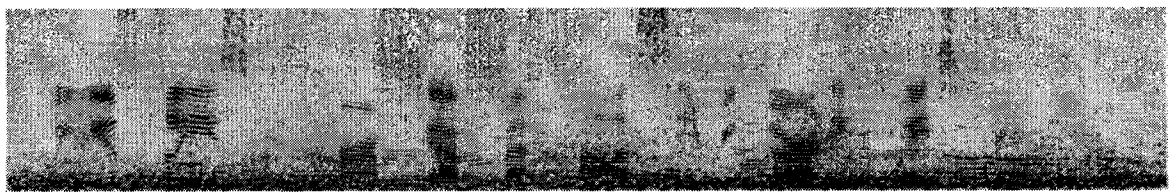
(b) Noisy Speech (Babble, SNR=0dB)



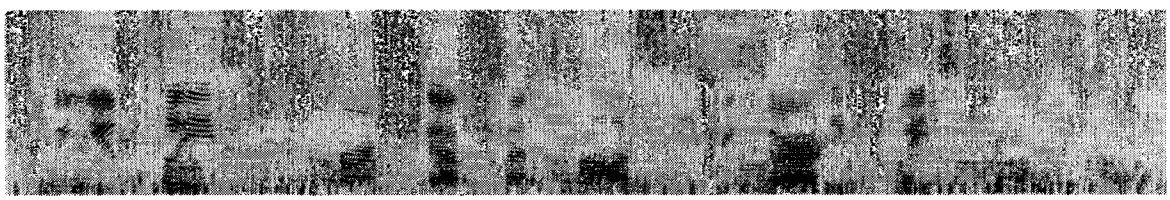
(c) Enhanced Speech (Standard Soft T)



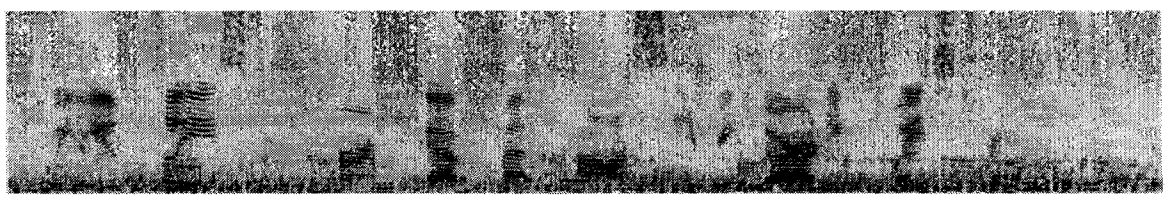
(d) Enhanced Speech (SURE)



(e) Enhanced Speech (QBNE)



(f) Enhanced Speech (PSNR)



(g) Enhanced Speech (Proposed System)

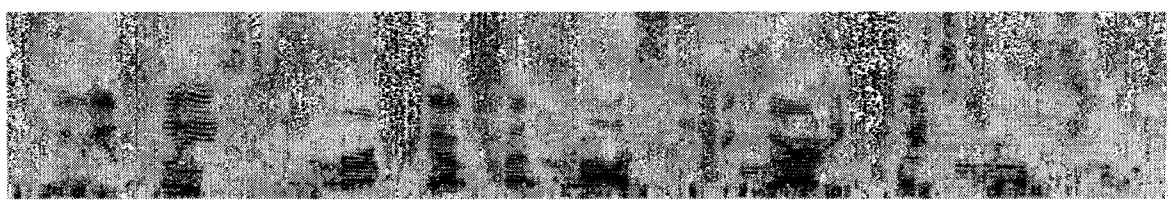


Figure 5-18 Spectrogram comparison with non-stationary noise (SNRinput=0dB)

The global SNR output (Figure 5-19) and segSNR output (Figure 5-20) with non-stationary noise also have the same tendency as those in the case of GWN. The proposed system takes the advantage over the traditional standard soft thresholding and SureShrink at the range of 0dB to 15dB (SNRinput).

