Speech enhancement in transform domain

Ding, Huijun

2011

Ding, H. (2011). Speech enhancement in transform domain. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/43536>

<https://doi.org/10.32657/10356/43536>

SPEECH ENHANCEMENT IN TRANSFORM DOMAIN

DING HUIJUN

School of Electrical & Electronic Engineering

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

2011

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me and has not been submitted for a higher degree to any other University or Institute.

.....

Date

.....

DING HUIJUN

*To Dad and Mom,
for their encouragement and love.*

Summary

This thesis focuses on the development of speech enhancement algorithms in the transform domain. The motivation of the work is first stated and various effects of noise on speech are discussed. Then the main objectives of this work are explained and the primary aim is to attenuate the noise component of a noisy speech in order to enhance its quality using transform based filtering algorithms. A literature review of various speech enhancement algorithms with an emphasis on those implemented in the transform domain is presented. Some of the important speech enhancement algorithms are outlined and various transform methods are compared and discussed.

Based on the discussion of transform methods, the use of Discrete Cosine Transform (DCT) is investigated. DCT has been proven to be a good approximation to Karhunen-Loeve Transform (KLT) and has similar properties to Discrete Fourier Transform (DFT). It also possess a better energy compaction capability which is advantageous for speech enhancement. However, frame to frame variations of DCT coefficients for a perfectly stationary signal can be observed. Therefore a DCT based adaptive time-shift analysis technique is introduced to overcome this problem. It reduces the drawbacks of the fixed window-shift and the amount of shift in the analysis window is now based on the pitch period, thus increasing the inter-frame similarities. Furthermore, a Wiener filter using the a-priori signal to noise ratio (SNR) with an adaptive parameter is also derived and implemented.

For DFT based speech enhancement, most algorithms using spectral filtering will result in residual noise which is musical in nature and is very annoying to human listeners. Besides, many speech enhancement approaches assume that the transform

coefficients are independent of one another and can thus be attenuated separately, thereby ignoring the correlations that exist between different time frames and within each frame. In order to exploit such correlations between the different time frames and to further reduce residual noise, a single channel speech enhancement based on 2D consideration in frequency domain is investigated. Unlike other 2D speech enhancement techniques which apply a post-processor after some classical algorithms such as spectral subtraction, the proposed approach uses a hybrid Wiener spectrogram filter containing a 1D Wiener filter and a 2D Wiener filter for effective noise reduction, followed by a multi-blade post-processor which exploits the 2D features of the spectrogram to preserve the speech quality and to further reduce the residual noise.

Despite the quality improvement of the speech signal with most DFT based noise reduction algorithms, the output is always distorted to some extent due to the over-attenuation of speech components. Weak speech components are usually regarded as noise in the noise reduction processing and are therefore highly suppressed. A post-processing technique which is based on the regeneration of both the voiced and unvoiced speech in the entire frequency domain is thus proposed to reduce this problem. A non-linear transform is first applied to obtain the excitation signal, and a smooth envelop is then estimated. To utilize the information of the clean speech contained in the envelop, the Wiener filtered output is combined with a weighted product of the excitation signal and the estimated envelop to generate the final synthesized speech which has less distortions on the speech signal compared to the Wiener filtered one. The synthesized speech is quite close to the clean speech and is more natural-sounding. Moreover, this algorithm can mask the residual musical noise effectively with the regenerated speech components.

The enhanced speech needs to be evaluated either by subjective measure or

objective measure. Objective measure is preferred since it is convenient and time-saving. Among all the existing objective measures, none is able to give a specific measure on speech distortion or noise reduction although speech distortion and noise reduction are two key metrics to evaluate the resultant speech quality. Two novel objective measures are thus proposed to evaluate the performance on speech distortion and noise reduction. They are calculated based on the residual signal which is the difference between the original clean speech and the processed speech.

Finally, all the proposed algorithms are compared in terms of their computational cost, inherent delay times and their output speech qualities. Five female and five male utterances are used as the test set. They are corrupted using four different noise types. Several objective measures, including segmental SNR, perceptual evaluation of speech quality (PESQ) and composite measures are performed on the test set. The strengths and weaknesses of the various proposed algorithms are analyzed and the effectiveness and significance of the proposed methods have also been verified. Based on that, the conclusion and some recommendations for future work are also presented.

Acknowledgements

In the first place I would like to express my most sincere gratitude to my supervisor, Associate Professor Soon Ing Yann, for all these years of conscientious supervision, sound advices, invaluable guidance, constant encouragement and support. He is an oasis of ideas and his passion in science has inspired me and enriched my training as a student and a researcher.

I gratefully acknowledge Professor Koh Soo Ngee and Associate Professor Yeo Chai Kiat for their advice and the enlightening discussion to improve my work. Their thoughtful comments and encouragement are highly appreciated.

I am deeply grateful to my family for their support and unfailing blessings. My mum, Ms. Lu Yun, my dad, Mr. Ding Yu and my husband, Mr. Liu Ming. Their endless love, support, encouragement and care are a strong motivation for me.

I would also like to express my special thanks to Jin Feng, Shen Minmin, Luo Xuewen and Qiu Mengran for their friendship and countless help rendered during my graduate study. My memorable days of life in NTU can never be complete without them.

Last but not least, I would like to thank all those who have helped me a lot in the past, especially those who have helped to do the listening tests.

Contents

Abstract	i
Acknowledgements	iv
List of Figures	ix
List of Tables	xi
List of Abbreviations	xii
List of Symbol	xiv
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	5
1.3 Contributions	7
1.4 Organization of this Thesis	8
2 Common Speech Enhancement Methods	9
2.1 Noise Model and Estimation	10
2.2 Time Domain Methods	13
2.2.1 Comb Filtering	13
2.2.2 Wiener Filtering and Linear Prediction	15
2.2.3 Kalman Filtering	16
2.2.4 HMM Filtering and Neural Networks	17
2.3 Transform Domain Methods	20

2.3.1	Spectral Filtering	20
2.3.2	Discrete Cosine Transform based Filtering	27
2.3.3	Subspace Filtering	28
3	DCT Based Speech Enhancement Filtering	30
3.1	Introduction	30
3.2	Motivation	34
3.3	Structure of Adaptive Time-shift Analysis	36
3.4	Windowing Function	37
3.5	Pitch Synchronization	39
3.6	Pitch Synchronization with Maximum Alignment	41
3.7	Wiener Filter with Adaptive Controller	45
3.8	Results and Discussion	48
3.9	Conclusion	52
4	Hybrid Wiener Filter for Spectral Filtering	55
4.1	Introduction	56
4.2	Structure of Proposed System	57
4.3	1D & 2D Hybrid Wiener Filter	59
4.3.1	1D Wiener and Two-step Noise Reduction	59
4.3.2	2D Wiener Filter	60
4.3.3	Hybrid Wiener Filter	61
4.4	Classification of Speech and Non-speech Components	62
4.5	Multi-Blade method	63
4.6	Results and Discussion	65
4.6.1	Comparison of Spectrograms	66
4.6.2	Comparison of Segmental SNR Results	67
4.6.3	Comparison of PESQ Scores	69
4.6.4	Comparison of Subjective Measure Results	70
4.7	Conclusion	72
5	A Spectral Post-processing Technique	74

5.1	Introduction	74
5.2	Structure of Proposed System	76
5.3	Traditional Noise Reduction Filter	77
5.4	Excitation Signal	80
5.5	Spectral Envelop Estimation	82
5.6	OSCR Filtered Speech	85
5.7	Results and Discussion	86
5.7.1	Voiced and Unvoiced Components Regeneration	87
5.7.2	Overall Speech Quality	93
5.8	Conclusion	95
6	Speech and Noise Distortion Measures	96
6.1	Introduction	96
6.2	Analysis for Speech Distortion	98
6.3	Analysis for Noise Distortion	99
6.4	Results and Discussion	101
6.4.1	DCT based Speech Enhancement Filtering	101
6.4.2	Hybrid Wiener Spectrogram Filtering	104
6.4.3	Spectral Post-processing	107
6.5	Conclusion	110
7	Evaluation of Speech Enhancement Algorithms	111
7.1	Introduction	111
7.2	Computational Complexity	112
7.3	Time Delay	114
7.4	Clean Speech and Noise Type	116
7.5	Performance Comparison	117
7.5.1	Objective Measures	117
7.5.2	Results Discussion	120
7.6	Conclusion	126
8	Recommendations and Conclusion	127

8.1 Conclusion	127
8.2 Recommendations for Further Research	131

Author's Publications	133
------------------------------	------------

Bibliography	134
---------------------	------------

List of Figures

2.1	Block diagram of VAD detector	12
2.2	Block diagram of comb filtering	14
2.3	Speech waveform and comb filter with constant pitch period T	14
2.4	Block diagram of the statistical filtering problem	15
2.5	Block diagram of HMM based MMSE estimator	18
2.6	The three layered neural network	19
2.7	Structure of transform domain based speech enhancement	20
3.1	Comparison of energy compaction between DCT and DFT	31
3.2	Upper limits of enhancement using DCT and DFT	32
3.3	Respective results of DFT and DCT by placing the analysis window in different positions	34
3.4	The block diagram of proposed adaptive time-shift analysis system . .	36
3.5	Magnitude of a rectangular window of length 100 with 1024-point transform	37
3.6	Basis functions for DCT-II, $N = 8$	42
3.7	Comparison of energy compaction among three different positions of starting point	43
3.8	5 impulses with pitch period	44
3.9	Segmental SNR results of noisy speech, Wiener filtered speech with rectangular window and Wiener filtered speech with Hann window . .	48
3.10	PESQ results of noisy speech, Wiener filtered speech with rectangular window and Wiener filtered speech with Hann window	49
4.1	Block diagram of the speech enhancement system in DFT domain . .	58
4.2	Sixteen blades of the different directions	64