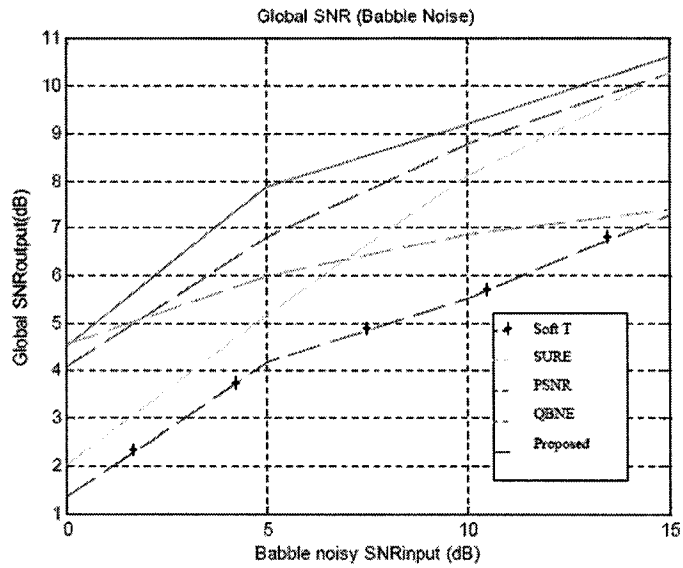


Figure 5-19 Global SNR output with non-stationary noise

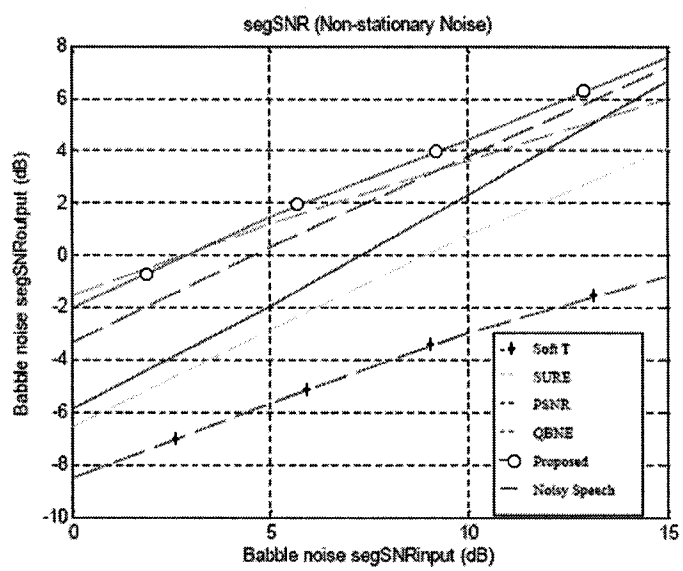


Figure 5-20 segSNR output with non-stationary noise

Other than objective measures, subjective evaluations derived from ITU-T recommendation P.835 were also adopted in this section. Table 5-6 shows the subjective evaluation scores of the sentence sx366 corrupted by the WGN and babble noise at 0dB, 5dB, and 10dB SNRinput respectively. The optimized system has the highest scores in

both noise environments with different SNRinput. It proves that the proposed system improves the perceptual speech enhancement evidently, comparing to the SureShrink.

Table 5-8 Subjective evaluation of SureShrink and proposed System

Speech	SNR_{input}	Noise Type	Noisy Speech			SureShrink			Proposed System		
			SIG	BAK	ORL	SIG	BAK	ORL	SIG	BAK	ORL
Sx366	0dB	WGN	4	1	2.572	4	1	2.572	4	2	2.938
		Babble	4	1	2.572	4	1	2.572	4	2	2.938
	5dB	WGN	4	1	2.572	4	2	2.938	4	3	3.304
		Babble	4	1	2.572	4	2	2.938	4	3	3.304
	10 dB	WGN	4	2	2.938	4	3	3.304	4	4	3.67
		Babble	4	2	2.938	4	3	3.304	4	4	3.67

CHAPTER VI

6. CONCLUSION AND FUTURE WORK

The purpose of this research is to summary and compare the latest wavelet de-noising algorithms, and to propose an optimal wavelet speech enhancement method. Chapter I and II introduced the basic research background and the problem, followed by Chapter III and IV discussing the principles of five perceptual and adaptive methods, including four musical noise suppression methods as well. Three parts of improvement were proposed. A novel adaptive wavelet speech enhancement system was also introduced. In Chapter V the Matlab simulation was determined to produce the best results of all the methods discussed in this thesis. The standard speech and noise database, TIMIT and Noise92 were selected as the signal sources. Both stationary and non-stationary noise environments were considered. The comparison consisted of subjective evaluation and several objective evaluation methods. Since it is highly correlated to subjective evaluation result, the spectrogram had been used as a major evaluation tool.

After large amounts of simulation and comparison, the advantages and disadvantages of these methods have been presented.

- **Traditional Wavelet De-noising:** To evaluate the standard methods, the simulation results of Standard Soft Thresholding and SureShrink were compared first. It is clear that, the standard soft Thresholding removes a lot of noise in both white noise and non-stationary noise environment, however, distorts the speech components badly in the whole spectrum due to the using of universal threshold algorithm. In addition, since the thresholding is only performed at the detail parts on the decomposition tree, the low frequency noise is left in the output signal. SureShrink is adaptive to signal, and provides smoother thresholding results. In other word, the speech signal is preserved better. Although SureShrink is not powerful enough for noise removing, it

shows a better performance in subjective evaluation. Thus, this method was picked as the benchmark of the following comparison.

- **Perceptual Wavelet Thresholding (PWT)**: A perceptual time-frequency analysis model were designed to map the critical bands. This model has approximate center frequency and bandwidth values to the parameters of auditory filter banks. The noise suppression in wavelet domain depends on a voice activity detector (VAD). The accuracy of the VAD has a great impact on the performance of this method. Because of the inaccuracy of the VAD, this method has led to an unsatisfied speech enhancement. Another noticeable disadvantage of this method is complicated calculation.

- **Adaptive Thresholding:** Within this group, four algorithms were discussed. The SHTAT (Smoothed Hard Thresholding with Aggravated Threshold) is a method depending on VAD as well. It has been proven that the aggravating standard threshold is not proper since the speech parts have been damaged too much.

  TEO is a good indicator of speech activity, useful for distinguishing the noise and speech components. However, the current algorithm of TEOAT is not accurate enough and shows an unstable effect in different cases.

  QBNE (Quantile-based Time-frequency Noise Estimate) performances well when the input SNR is lower than 5 dB, but worse when the background noise is weak. It reduces the residual noise evidently, however, over suppresses speech components due to universal statistic algorithm.

  PSNRAT(Posteriori SNR Time-Adaptive Thresholding) performances well in tracking the local SNR. It yields more residual noise than QBNE does, but also reduces the speech distortion most effectively.

It has been proven that the adaptive threshold algorithms work better than the traditional wavelet de-noising algorithms, especially under low SNR situation.

- **Musical Noise Suppression:** The Floor Construction is simple but not effective. Adaptive Minimizing is time adaptive, but depending on the VAD, introducing errors in local noise estimate. Silent Segment Musical Noise Suppression was designed to remove all the noise in the silent segment, however, depending on the VAD. Thus, it takes a high risk of cutting off the speech frames and introducing heavier musical noise. The new adaptive algorithm of musical noise suppression based on the local energy analysis possesses the advantage of time-adaptive algorithm, describing the properties of the musical noise, has been proven effective when the input SNR is in the range of 0dB to15dB.

Based on the above work, a new speech enhancement system using adaptive wavelet de-noising was proposed. Each step of the standard wavelet thresholding was improved by optimized adaptive algorithms. The Quantile based adaptive noise estimate and the posteriori SNR based threshold adjuster are compensatory to each other. The combination of them has achieved a very good tradeoff between noise suppression and speech reserving, in both stationary and non-stationary noise environments. Another contribution of this paper is introducing a successful innovative musical noise analysis and suppression algorithm. The TEO based silent segments smoothing has also been demonstrated to increase the perceptual quality of the output speech. The experimental results demonstrated the capability of the proposed system in both stationary and non-stationary noise environments.

For the future work, TEO is a simple but effective technology worth for further study too. The idea of local energy analysis could be used to distinguish the speech signal and noise signal. In addition, the performance of the quantile-based time-frequency noise estimate in a very low SNR environment is also impressive. We can infer that this statistical algorithm could be modified to meet the request of higher SNR environment. An adaptive

quantile-based noise estimator can probably help us to achieve more accurate noise suppression.

One limit of the wavelet thresholding is that it cannot be exactly accurate. It assumes that the noise coefficients have smaller abstract values than the speech coefficients, but it may not be the truth in real life, especially in a low SNR environment. Extracting a satisfying estimate from the corrupted signal is always very hard. Therefore, incorrect estimate leads to improper results. If the goal of the speech processing is much higher perceptual quality, more aspects of human speech characteristics have to be considered, and more complicated speech models should be built up to recover the speech from the noise mixed signal.

REFERENCES

[Bahoura2001] Mohammed Bahoura and Jean Rouat, "Wavelet speech enhancement based on the teager energy operator", *IEEE Signal Processing Letters*, vol. 8, No. 1, January, 2001

[Bai2003] Houwu Bai Eric A. Wan," Two-pass quantile based noise spectrum estimation", *Center of Spoken Language Understanding, OGI School of Science and Engineering at OHSU*, 2003

[Black1995] Mark Black, Mehmet Zeytinog'lu, "Computationally efficient wavelet packet coding of wide-band stereo audio signals", *Acoustics, Speech, and Signal Processing*, ICASSP-95. 1995.