4.3	Spectrograms comparison of clean speech and noisy speech with white noise at SNR=5dB	67
4.4	Resulting spectrograms comparison of three algorithms	68
5.1	Block diagram of the proposed speech enhancement system	77
5.2	Voiced spectrum of (a) Clean speech (b) Wiener filtered speech (c) SS filtered speech (the latter two are with input white noise at 5dB); Spectral Envelop of (d) Clean speech smoothed by $H(\omega)$ (e) Wiener filtered speech smoothed by $H(\omega)$ (f) Proposed method $C(m, \omega)$. . .	83
5.3	Unvoiced spectrum of (a) Clean speech (b) Wiener filtered speech (c) SS filtered speech (the latter two are with input white noise at 5dB); Spectral Envelop of (d) Clean speech smoothed by $H(\omega)$ (e) Wiener filtered speech smoothed by $H(\omega)$ (f) Proposed method $C(m, \omega)$. . .	83
5.4	Correlation factors of <i>Noisy</i> , <i>SS</i> , <i>EnvN</i> and <i>EnvSS</i> under different noise types.	92
5.5	Overall PESQ results comparison	93
6.1	Spectral energies of pure noise from the different speech enhancement algorithms: WF, PSWF, PSMax and ATSA	103
6.2	Spectral energies of pure noise from the different speech enhancement algorithms: 2DFT, GSSB and HWSF	105
6.3	Spectral energies of pure noise from the different speech enhancement algorithms: WF, PHFR, HRNR and OSCF	108
7.1	Block diagram of the PESQ measure	119
7.2	Comparison results of the SegSNR measure	121
7.3	Comparison results of the PESQ measure	122
7.4	Comparison results of the SIG measure	123
7.5	Comparison results of the BAK measure	124
7.6	Comparison results of the OVL measure	125

List of Tables

3.1	Comparison Results of SegSNR and PESQ measures	51
3.2	Comparison of Segmental SNR Results	52
3.3	Comparison of PESQ Scores	53
4.1	SegSNR Results with White Noise	69
4.2	SegSNR Results with Car Noise	69
4.3	SegSNR Results with F16 Aircraft Noise	70
4.4	PESQ Results with White Noise	70
4.5	PESQ Results with Car Noise	71
4.6	PESQ Results with F16 Aircraft Noise	71
4.7	Results of Subjective Measure and Objective Measure with White Noise at SNR=2dB	72
5.1	Mean log-spectral distance comparison for unvoiced parts	89
5.2	Mean log-spectral distance comparison for voiced parts	90
5.3	Overall mean log-spectral distance comparison	91
6.1	Normalized cross-correlation λ between clean speech and residual noise	102
6.2	Normalized cross-correlation λ between clean speech and residual noise	104
6.3	Normalized cross-correlation λ between clean speech and residual noise	106
7.1	Comparison Results of Computational Time	113
7.2	Delay time for different frame size at sampling frequency of 8 kHz and 16 kHz	114
7.3	Delay time for different number of frames with different amount of overlap	115

List of Abbreviations

ABBREVIATIONS

1D
2D
2DFT
AR
ASR
ATSA
BAK
CD
DCT
DFT
DWT
FFT
FIR
GSSB
HMM
HRNR
HWSF
HWSF-C
HWSF-T
IIR
IS
KLT
LLR
LPC
LSD
MMSE
MOS
MSE

FULL EXPRESSIONS

One Dimensional
Two Dimensional
2D Fourier Transform Filtering
Autoregressive
Automatic Speech Recognition
Adaptive Time-shift Analysis
An objective measure for noise distortion
Cepstral Distance
Discrete Cosine Transform
Discrete Fourier Transform
Discrete Wavelet Transform
Fast Fourier Transform
Finite Impulse Response
General Spectral Subtraction with Blade Post-processor
Hidden Markov Model
Harmonic Regeneration Noise Reduction Filter
Hybrid Wiener Spectrogram Filtering
HWSF without statistical classification
HWSF without Two-Step Noise Reduction
Infinite Impulse Response
Itakura-Saito Distance
Karhunen-Loeve Transform
Log-Likelihood Ratio
Linear Predictive Coefficients
Log-spectral Distance
Minimum Mean Square Error
Mean Opinion Score
Mean Square Error

OSCR	Over-attenuated Speech Components Regeneration
OVL	An objective measure for signal distortion
PESQ	Perceptual Evaluation of Speech Quality
PHFR	Post-processing for High-frequency Regeneration
PSMax	PSWF with Maximum Alignment
PSWF	Pitch Synchronized Wiener Filtering
SegSNR	Segmental Signal to Noise Ratio
SIG	An objective measure for signal distortion
SNR	Signal to Noise Ratio
SS	Spectral Subtraction
STFT	Short-time Fourier Transform
SVD	Singular Value Decomposition
TNR	Traditional Noise Reduction Filters
TSNR	Two-Step Noise Reduction
VAD	Voice Activity Detection
WF	Wiener Filtering
WFHO	Wiener Filtering with Higher Fixed Overlap
WSS	Weighted-slope Spectral Distance

List of Symbols

SYMBOL	MEANING
$\overline{(\cdot)}$	Mean function
α	Parameter in decision-directed approach normally set to 0.98
$\alpha(m, k)$	Adaptive parameter α
$\delta(n)$	Delta function = $\begin{cases} 1, & n = 0 \\ 0, & \text{others} \end{cases}$
$\epsilon(t)$	Residual noise signal, difference between clean and enhanced speech
$\varepsilon(\omega)$	Spectral expression of residual noise signal $\epsilon(t)$
$\varepsilon_N(\omega)$	Spectral noise distortion
$\varepsilon_S(\omega)$	Spectral speech distortion
$\gamma(\omega)$	The a-posteriori signal to noise ratio at frequency ω
θ_Y	Phase information of noisy speech signal
λ	Degree of speech distortion
λ_m	Degree of speech distortion in each speech frame m
$\lambda_{NN}(\omega)$	Expectation of noise signal power spectrum
$\lambda_s(k)$	k th eigenvalue of the autocorrelation of a speech signal

$\lambda_{SS}(\omega)$	Expectation of clean speech power spectrum
$\lambda_{SY}(\omega)$	Cross-power spectrum of clean speech and noisy speech
$\lambda_{YY}(\omega)$	Expectation of noisy speech power spectrum
Λ_s	Eigenvalue matrix of the autocorrelation of a speech signal
Λ_μ	Diagonal matrix of Lagrange multipliers
$\mu(k)$	k th Lagrange multiplier
$\nu(m)$	A gain function
$\xi(\omega)$	The a-priori signal to noise ratio (SNR) at frequency ω
$\hat{\xi}(\omega)$	The a-priori SNR estimated by decision-directed approach
σ_n^2	Noise Variance
σ	The variance of Gaussian distribution
ω_s	Sampling frequency
$\text{diag}(\cdot)$	Diagonal matrix operator
$C(m, \omega)$	Estimated spectral envelop
e	Base of natural logarithm
$E(\cdot)$	Expectation Function
$E_{PureNoise}$	Pure noise contained in residual signal
$g(k)$	k th diagonal element of G
$g(t)$	Traditional noise reduction Filter in time domain
G_1	Full rank $M \times M$ matrix
G	$= \begin{bmatrix} G_1 & 0 \\ 0 & 0 \end{bmatrix}$
$G(\omega)$	Spectral expression of traditional noise reduction filter $g(t)$
H_0	Hypothesis of speech absence

H_1	Hypothesis of speech presence
$H(\omega)$	A low-pass finite impulse response (FIR) filter
$H(m)$	A low-pass filter implemented by a Gaussian mask
H_{sub}	Linear filter of subspace speech enhancement
$im(n)$	Impulse sequence
I_0	Zero order modified Bessel function
J_α	Error between estimated a-priori SNR $\hat{\xi}(m, k)$ and real one $\xi(m, k)$
$\max(\cdot)$	Maximum operator
$M(; ;)$	Confluent hypergeometric function
M_n	Average Magnitude Function at time n
$n(t)$	Noise signal at time t
$N(\omega)$	Noise signal at frequency ω
$N(k)$	k th transformed coefficient of noise signal
$N(m, k)$	Time-frequency labeled coefficient of noise speech signal
$R(m, \omega)$	Regenerated speech components
$R(n)$	Autocorrelation function
$\text{Re}(\cdot)$	Real-part operator
R_s	Autocorrelation matrix of clean speech signal
$R_{\hat{sim}}$	Discrete cross-correlation of signals $\hat{s}(n)$ and $im(n)$
$s(n)$	Speech signal sequence at time n
$\hat{s}(n)$	Estimated speech signal at time n
$s(t)$	Speech signal at time t
$\hat{s}(t)$	Estimated speech signal at time t
$\tilde{s}(t)$	Restored speech with OSCR in time domain