[Boll1979] Steven F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.ASSP-27, No. 2, April 1979

[Cairns1996] D. A. Cairns and J. H. Hansen and J. F. Kaiser, "Recent advances in hypernasal speech detection using the nonlinear TEO", *Proc. {ICSLP} '96*, vol2, Philadelphia, PA, pp 780—783, 1996

[Carnero1999] Benito Carnero, and Andrzej Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms", *IEEE Transactions on Signal Processing*, Vol. 47, No. 6, June 1999

[Chen2004] Shi-huang Chen, Jhing-fa Wang, "Speech enhancement using perceptual wavelet packet decomposition and teager energy operator", *Journal of VLSI Signal Processing 36*, 125–139, 2004

[Donoho1995a] David L. Donoho, "De-noising by soft-thresholding", *IEEE Transactions on Information Theory*, vol. 41, No. 3, May, 1995

[Donoho1995b] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: Asymptopia?" J. *Roy. Srat. Soc., serB*, vol. 57, pp. 301-369, 1995.

[Donoho1995c] D.L. Donoho and I.M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage", *Journal of American Statistical Assoc.,* Vol. 90, no. 432, pp 1200-1224, Dec. 1995.

[Ephraim2003] Yariv Ephraim, Hanoch Lev-Ari and William J.J. Roberts, "A brief

survey of speech enhancement", *the electronic handbook, CRC Press*, 2003.

[Fu2003] Q. Fu and E. A. Wan, "A novel speech enhancement system based on wavelet denoising", *Technical Report, OGI, School of Science and Engineering, Oregon*, 2003.

[Ghanbari2005] Yasser Ghanbari, Mohammad Reza Karami-Mollaei,Esfandiar Zavarehei, and Mojtaba Lotfizad, "A New Approach for speech enhancement based on adaptive thresholding of wavelet packet", *ICASSP*, Paper #3097, 2005

[Gustafsson2001] Harald Gustafsson, Sven Nordholm and Ingvar Claesson, "Spectral subtraction using correct convolution and a spectrum dependent exponential averaging method", *IEEE Transactions on Speech and Audio Processing*, Vo9. pp. 799-807, Nov. 2001

[Gui2005] Yang Gui, "Speech enhancement using auditory filterbank" , *Master Thesis, Department of ECE, University of Windsor*, 2005

[Hu2004] Yi Hu and Philipos C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum", *IEEE Transactions on Speech and Audio Processing*, Vol. 12, No. 1, January 2004

[Hu2006] Yi Hu and Philipos C. Loizou, "Evaluation of objective measures for speech enhancement" *Department of Electrical Engineering University of Texas at Dallas Richardson, TX, USA*.2006

[Jabloun1999] Firas. Jabloun, A. Enis. Cetin, "The Teager energy based feature parameters for robust speech recognition in car noise", *ICASSP1999*, vol.1, pp 273-276, 1999.

[Johnstone1997] Iain M. Johnstone and Bernard W. Silverman," Wavelet threshold estimators for data with correlated noise", *J.Roy. Stat. Soc.* B (59), pp.319-351, 1997.

[J.D.Johnston1988] J. D. Johnston, "Transform Coding of Audio Signals Using the Perceptual Noise Criteria," *IEEE J. Selected Areas in Comm.*, vol. 6, pp. 314–323, Feb. 1988.

[Kaiser1990] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," *Proc IEEE Int. Conf. Acoust. Speech, Signal Processing, Albuquerque, NM*, (1990), pp. 381-384.

[Kaiser1993] J. F. Kaiser, "Some useful properties of Teager's energy operators", *in Proc. IEEE ICASSP, Minneapolis, Minnesota*,. vol.3, April 1993, pp.149-152.

[Lee2004] S. W. Lee, P. C. Ching, and Tan Lee, "Noise-robust automatic speech recognition using mainlobe-resilient time-frequency quantile-based noise", *ISCAS,* Vol 3, Page(s): III - 425-8, 2004

[Lei2005] Sheau-Fang Lei and Ying-Kai Tung, "Speech enhancement for nonstationary noises by wavelet packet transform and adaptive noise estimation", *Intelligent Signal Processing and Communication Systems,* 2005

[Lin2003] L. Lin, W.H. Holmes and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement", *Electronics Letters 1st*, May, 2003 Vol. 39 No. 9

[Manikandan2006] S. Manikandan, "Speech enhancement based on wavelet denoising", *Academic Open Internet Journal*, Vol 17, 2006

[Martin1993] Rainer Martin, "An efficient algorithm to estimate the instantaneous SNR of speech signal", *EuroSpeech'93,* pp1093-1096, 1993

[Martin1994] Rainer Martin, "Spectral subtraction based on minimum statistics", *Eur. Signal Processing Conf.*, pp1182-1185, 1994.

[Moore1996] B. C. J. Moore, "Masking in the Human Auditory System," in *Collected Papers on Digital Audio Bit-Rate Reduction* (N. Gilchrist and C. Grewin, eds.), pp. 9–19, Audio Engineering Society, 1996.

[Pinte'r1996] Istva'n Pinte'r, "Perceptual wavelet-representation of speech signals and its application to speech enhancement", *Computer Speech and Language* (1996) 10, 1–22

[Quackenbush1988] S. R. Quackenbush, T. P. Barnwell III, M. A. Clements, "Objective Measures of Speech Quality" , *Prentice Hall*, 1988.

[Ris2001] Christophe Ris, Stephane Dupont, "Assessing local noise level estimation methods: application to noise robust ASR", *Speech Communication*, v34, i1-2, pp141-158, Apr. 2001.

[Shao2005] Yu Shao and Chip-Hong Chang,Center for Integrated Circuits and Systems, "A versatile speech enhancement system based on perceptual wavelet denoising", *Circuits and Systems*, 2005. *IEEE International Symposium on Circuits and Systems,* Vol. 2, pp. 864 –867, 2005.

[Stahl2000] Volker Stahl, Alexander Fischer and Rolf Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering", *Acoustics, Speech, and Signal*

Processing, 2000. ICASSP '00. Vol 3, 5-9 June 2000 Page(s):1875 - 1878 vol.3

[Stein1981] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.,* vol. 9, pp. 1135–1151, 1981.

[Teager1990] H.M. Teager and S. M. Teager, "Evidence for nonlinear production mechanisms in the vocal tract", *in Speech Production and Speech Modeling*, pp. 241-261, 1990

[Terhardt1982] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for Extraction of Pitch and Pitch Salience from Complex Tonal Signals," *J. Acoust. Soc. Am.*, vol. 71, pp. 679–688, Mar. 1982.

[Virag1999] Nathalie Virag, "Single channel speech enhancement based on masking properties of the human auditory system", *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 2, March 1999

[Web1] http://www.mydr.com.au/default.asp?article=3434

[Web2] http://www.mathworks.com/access/helpdesk/help/toolbox/wavelet/index.html?/access/helpdesk/help/toolbox/wavelet/ch01_i10.html&http://www.mathworks.com/cgi-bin/texis/webinator/search?pr=Whole_site&db=MSS&prox=page&rorder=750&rprox=750&rdfreq=500&rwfreq=500&rlead=250&sufs=0&order=r&whole=Whole_site&entire_flag=1&is_summary_on=1&ResultCount=10&query=wavelet+transform

[Web3] http://www.mathworks.com/access/helpdesk/help/toolbox/wavelet/index.html?/access/helpdesk/help/toolbox/wavelet/ch01_i10.html&http://www.mathworks.com/cgi-bin/texis/webinator/search?pr=Whole_site&db=MSS&prox=page&rorder=750&rprox=750&rdfreq=500&rwfreq=500&rlead=250&sufs=0&order=r&whole=Whole_site&entire_flag=1&is_summary_on=1&ResultCount=10&query=wavelet+transform

[Web4] http://en.wikipedia.org/wiki/Discrete_wavelet_transform

[Web5] http://en.wikipedia.org/wiki/Wavelet_packet_decomposition

[Web 6] http://wiki.hydrogenaudio.org/index.php?title=ATH

[Web 7] http://www.mathworks.com/products/wavelet/demos.html?file=/products/demos/shipping/wavelet/denoisingsignalsdemo.html#1

[Zhang2003] Wei-Qiang Zhang and Guo-Xiang Song, "A translation=invariant wavelet de-noising method based on a new thresholding function", *Proceedings of the Second International Conference on Machine Learning and Cybernetics,* xi'an, 2-5 November 2003

[Zwicker1999] E. Zwicker and H. Fastl, Psychoacoustics: Facts and Models. *Springer-Verlag,* 1999.

VITA AUCTORIS

NAME

Lan Xu

PLACE OF BIRTH

TaiYuan, ShanXi, China

EDUCATION

M. A. Sc.

Department of Electrical Engineering

University of Windsor

Windsor, Ontario, Canada

2005-2007

B. Eng.

Department of Automation

University of Electronic Science and Technology of China

ChengDu, SiChuan, China

1994-1998