$S(\omega)$	Speech signal at frequency ω
$S(k)$	k th transformed coefficient of speech signal
$S(m, k)$	Time-frequency labeled coefficient of speech signal
$\hat{S}(m, k)$	Estimated time-frequency labeled coefficient of speech signal
$\tilde{S}(m, \omega)$	Synthesized speech with OSCR in frequency domain
$\hat{S}_{1D}(m, \omega)$	Estimated speech filtered by 1D Wiener filter with TSNR
$\hat{S}_{2D}(m, \omega)$	Estimated speech filtered by 2D Wiener filter
U	Orthogonal eigenvector matrix of a speech signal
$w(\cdot)$	Window function
$x(t)$	Excitation signal
$\hat{X}(m, \omega)$	Whitened excitation signal in frequency domain
$y(n)$	Noisy speech signal sequence at time n
$y(t)$	Noisy speech signal at time t
$Y(\omega)$	Noisy speech signal at frequency ω
$Y(k)$	k th transformed coefficient of noisy speech signal
$Y(m, k)$	Time-frequency labeled coefficient of noisy speech signal

Chapter 1

Introduction

1.1 Motivation

Speech has been the most important form of human communication. Once a child is born, he or she will be immediately exposed to human language and may start learning to speak as early as 12 months old. This process makes speaking the most natural and straightforward way for conveying information when compared to other forms of communication such as body languages, text messaging or letter writing. Based on that, lots of applications have been developed to help people physically separated to have conversations. For example, the telephone invention has revolutionized the development of simultaneous information exchange across distance. Speech processing is developing into an indispensable part of this modern world for numerous applications ranging from hearing aids to cellular radio telephony system. Owing to the imperfections in speech acquisition, transmission and storage, speech signal is often degraded severely. The contamination of the speech signal not only affects its audio quality but also precludes many further high level speech processing tasks such as speech signal coding and recognition [1–3]. Speech enhancement

algorithms can therefore be used to improve the quality of speech as a preprocessor in speech coding or recognition systems [4–6].

In this thesis, noise reduction is the main objective among the various fields of speech enhancement fields. It mainly focuses on reducing or suppressing the background noise to some degree while maintaining the embedded speech signal as much as possible. Thus the unwanted signal, noise, should be understood prior to designing enhancement algorithms. We are surrounded by noise wherever we go in daily life. When conversation is ongoing in an noisy environment, it is definitely an undelightful experience, not to mention recorded noisy speech broadcasts which result in listening fatigue and lower intelligibility. Noise appears in different forms in this real world. However, it still can be roughly characterized by two main features from an academic point of view [7].

- Noise can be stationary or non-stationary: Stationary noise means it remains unchanged over a long time period, such as the engine noise in the car or aircraft, the fan noise from PCs and air conditioning noise in the office. The non-stationary noises include telephone ringing noise, train noise and restaurant noise, i.e people talking at neighboring tables and waiters interacting and conversing with consumers. Clearly, it is a more difficult challenge to suppress noise that is non-stationary than its stationary counterpart.
- Noise spectrum can be narrow-band or wide-band: Narrow-band or wide-band means the noise energy occupies a different range in the frequency domain. For example, the energy of wind noise and car noise are concentrated in the low frequencies typically below 500 Hz, while the restaurant noise and aircraft noise occupy a wider frequency range.

All the examples illustrated above are the kind of additive noise. Actually there is

another special noise type among the non-stationary noise, which is reverberation. A reverberation is created when a sound is produced in an enclosed space causing a large number of echoes which are the delayed versions of the original sound because of the reflections from the walls. This is most noticeable when the sound source stops but the reflections continue, and then they decrease in amplitude until they can no longer be heard. Reverberation is thus a convolutional noise type but it is not the focus of this thesis.

Different types of noise require different solutions. A simpler noise corruption situation is considered where the background noise is assumed to be an additive and wide-sense stationary signal which is statistically independent of the clean speech. Single channel speech enhancement is really a tough topic since only a single source of noisy signal is available for processing. The number of microphones available can influence the performance of speech enhancement algorithms. Typically the larger the number of microphones, the easier the speech enhancement task. However, such multi-sensor algorithms are very often impossible to apply on commonly available speech source, e.g. telephony speech, recorded speech, broadcast speech, etc. Thus single channel speech enhancement problem has been one of the most active topics in speech processing for decades. It is a long standing and challenging problem and a great number of algorithms have been proposed. However, the results are still not ideal for now.

Generally speaking, speech enhancement has three major objectives:

- to improve perception for human listeners, i.e. to make the processed speech sounds better or clearer.
- to improve the performance of speech coders, i.e. to be robust when intense noise is present.

- to increase the accuracy of speech recognition, i.e. to reduce the error rates of automatic speech recognition (ASR) system operating under adverse situation.

The first objective is the most common focus in single channel speech enhancement as well as this thesis. The perception of a speech signal is usually measured in terms of its quality and intelligibility. The quality is a subjective measure which reflects individual preferences of listeners. Intelligibility is an objective measure which predicts how much information can be accurately identified by listeners. Ideally speech enhancement is supposed to improve both quality and intelligibility. However, most algorithms improve only the quality of speech in practice. There are two main reasons. Firstly, the two characteristics, quality and intelligibility are independent of each other. It is possible to reduce the background noise, but at the expense of introducing some speech distortion which in turn may impair speech intelligibility. In other words, one can never learn from the enhanced signal more than one can learn from the noisy signal about the clean signal based on the information theory [8]. On the other hand, a low quality speech may be high in intelligibility. This can be achieved by over-emphasizing the high frequencies of the noisy signal [9]. Hence the main challenge in designing effective speech enhancement algorithms is to suppress the noise without introducing any perceptible distortion in the signal. Secondly, intelligibility measurements are more complicated to be conceived than the quality measurements which commonly utilize the Mean Opinion Score (MOS). MOS is a subjective listening test in which listener rates the speech quality from 1 to 5. This can also be predicted by some objective measurements such as Perceptual Evaluation of Speech Quality (PESQ). Intelligibility measurements are conducted differently as it is focused on the accuracy of identified words. In this test, volunteers are asked to listen to various sentences or isolated words and write down the words they heard. Based on the percentage of correctly recognized words, an intelligibility

score is obtained.

A number of algorithms have been proposed in the literature for speech enhancement with additive noise model. These algorithms can be divided into two main classes:

- Time domain algorithms: filtering is performed directly on the time sequence. This class includes Linear Predictive Coding (LPC) based algorithms [10], Kalman filtering [11–13], and Hidden Markov Model (HMM) filtering [14–16], etc.
- Transform domain algorithms: filtering is performed on the transformed coefficients. This class includes Discrete Fourier Transform (DFT) based algorithms [17–20], Discrete Cosine Transform (DCT) based algorithms [21–23], Discrete Wavelet Transform (DWT) based algorithms [24–26], and subspace algorithms (or Karhunen-Loeve transform) [27–29].

The transform domain speech enhancement algorithms are more popular since the transform coefficients offer an easy and efficient way for separating noise from clean speech signal. Therefore the transform domain speech enhancement algorithms will be further examined in this thesis.

1.2 Objectives

In this thesis, single channel speech enhancement algorithms in the transform domain are studied. A large number of papers in the literature show that this issue is theoretically challenging and practically important. With only one microphone, the information available is much less than in the case of two or more microphones. Consequently, noise suppression has to be achieved through estimation of noise and

signal statistics, modeling of signal and noise and statistical parameters, etc. Transform domain is a more efficient way of representing the speech and noise signals and it also allows better estimation of speech statistics. Hence all the algorithms developed in this thesis are based in the transform domain.

As mentioned in Section 1.1, there are three main objectives of speech enhancement and the first one, perceptual improvement is the main focus of this research. A successful algorithm must not only be able to reduce the magnitudes of corrupting noise components, but also to preserve the original embedded speech quality as much as possible. However, common noise reduction techniques emphasize on suppressing the background noise with speech distortion as the tradeoff. One extreme example is, when noise energy is very much higher than speech energy, there may be little speech left after simply cutting down the noise using a noise reduction technique. Furthermore, how successful a speech enhancement algorithm is still an open question for discussion. All the traditional objective measures or subjective measures give one score to show how similar the resulting speech is to the clean speech. They never quantify how much noise is reduced and how badly the underlying speech is disturbed. This additional information will be able to direct the speech enhancement algorithms to pursue a good trade-off between noise reduction and speech distortion, so as to result in subjectively pleasant enhanced speech for individuals. Hence the main objectives of this thesis are:

- to propose new algorithms for noise reduction in transform domain
- to investigate novel algorithms which emphasize on preserving the original clean speech without degrading the performance of noise reduction filter
- to quantify two aspects, speech distortion and noise reduction of overall speech quality for evaluating how successful the speech enhancement algorithms are.

- to compare various proposed algorithms using traditional objective and subjective measures and to show the comparison results.

1.3 Contributions

The major contributions of this dissertation are summarized as follows:

- A new Discrete Cosine Transform (DCT) based enhancement method is proposed and implemented. Some differences and advantages of DCT as compared to Discrete Fourier Transform (DFT) are discussed to show the feasibility of utilizing the transform.
- A new spectral filtering method is proposed and implemented. Unlike traditional spectral filtering, this method incorporates image processing type of technique to work on the speech spectrogram to obtain a cleaner enhanced speech.
- A novel post-processor is proposed for over-attenuated speech components regeneration. This method attempts to solve one of the best known problems in traditional speech enhancement: the unvoiced portions are severely attenuated along with the noise. The proposed algorithm is a pioneering work which can restore both voiced and unvoiced embedded speech signal which are over-suppressed by noise reduction process.
- Two novel objective measures are proposed and implemented. They report the amount of speech distortion and noise reduction respectively, instead of an overall quality figure given by all traditional measures.
- Comparison of all the proposed techniques and some other current techniques

are performed using the various objective measures.

1.4 Organization of this Thesis

This thesis consists of eight chapters. In Chapter 2, both time and frequency domain algorithms for speech enhancement are reviewed. Discrete Fourier Transform (DFT) based algorithms will be emphasized since they are more popular. In Chapter 3, the Discrete Cosine Transform (DCT) based speech enhancement algorithm will be introduced. In Chapter 4, a new approach to the problem of speech enhancement in time-frequency domain will be presented which utilizes 2D techniques on the spectrogram. This approach is quite good at suppressing the residual noises. Chapter 5 will present a novel post-processor that can be used in conjunction with any noise reduction filters. This proposed scheme attempts to minimize the distortion of speech signal without compromising the effectiveness of the noise reduction process. In Chapter 6, two novel objective measures will be introduced for respective predictions of speech distortion and noise reduction. Detailed derivations and explanations are given in this chapter as well. In Chapter 7, the performance comparison of the proposed enhancement algorithms is given in terms of the various objective measures. In Chapter 8, conclusion and recommendations for future work are presented.

Chapter 2

Common Speech Enhancement Methods

In the earlier years from 1940s to 1970s, speech enhancement began to attract the interest of researchers. In 1947, infinite peak clipping was reported to be able to improve the intelligibility despite introducing some undesirable distortions [30]. Then single channel speech enhancement has received considerable attention since mid-1970s. The famous spectral subtraction algorithm was initially proposed by Schroeder [31] in 1965, followed by the investigation in the correlation domain of Weiss et al. [32], and became very popular after the work of Boll [17] in the Fourier transform domain. Since the input speech signal is randomly selected by the user and noise signal is naturally a random process, the enhancement problem is a statistical estimation problem based on the modeling of speech signals and noise. Work in this area was initiated by MacAulay and Malpass [19] in 1980 where a maximum-likelihood approach is proposed for estimating the clean speech, and was followed by the outstanding work by Ephraim and Malah [18] in 1984, which not only produced more noise reduction but also reduced the level of tonal noise as compared to spectral

subtraction. Much of the work with the Wiener algorithm was initiated in the speech enhancement field by Lim and Oppenheim [10, 33]. Besides, the survey paper [33] classified the different speech enhancement systems and compared these systems. Other survey papers are found in [14, 34].

Speech enhancement can be performed both in time domain and frequency domain. Time domain filters include those utilizing Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) filters, Linear Predictive Coefficients (LPC) and Hidden Markov Model (HMM), etc. Transform domain filters are those which calculate the transform coefficients first and followed by the enhancement process. In the end, inverse transform must be applied to obtain the final resulting speech. Before reviewing the various specific speech enhancement algorithms, it is crucial to understand the behavior of noise and its statistical properties.

2.1 Noise Model and Estimation

Additive noise model is used in most of speech enhancement algorithms. Based on this model, the noisy speech, $y(t)$, is the result of clean speech, $s(t)$, plus noise signal, $n(t)$, which is defined in the following equation:

$$y(t) = s(t) + n(t) \quad (2.1)$$

where t is the time index.

So far, most of speech enhancement algorithms assume that an estimate of the noise is available. Such an estimate is critical for the performance of speech enhancement algorithms since it is required in many algorithms, for example, to estimate the a-priori SNR for the Wiener filter and Minimum Mean Square Error algorithms, or to estimate the noise covariance matrix for the Kalman filtering or subspace al-

gorithms. The noise estimate does have an impact on the quality of the enhanced signal. If the noise estimate is too low, annoying residual noise will be heard. If the noise estimate is too high, speech will be distorted, possibly resulting in intelligibility loss.

The simplest approach will be to manually adjust until an optimal enhanced speech can be heard. However this is only suitable for white noise and is not applicable in general cases. An alternative will be to estimate and update the noise estimate during the silence or speech pause period using a Voice Activity Detection (VAD) algorithm. It should be noted that although this kind of approach may work satisfactorily under stationary noise, it will not work well in more realistic environments where the noise characteristics change constantly.

A number of VAD algorithms were proposed based on various types of features extracted from the speech. The earliest work was based on energy levels and zero-crossings [35, 36], the Itakura LPC spectral distance measure [37], the periodicity measure [38] and cepstral features [39]. Some of the VAD algorithms are used in commercial applications such as audio-conference. Sometimes up to 70% of frames are silence during a phone conversation and using a VAD algorithm can extend the battery life of cellular phones. A good survey on commercial applications can be found in [40]. All the VAD algorithms generally output a binary decision: the tested speech segment is voice activity ($VAD=1$) if the measured values exceed a predefined threshold, otherwise it is noise ($VAD=0$).

A specific VAD algorithm will be described which is a portion of the clipping autocorrelation pitch detector described in [41]. This algorithm works well in the noisy environment when the noise is stationary or near stationary. The block diagram of this simple energy discriminant algorithm is shown in Figure 2.1. In this figure, the external initiation can be selected manually. A certain time period, such

Figure 2.1: Block diagram of VAD detector

as 59 ms is chosen as noise only period for calculating the noise energy, so-called the silence level threshold. Then the energy of speech segment is computed using a 10ms rectangular window. If the energy is above threshold, the segment is considered to be speech, otherwise the segment is classified as silence. The segment energy at the time position of n is computed by an average magnitude function M_t which is defined as follow:

$$M_n = \sum_{m=n-N+1}^n |y(m)|w(n-m) \quad (2.2)$$

where N is the number of samples, n and m are the time sample indices, and $w(.)$ is a window function and it can be a simple rectangular window here. Instead of using the standard square of magnitude, the magnitude is used. This is due to the reason that the squaring operation is very sensitive to large signal amplitudes which may lead to over-emphasis on large variations in y . Therefore a simplified magnitude function is used to compute the energy.

Besides the VAD algorithms, there are also many algorithms proposed in the literature for noise estimations. For example, in the frequency domain, the signal is analyzed using short-time spectra computed from short overlapping analysis frames, typically 20-30 ms in length with 50%-75% overlap between adjacent frames. After

that, the noise variance can be obtained by three main methods:

- Minimal-tracking algorithms: the silent segments not only appear in silent period, but also occur during speech activity as short pauses. Thus this kind of algorithms track the minimum of noisy speech power spectrum within a finite window [42].
- Time-recursive averaging algorithms: the power of noisy speech in individual frequency bands often decays to the power level of the noise even during speech activity. Thus this kind of algorithms track the minimum of noisy speech spectrum in individual frequency bands [43, 44].
- Histogram-based algorithms: histogram of energy values in individual frequency bands reveal that the most frequent value corresponds to the noise level of the specified frequency band. Thus this kind of algorithms estimate the noise based on the histogram of past power spectrum [45].

Apart from these algorithms, several other algorithms were proposed which were not included in these three categories. Since noise estimation is not the main focus, the variance of noise is assumed to be a known parameter for speech processing throughout this thesis.

2.2 Time Domain Methods

2.2.1 Comb Filtering

Comb filtering is designed based on the quasi-periodic nature of voiced sound in the speech waveform [46, 47]. It is one of the earliest attempts to reduce the noise in the literature of speech enhancement. Normally the number of filter coefficients range

Figure 2.2: Block diagram of comb filtering

Figure 2.3: Speech waveform and comb filter with constant pitch period T

Figure 2.4: Block diagram of the statistical filtering problem

from 3 to 7 and the shape of the filter is like a comb. The basic idea can be described in the block diagram in Figure 2.2. The voiced/unvoiced decision should be made first. For the unvoiced speech, besides the silence period, a simple factor is applied to suppress the noise. For the voiced speech, the impulse response of a FIR filter guided by the extracted pitch interval is applied. The output is the weighted sum of successive points separated by a constant pitch period of desired signal. Hence, successive periods of voiced sound will add constructively, whereas the output caused by background noise or unvoiced sound with a different fundamental frequency will not. The speech waveform with the corresponding comb filter is illustrated in Figure 2.3. However, the comb filter is not very effective as the amount of noise reduced is dependent on the number of coefficients. A larger number of coefficients will work better for noise reduction but is bad for the situation when speech characteristics changes rapidly. The accuracy of pitch detection is also a key point which may affect the results a lot especially under the low SNR input. Therefore, comb filter is not competitive nowadays.

2.2.2 Wiener Filtering and Linear Prediction

A statistical filtering problem is considered in Figure 2.4. The Linear time-invariant filter is designed in such way that the output signal, enhanced speech, $\hat{s}(t)$, is as close

to the desired clean speech, $s(t)$, as possible. This can be done by minimizing the estimate error, $e(t)$. The optimal filter with MMSE is the well-known Wiener filter, named after the mathematician Norbert Wiener. The filter can be implemented using FIR or IIR filters, but FIR filters are often used since they are inherently stable and the resulting solution is linear and computationally easy to evaluate.

Another interesting thing is that, if the desired signal in Figure 2.4 is equal to the input signal, i.e. $s(t) = y(t)$, then the Wiener filter becomes a linear prediction filter. In other words, the Wiener filter predicts the present input value based on a few past input values. Thus the LPC can also be computed by the autocorrelation method or covariance method based on the autoregressive (AR) model of speech. This is the kind of LPC based filtering for speech enhancement and one of the most quoted work is proposed by Lim and Oppenheim [10] which is an iterative scheme to estimate the AR parameters from the noisy speech. The main drawback of this algorithm is that the speech formants shift in location and decrease in formant bandwidth as additional iterations are performed, although this can be improved by adding spectral constraints [3]. Wiener filter is known to be a linear solution to estimate the clean speech in the mean-square sense.

2.2.3 Kalman Filtering

Wiener filter is derived under the assumption that the underlying signals are stationary, while Kalman filter is an extension of the Wiener filter to handle non-stationary signals and noise. Although they both try to minimize the mean square error, the Kalman filter can be viewed as a sequential MMSE estimator. It characterizes the signal dynamically using the speech and noise production models. In this way, the enhanced signal is obtained on a sample-by-sample basis rather than on a frame-by-

frame basis as performed by the Wiener filter. In [11], Paliwal and Basu estimated the speech parameters from clean speech first, then used a delayed version of Kalman filter to estimate the speech signal from a given noisy version contaminated by white noise. In [48, 49], it was demonstrated that Kalman filtering with white noise assumption is not very effective in a colored noise environment, while the Kalman filter with colored noise model achieves expectedly better performance. Work in [50] is also proposed to solve the colored noise problem where a better estimate of autoregressive (AR) coefficients for Kalman filter is obtained by a pre-filter. In [13], a more complex speech model which differentiates between voiced and unvoiced speech is proposed. In [51], it is shown that Kalman filtering-based speech enhancement in the sub-band domain not only greatly reduces the computation complexity, but also achieves a better performance as compared to that in the full-band domain.

2.2.4 HMM Filtering and Neural Networks

The hidden Markov model (HMM) and neural network all possess hidden states and can be considered to be non-linear statistical data modeling techniques, which are supposed to yield better performance than linear estimators such as Wiener filter. The main difference is that, the process of HMM must depend only on the last state in order to use a Markov chain, while a neural network may need a lot of past data. Furthermore, neural network is an adaptive model in most cases. Ideally it can train itself to capture online changes. In literature, both these two statistical models are adopted for speech enhancement.

HMM can be considered to be the simplest dynamic Bayesian network, just like the Kalman filter. It is commonly used in speech recognition [52, 53] and was first applied in speech enhancement by Ephraim [14]. One of such algorithm uses two

Figure 2.5: Block diagram of HMM based MMSE estimator

HMMs to model the statistics of speech and noise. For each frame of input noisy speech, this composite model of two HMMs estimate the clean speech and the noisy speech. After that this problem is converted to a MMSE estimation where Wiener filter can be adopted. The final enhanced speech is obtained by a weighted sum of all the estimated results. The weights are based on the posterior probability of the composite states obtained from the HMM models. This process is shown in Figure 2.5, and this HMM based MMSE estimator can be considered as a soft decision filter as well. Some other HMM based work can also be found such as the ones incorporating pitch period and harmonic model [15, 16]. Non-stationary noise corruption can be solved by this technique as well [54].

Artificial neural network is actually a mathematical model or computational model trying to simulate the structure or functions of a biological neural network. The most commonly used neural network is the three layered network which is shown in Figure 2.6. Each layer has a set of neurons and each neuron is a nonlinear weighted sum of previous layer of neurons. Then the input layer is fully connected to the hidden layer through a set of weights. Similarly the hidden layer is fully connected to the output layer through another set of weights. Thus all the input

Figure 2.6: The three layered neural network

data can be passed through nonlinear transfer functions to get the final output. For the case of speech enhancement, a noisy speech sequence corresponds to the input layer data and the enhanced speech sequence corresponds to the output layer results when the neural network has been trained properly.

Although the training process can be lessened by the learning capability of neurons, it is still quite important since the artificial neurons are not as intelligent as the biological neurons of human beings. Training process gives the proper set of weights which strongly affects the final output. A large set of input noisy speech training set and its corresponding clean speech are thus required for training. The necessity of training creates an operational problem. A neural network trained for a particular noise type cannot be applied directly to other noise types or at other signal to noise ratios (SNRs) without severe performance degradation. Such limitations can be found in [55] where only white noise at 7.4 dB segmental SNR is considered, and in [56] only the impulsive noise type is considered. For some other works, neural network is combined with other techniques such as Kalman filtering [57] and

Figure 2.7: Structure of transform domain based speech enhancement

HMM [58].

2.3 Transform Domain Methods

Many speech enhancement algorithms prefer operating in the transform domain since the speech energy is not present in all the transform coefficients and it is therefore easier to filter off the noise especially for the noise-only coefficients. As mentioned in the beginning of this chapter, transform domain filters are applied on the transform coefficients instead of input noisy speech sequence in time domain. A transform is applied before noise reduction filters and the corresponding inverse transform is used to revert to the enhanced speech sequence. The process of transform domain based speech enhancement algorithms is shown in Figure 2.7. Different transforms may have different analysis methods.

2.3.1 Spectral Filtering

Prior to introducing speech enhancement algorithms in frequency domain, so-called spectral filters, the discrete Fourier transform (DFT) should be reviewed first.

The discrete-time Fourier transform of a discrete-time sequence is a continuous function of frequency with a periodicity of 2π . In practice, the time sequence is finite in duration, e.g. consisting of N samples, and then the continuous frequency is sampled at N uniformly spaced frequencies which is referred to as Discrete Fourier

Transform (DFT). Thus a discrete finite signal can be represented in frequency domain. Based on DFT properties, this representation is more appropriate for stationary signals while speech signal is quasi-stationary with temporal and spectral characteristics changing over time. However, the DFT may be applicable to short speech segments (e.g. 10-30 ms) during which the properties of speech do not change much. This is the so-called Short-time Fourier Transform (STFT).

Short-time Fourier Transform

The Short-time Fourier Transform (STFT) of a sequence $x(n)$ is given by:

$$X(m, \omega) = \sum_{n=-\infty}^{\infty} x(n)w(m-n)e^{-j\omega n} \quad (2.3)$$

where $w(n)$ is the analysis window which is time-reversed and shifted by m samples. The STFT is a function of two variables, the time index, m , and the continuous frequency variable, ω . If m is fixed but ω varies, then $X(m, \omega)$ can be viewed as the Fourier transform of a windowed sequence. If ω is fixed and the time m varies, a filtering interpretation emerges. Note that this two-dimensional (2D) function provides the spectrogram of speech signal which is a widely used tool employed for studying the time-varying spectral characteristics of speech. For convenience, the time shift m can be considered to be the time index of spectrogram.

Similar to the discrete Fourier transform (DFT), a discrete version of the STFT can be obtained by sampling the frequency variable ω at N uniformly spaced frequencies. The resulting discrete STFT is defined as:

$$X(m, k) = \sum_{n=-\infty}^{\infty} x(n)w(m-n)e^{-j\frac{2\pi}{N}kn} \quad (2.4)$$

where $k = 0, 1, \dots, N-1$. These transform coefficients are complex numbers and can be separated into phase θ and magnitude $|\cdot|$, e.g. $X(m, k) = |X(m, k)|e^{j\theta_X(m, k)}$.

The inverse STFT process is achieved by the inverse DFT transform and an overlap & add process [59]. The inverse DFT is given by:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j \frac{2\pi}{N} kn} \quad (2.5)$$

The DFT of a real sequence exhibits conjugate symmetry about $N/2$ which means that only half of the data need to be processed. The phase is normally left untouched and filtering is performed on the magnitude coefficients alone. Furthermore, a fast computational algorithm, Fast Fourier Transform (FFT), exists for calculating the DFT and is able to reduce the computational effort.

Frequency Domain Additive Noise Model

Since Fourier Transform is a linear transform, the additive noise model is still applicable in the frequency domain. Therefore the noisy speech from Equation (2.1) can be expressed in the frequency domain as follow:

$$Y(\omega) = S(\omega) + N(\omega) \quad (2.6)$$

where Y , S and N are the complex Fourier transformed signals of noisy speech, $y(t)$, clean speech, $s(t)$, and noise signal, $n(t)$, respectively, ω is the frequency index.

As mentioned in Section 2.1, background noise is assumed to be a stationary process. Thus the expectation of noise spectrum can be estimated and updated during periods when speech is absent.

Spectral Subtractive Filtering

The spectral subtractive filtering is one of the simplest algorithms proposed for noise reduction. The basic principle is described in [17]. An estimate of the clean speech

spectrum is obtained by subtracting an estimate of the noise spectrum from the noisy speech spectrum. This process can be expressed as follows:

$$\hat{S}(\omega) = \max [|Y(\omega)| - E(|N(\omega)|), 0] e^{j\theta_Y(\omega)} \quad (2.7)$$

where $\hat{S}(\omega) = |\hat{S}(\omega)|e^{j\theta_Y(\omega)}$ is a complex number with noisy phase information, $\theta_Y(\omega)$ is the phase information of noisy speech which is not processed during noise reduction filtering in most cases, $E(.)$ is the expectation function and $E(|N(\omega)|)$ is the average noise magnitude of coefficient ω . Ideally it should be $|N(\omega)|$ instead of $E(|N(\omega)|)$ but there is no way of obtaining it. The half-wave rectification process is carried out by the $\max(.)$ function. It is one of the many ways (such as full-wave rectification) of ensuring nonnegative $|\hat{S}(\omega)|$.

In some cases, it might be best to work on the power spectra rather than magnitude spectra. The magnitude spectral subtraction algorithm can be easily extended to the power spectrum domain by multiplying the conjugate $Y^*(\omega)$ to Equation (2.6), which becomes as follows:

$$|Y(\omega)|^2 = |S(\omega)|^2 + |N(\omega)|^2 + 2\text{Re}[S(\omega)D^*(\omega)] \quad (2.8)$$

where $\text{Re}(\cdot)$ is the function to return the real part of a complex number. If the noise is assumed to be zero mean and uncorrelated with the clean speech, then the term $\text{Re}[S(\omega)D^*(\omega)]$ reduces to zero. Thus the estimate of clean speech power spectrum can be obtained as follows:

$$|\hat{S}(\omega)|^2 = \max \{|Y(\omega)|^2 - E[|N(\omega)|^2], 0\} \quad (2.9)$$

Similar to $E[|N(\omega)|]$, $E[|N(\omega)|^2]$ can be estimated and be used to replace the unknown part $|N(\omega)|^2$. Half-wave rectification is also used to filter off the negative values. Equation (2.9) thus describes the Power Spectrum Subtraction algorithm.

From the above discussion it can be seen that spectral subtractive algorithms are computationally simple to implement. However, this simple subtraction comes with the price of introducing a distortion in the enhanced speech known as musical noise. The technique proposed in [20] mitigates the musical noise distortion to some degree by using power spectral subtraction together with the a-priori signal to noise ratio. Different variations of spectral subtraction have been developed over the years. The most common variation involved the use of an over-subtraction factor that controlled the amount of speech spectral distortion caused by the subtraction process. Different methods were proposed for computing the over-subtraction factor based on different criteria that included linear [60] and nonlinear [61] functions, and psychoacoustic masking thresholds [62], etc.

Wiener Filtering

Wiener filter can be implemented both in time domain (Section 2.2.2) and frequency domain. To minimize the mean square error, the general form of the Wiener filtered speech in the frequency domain can be obtained by:

$$\hat{S}(\omega) = \frac{\lambda_{SY}(\omega)}{\lambda_{YY}(\omega)} Y(\omega) \quad (2.10)$$

where $\lambda_{SY}(\omega)$ is the cross-power spectrum and $\lambda_{YY}(\omega)$ is the expectation of the power spectrum of the noisy speech. They are defined as follows:

$$\lambda_{SY}(\omega) = E[S(\omega)Y^*(\omega)] \quad (2.11)$$

$$\lambda_{YY}(\omega) = E[|Y(\omega)|^2] \quad (2.12)$$

If the same assumption in spectral subtractive filtering is made that background noise is zero mean and uncorrelated with clean speech, the estimated clean speech

spectrum can be converted to:

$$\hat{S}(\omega) = \frac{\xi(\omega)}{\xi(\omega) + 1} Y(\omega) \quad (2.13)$$

where ξ is the a-priori Signal to Noise Ratio (SNR) and be defined as:

$$\xi(\omega) = \frac{\lambda_{SS}(\omega)}{\lambda_{NN}(\omega)} \quad (2.14)$$

where

$$\lambda_{SS}(\omega) = E[|S(\omega)|^2] \quad (2.15)$$

$$\lambda_{NN}(\omega) = E[|N(\omega)|^2] \quad (2.16)$$

The a-priori SNR is unknown in practice and can be estimated by the well-known decision-directed approach [18]:

$$\hat{\xi}(m, \omega) = \alpha \frac{|\hat{S}(m-1, \omega)|^2}{\lambda_{NN}(\omega)} + (1 - \alpha) \max \left[\frac{|Y(m, \omega)|^2}{\lambda_{NN}(\omega)} - 1, 0 \right] \quad (2.17)$$

where m is the time frame index and the current frame of the a-priori SNR is computed based on the estimated clean speech energy of previous frame. Parameter, α , controls the behavior of the SNR estimator and is normally set to 0.98. The effect of varying the value of α is discussed in detail in [63]. Amongst the simpler filters, Wiener filter produces the highest noise reduction.

Minimum Mean Square Error Filtering

The previous part clearly shows that Wiener filter can be considered as the optimal complex spectral estimator in the mean square error sense. However, it is not the optimal spectral magnitude estimator. Acknowledging the importance of the magnitude on speech quality, several authors have proposed optimal methods for obtaining the spectral amplitude from noisy observations. One such excellent work, Minimum Mean Square Error (MMSE) estimator is proposed by Ephraim and Malah

[18]. The magnitude is assumed to follow a Rayleigh distribution and the phase is assumed to follow a uniform distribution in this part. By minimizing the mean square error, the enhanced speech is obtained by the following equation:

$$|\hat{S}(\omega)| = \frac{\sqrt{\pi}}{2} \frac{\sqrt{v(\omega)}}{\gamma(\omega)} M[-0.5; 1; -v(\omega)] |Y(\omega)| \quad (2.18)$$

where $\xi(\omega)$ is the a-priori SNR and can be estimated by decision-directed approach as shown in Equation (2.17), $v(\omega)$ is defined in Equation (2.19), $\gamma(\omega)$ is the a-posteriori SNR and is defined in Equation (2.20) by incorporating Equation (2.16).

$$v(\omega) = \frac{\xi(\omega)}{1 + \xi(\omega)} \gamma(\omega) \quad (2.19)$$

$$\gamma(\omega) = \frac{|Y(\omega)|^2}{\lambda_{NN}(\omega)} \quad (2.20)$$

Function $M(; ;)$ in Equation (2.18) is the confluent hypergeometric function which is defined in [64]. It can be computed efficiently using the following series summation:

$$M(a; b; c) = \sum_{m=0}^{\infty} \frac{(a)_m c^m}{(b)_m m!} = 1 + \frac{ac}{b1!} + \frac{a(a+1)c^2}{b(b+1)2!} + \frac{a(a+1)(a+2)c^3}{b(b+1)(b+2)3!} + \dots \quad (2.21)$$

Although the principle based on the squared error of the magnitude spectra is mathematically tractable, it may not be subjectively meaningful. It has been suggested that a principle based on the squared error of the log-magnitude spectra may be more suitable for speech processing in the biological auditory sense. An estimator that minimizes the mean square error of the log-magnitude spectra is derived in [65]. This optimal log-MMSE estimator is given by the following equation:

$$|\hat{S}(\omega)| = \frac{\xi(\omega)}{1 + \xi(\omega)} \exp\left(\frac{1}{2} \int_{v(\omega)}^{\infty} \frac{e^{-t}}{t} dt\right) |Y(\omega)| \quad (2.22)$$

where $v(\omega)$ is defined in Equation (2.19), and $\xi(\omega)$ is defined in Equation (2.14) which can be estimated by Equation (2.17).

MMSE based speech enhancement algorithms are known to be good at noise reduction without leaving obvious annoying musical tones in contrast to the spectral subtractive algorithms.

2.3.2 Discrete Cosine Transform based Filtering

The Discrete Cosine Transform (DCT) is widely used in image compression because of its excellent energy compaction property which is comparable with Karhunen-Loeve transform (KLT) but with a more efficient computational load. KLT has been successfully applied in speech enhancement area and is used in one of the subspace speech enhancement algorithms which will be introduced later. DCT, as a good approximation to the KLT and being a fast transform algorithm, should be a good choice for speech enhancement and this has been shown in [21].

Unlike DFT, the DCT coefficients are real numbers without phase information. The forward DCT of sequence $x(n)$ with N samples is given as follow:

$$X(k) = \alpha(k) \sum_{n=0}^{N-1} x(n) \cos \left[\frac{\pi(2n+1)k}{2N} \right], k = 0, 1, \dots, N-1 \quad (2.23)$$

where

$$\alpha(0) = \sqrt{\frac{1}{N}}, \alpha(k) = \sqrt{\frac{2}{N}}, k = 0, 1, \dots, N-1 \quad (2.24)$$

and then the inverse DCT is given by:

$$x(n) = \sum_{k=0}^{N-1} \alpha(k) X(k) \cos \left[\frac{\pi(2n+1)k}{2N} \right] \quad (2.25)$$

The capability of high energy compression is an advantage for noise reduction since if the speech energy can be compressed into fewer coefficients while the noise energy remains white, reduction of noise can be achieved easily. Some classical techniques, such as Wiener filter [23], minimum mean square error estimator [21], hard and soft thresholding criteria [66], etc, can also be applied by making the necessary modifications to the derivations in the DCT domain.

2.3.3 Subspace Filtering

Subspace filtering is quite a different class of speech enhancement algorithms that are largely based on linear algebra theory. More specifically, these algorithms are based on the principle that the clean signal might be confined to a subspace of the noisy Euclidean space. The decomposition of the vector space of the noisy speech into speech and noise subspaces can be done using orthogonal matrix factorization techniques, in particular, the Singular Value Decomposition (SVD) or eigenvector-eigenvalue factorization. The orthogonal matrices obtained from the SVD or eigenvalue decomposition can be viewed as signal dependent transforms. This transform can be applied to either a matrix containing the signal or to a speech vector. The well-known Karhunen-Loeve transform (KLT) is actually the eigenvector matrix of the signal covariance matrix which will be briefly introduced as an example in this section.

A KLT based speech enhancement algorithm is the process whereby KLT components are filtered by a gain function and then undergo the inverse KLT to restore the enhanced speech. Let R_s denote the autocorrelation matrix of clean speech $s(n)$ with K samples, the eigen-decomposition of R_s is given by:

$$R_s = U\Lambda_s U^T \quad (2.26)$$

where $\Lambda_s = \text{diag}[\lambda_s(1), \lambda_s(2), \dots, \lambda_s(K)]$ and U are the eigenvalue matrix and the orthogonal eigenvector matrix of R_s respectively. As analyzed in [28], if the number of basis vectors is smaller than the dimension of the vector, then the covariance matrix of the vector must have zero eigenvalues. On the other hand, all eigenvalues of white noise are positive and fill the entire vector space of the noisy signal. Hence, the vector space of a noisy speech is composed of a signal-plus-noise subspace, which should be processed by speech enhancement algorithms, and a noise

dominated subspace which can be simply discarded. According to the spectral domain constrained estimator proposed in [28], a linear filter of the subspace speech enhancement method H_{sub} can be obtained by:

$$H_{\text{sub}} = UGU^T = U \begin{bmatrix} G_1 & 0 \\ 0 & 0 \end{bmatrix} U^T \quad (2.27)$$

G_1 is a full rank $M \times M$ matrix where M is the assumed number of non-zero eigenvalues of the autocorrelation matrix R_s . A possible solution of the spectral domain constrained estimator is that G is a diagonal matrix given by:

$$G = \Lambda_s(\Lambda_s + \sigma_n^2 \Lambda_\mu)^{-1} \quad (2.28)$$

where σ_n^2 is the noise variance, $\Lambda_\mu = \text{diag}[\mu(1), \mu(2), \dots, \mu(K)]$ is a diagonal matrix of Lagrange multipliers when deriving the estimator, H_{sub} , based on the criterion of minimizing the speech signal distortion with a permissible residual noise level [28]. The k th diagonal element of G is given by:

$$g(k) = \begin{cases} \frac{\lambda_s(k)}{\lambda_s(k) + \mu(k)\sigma_n^2}, & k = 1, 2, \dots, M \\ 0, & k = M + 1, \dots, K \end{cases} \quad (2.29)$$

The method above is based on the white noise model. For the colored noise type, a technique proposed in [67] is utilized depending on a speech/noise dominated frame detection, so-called signal/noise KLT-based approach. An adaptive algorithm [68] can also be used to solve this problem.

Chapter 3

DCT Based Speech Enhancement Filtering

3.1 Introduction

Discrete Cosine Transform (DCT) has been proven to be an acceptable alternative to the Discrete Fourier Transform (DFT) for speech enhancement [21, 66, 69–71]. Work in this area was initiated by Soon et al. [21] who presented several advantages of DCT for enhancing speech as compared to DFT. These advantages of DCT can be summarized as follows:

- It provides a significantly higher energy compaction capability.
- It is a real transform without phase information.
- It provides a higher resolution for analyzing the transform coefficients in the same window length.

Figure 3.1: Comparison of energy compaction between DCT and DFT

The higher energy compaction capability is the main advantage for speech enhancement since if the speech energy can be compressed into fewer coefficients while the background noise remains white, then the noise can be easily removed and there is less chance to distort the speech signal during the noise reduction process. Several comparisons have been implemented for illustrating the energy compaction capability of DCT in [21]. Among those, one simple experiment is to use actual speech signal divided into frames with 50% overlap, and then the transform is performed. The transform magnitudes are sorted in descending order and the last n coefficients are set to zero. The resulting speech is then reconstructed using the weighted overlap & add process. The mean square error between the original and the resulting

Figure 3.2: Upper limits of enhancement using DCT and DFT

speech is computed and plotted against n for both the DFT and DCT in Figure 3.1. It can be clearly seen that the DCT provides better energy compaction.

For the second advantage, the DCT coefficients only have sign information instead of the phase information in the DFT. Most of the DFT based algorithms only attempt to filter the spectral magnitude while leaving the noise corrupted phase information intact, since it has been reported that the best estimate of the phase is the corrupted one itself [18]. Such no-action actually results in an upper bound on the maximum possible improvement in signal to noise ratio (SNR). The comparison of upper limits of enhancement using DCT and DFT can be found in Figure 3.2 which is cited from [21], and the detail experimental procedure can also be found

in [21].

The last advantage is very straight-forward based on the transform properties which can be found in Sections 2.3.1 and 2.3.2. DFT only produces about half the independent spectral components as the other half are complex conjugates, while DCT produces fully independent spectral components. Therefore, a higher resolution for analyzing spectral coefficients can be achieved by using DCT.

This chapter will focus on a zoom-in portion of DCT during the frame based analysis together with an improved noise reduction filter. In traditional DCT based speech enhancement algorithms, the transform is performed by a short-term cosine transform which is almost the same as short-term Fourier transform (STFT) described in Section 2.3.1 except for a different matrix transform. In other words, DFT can be simply replaced by DCT during the frame-by-frame processing. In such algorithms, the observed speech is divided into fixed overlapping frames ranging from 50% to 75% and then processed by DCT. Following which a noise suppression filter is applied on the DCT coefficients. However, this procedure does not take into account the differences between DCT and DFT. One of the major differences is that the DCT coefficients are real, while the DFT coefficients are complex and comprises a magnitude and a phase representation. Without a phase representation, DCT coefficients magnitudes obtained by a standard window-shift show much higher variation as compared to the magnitudes of DFT coefficients for a strictly stationary signal. This will impact negatively on the inter-frame techniques such as the decision-directed approach [18] for the estimation of a-priori SNR. Therefore, an adaptive time-shift analysis (ATSA) system is proposed to compensate for the difference between DCT and DFT. It helps to improve the performance of DCT based speech enhancement algorithms especially those using inter-frame techniques.

Figure 3.3: Respective results of DFT and DCT by placing the analysis window in different positions

3.2 Motivation

Some frequency domain speech enhancement algorithms can be correspondingly applied in the DCT domain such as the decision-directed approach for Wiener filter [21], since DCT expresses a signal in terms of a sum of sinusoids with different frequencies and amplitudes, which is similar to the Fourier transform. However, DCT does have some differences from DFT which leads to uniquely useful properties for various applications, e.g. a good performance for image compression.

The obvious distinction between DCT and DFT is that the former uses only cosine functions which means that only real numbers are obtained, while the latter

uses both cosines and sines. Thus DCT is equivalent to the DFT of real and odd functions without phase information, and only a sign signal is embedded instead. Furthermore, it operates on a signal over a finite number of samples by an even extension of the original signal, while the DFT is based on a periodic extension. Therefore, how the observed signal is truncated strongly affects the DCT coefficients even if the signal is highly stationary. This is in contrast to hardly noticeable differences in the DFT coefficients for the same situation. This can be verified by a simple experiment. A pure sine wave with frequency of 100 Hz (80 samples in one period) is generated. Three different positions within one period of signal are selected for the analysis window beginning with the first dotted line as shown in Figure 3.3. Then the truncated frame undergoes the standard DFT and DCT, and the respective results are shown in the second and third rows in Figure 3.3. From this figure, it clearly shows that DFT coefficients are not as sensitive to the position of analysis window as compared to the DCT coefficients. For the unvoiced sounds which are considered as non-periodic random signals, the transformed coefficients are too random to distinguish the differences between different analysis windows. Hence when unvoiced sounds are encountered, the algorithm falls back to the standard fixed length shift.

This observation indicates a shortcoming in DCT basis that needs to be overcome. For frame by frame based speech enhancement algorithms, it is better to generate a more consistent DCT amplitude as inter-frame techniques can then be accurately applied. If the shift of the analysis window is exactly equivalent to the pitch period of the signal, the DCT coefficients will be the same from frame to frame for a stationary signal. This will result in a more accurate a-priori SNR estimation.

Figure 3.4: The block diagram of proposed adaptive time-shift analysis system

3.3 Structure of Adaptive Time-shift Analysis

In DCT based speech enhancement algorithms, input speech sequence should be truncated into frames first. The basic idea of the proposed adaptive time-shift analysis (ATSA) is to detect the pitch periodicity first and then shift the analysis window by the pitch period to obtain the speech segments. This method will theoretically produce constant DCT coefficients for stationary signals. However the variable shift used above is only applicable to voiced speech signals, and for unvoiced/silence parts of the speech signal, the algorithm falls back to the standard time-shift for the analysis window. Despite this, significant improvements can still be obtained as voiced speech are more dominant than unvoiced speech.

The structure of this proposed adaptive time-shift analysis is shown in Figure 3.4. The initial speech frame is filtered by a noise reduction technique, and then a

Figure 3.5: Magnitude of a rectangular window of length 100 with 1024-point transform

voiced/unvoiced decision is made. If it contains voiced signal, the time-shift will be changed to one pitch period. Otherwise, the time-shift will fall back to the original fixed value. In this way, the analysis window shift adapts to the underlying speech properties and it is no longer fixed.

3.4 Windowing Function

In signal processing, if a signal is to be observed over a finite duration, then a window function has to be applied to truncate this signal. The simplest window function is the rectangular window which causes the well-known problem, spectral leakage

effect. That is, if there are two sinusoids with similar frequencies, leakage interferes with one overwhelmed by the other. If their frequencies are dissimilar, leakage interferes when one sinusoid is much weaker in amplitude than the other. The main reason is that the rectangular window represented in the frequency domain has strong side-lobes where the first side-lobe is only around 13 dB lower than the main lobe. Similar to Fourier transform, DCT has the same problem with rectangular window. These side-lobes effects can be seen in Figure 3.5. The rectangular window also has some disadvantages such as the discontinuities at the endpoints or maximum scalloping loss for frequency component that is exactly in the middle of two FFT coefficients. Thus some other window functions are used instead in many DCT applications. For instance, a sine window is widely used in audio coding because it offers good stop-band attenuation for high quality coding, e.g. MP3 and MPEG-2 ACC. Some other window shapes, such as Kaiser-Bessel derived window, are used for Vorbis, AC-2, etc.

Apart from this, rectangular window does have some advantages. It has a narrower main-lobe which is able to resolve comparable strength signals. Despite its 3 dB potential scalloping loss, it is the best choice for detecting a sinusoid at low signal to noise ratio (SNR). Besides, one advantage of using the DCT as compared to DFT is that there is no discontinuity problem caused by rectangular window at the endpoint, since DCT is based on an even symmetrical extension during the transform of a finite signal. Therefore, the selection of the window is based on a trade-off between spectral resolution and leakage effects. In the literature of DCT based speech enhancement algorithms, the Hann window is very popular [21]. Trapezoidal window is another possible window and has been applied in [69]. In this chapter, a bold attempt is made by choosing a rectangular window for DCT based speech enhancement although it exhibits some unsatisfactory aspects, but it still has some

interesting properties which is worth exploiting.

3.5 Pitch Synchronization

In order to implement the ATSA algorithm, the pitch period should be extracted first. There are many ways to estimate the pitch periodicity of a speech signal. But almost all the algorithms work well only in clean situations. Therefore, the noise reduction filtering should be performed first. To simplify this process and emphasize on the main contribution of proposed algorithm, a basic Wiener filter [21] is adopted in this chapter. There are many other more advanced algorithms [21, 66, 69–71] that can be applied too.

Let the DCT signals of noisy speech, clean speech and noise signal be $Y(m, k)$, $S(m, k)$ and $N(m, k)$, respectively, which have been transformed by Equation (2.23), where m is reframed as the time frame index and k is position of coefficient in the DCT transform domain. The additive noise model described in Equation (2.1) is considered, and it is still valid in the DCT domain since DCT is a linear transform. Wiener filter has been proven to be the optimal minimum mean square error (MMSE) amplitude estimator for the DCT transform [21] based on the Gaussian statistical model used. Then the Wiener filtered speech $S(\hat{m}, k)$ can be given by:

$$\hat{S}(m, k) = \frac{\hat{\xi}(m, k)}{\hat{\xi}(m, k) + 1} Y(m, k) \quad (3.1)$$

where the a-priori SNR $\hat{\xi}(m)$ can be estimated by the decision-directed approach in Equation (2.17). Following this noise reduction filtering, the enhanced speech after inverse DCT, $\hat{s}(n)$, is utilized for pitch detection to obtain a more accurate estimation.

A variety of algorithms are proposed in the past to detect the pitch period. The

time domain autocorrelation method is quite a common approach for solving this problem, since it is simple and robust to some noise corruption conditions. It is selected in this chapter for extracting the pitch period to be used for the time-shift. Firstly, the observed signal frame has its DC removed. Next clipping is applied for a better pitch period estimation. The clipping level is determined as a fixed percentage (e.g. 68%) of the minimum of the maximum absolute values in the first and last 85 samples of the speech frame. Using this clipping level, the speech signal is processed by a 3-level center clipper and the correlation function is computed over a range spanning the expected range of pitch periods. The autocorrelation function of the resulting signal $\hat{s}(n)$ can be defined as:

$$R(n) = \sum_{m=0}^{N-m-1} \hat{s}(m)\hat{s}(n+m) \quad (3.2)$$

Since fundamental frequency in spoken English language is range bound between 80Hz to 500Hz, a distinct peak in this range implies the presence of voiced signals. A distinct peak is defined to be greater than 0.5 times of $R(0)$. If no distinct peak can be found, it would mean that it is likely to be a silence or unvoiced frame. For voiced frames, the pitch period is extracted and used as the analysis window shift. It should be noted that, in this method, the window length needs to be at least twice as long as the longest pitch period of the observed speech signal.

The final enhanced speech is obtained by overlap & add process. Actually this process is a little different from the original process due to the adaptive window shifting. A convenient solution is to produce a weighting function which records all the windows frame by frame and calculates the net weighting function. After that, the enhanced speech has to be normalized by the weighting function.

3.6 Pitch Synchronization with Maximum Alignment

During the process of adaptive time-shift analysis (ATSA), the pitch synchronization step can be further improved by a maximum alignment technique which means the speech segments start from the maximum and the time shift equals to one period.

As mentioned previously, DCT coefficients are the representation of a signal in terms of a sum of sinusoids with different frequencies and amplitudes. From the mathematical point of view, DCTs are organized by different basis functions together with certain boundary conditions. There are four standard types of DCT based on their basis functions. The one used in Equation (2.23) is the DCT-II which is commonly used and initially derived as approximations of the eigenvectors which would give an optimal Karhunen-Loeve (KL) basis. Figure 3.6 shows the basis functions for DCT-II with 8 sample points ($N = 8$). From this figure, it can be seen that all the basis functions start from the positive values (non-zero points) which can be expanded to different number of points based on the Equation (2.23). If the truncated signal also starts from the short-term maximum, then the product of these two will combine their positive efforts: one is the variance reduction achieved by pitch synchronization, the other is the further energy compaction achieved by the maximum alignment.

The capability of further energy compaction can be illustrated by the simple experiment which has been described in Section 3.1. A few transform coefficients with lowest values are set to zero, and then the mean square error (MSE) between reconstructed signal and original signal is calculated. Lower value of MSE indicates a better energy compaction. A real speech segment of voiced sound with three different positions of analysis window (window length is 32 ms) is adopted here and the

Figure 3.6: Basis functions for DCT-II, $N = 8$

Figure 3.7: Comparison of energy compaction among three different positions of starting point

MSE results are shown in Figure 3.7. It clearly shows that starting from maximum performs the highest capability of noise compaction while starting from other positions perform worse. Based on the DCT basis functions, the speech segment with maximum at the beginning is better represented by the lower basis vectors while the segment with zero at the beginning cannot be so efficiently represented by the basis vectors. Actually an ideal case can be obtained artificially. If the window size is the integer multiple of the period and a pure sine wave segment has its maximum on both sides, then one distinctive impulse appears in the lowest frequency. Therefore, a more compressed speech energy representation in DCT domain can be obtained

Figure 3.8: 5 impulses with pitch period

by this maximum alignment technique. Combined with pitch synchronization technique which is able to reduce the frequency variances, an advanced speech analysis technique is generated for DCT based speech enhancement algorithms. Of course the above technique is still based on the voiced sound, and the analysis window is rectangular window which will not suppress the more dominant values on the starting point in contrast to other smooth window functions.

To implement this scheme, Wiener filtered speech is used to extract a more accurate pitch period compared with noisy speech. Several impulses with period equal to the pitch period of the current voiced frame are generated to calculate the cross-correlation with Wiener filtered speech. Let $\hat{s}(n)$ denote the Wiener filtered speech

sequence and $im(n)$ denotes the impulse sequence, the discrete cross-correlation of these two real signals is given by:

$$R_{sim}(n) = \sum_{m=-\infty}^{\infty} \hat{s}(m)im(n+m) \quad (3.3)$$

where the impulse sequence $im(n)$ can be expressed as:

$$im(n) = \sum_{m=-N}^N \delta(n-mT) \quad (3.4)$$

where $\delta(n)$ is the delta function which equals to 1 only if $n = 0$, else $\delta(n)$ equals to 0. T is the period and N could be ∞ or a fixed value which indicates the impulse train is infinite or finite respectively. In this thesis, the number of impulses are empirically fixed to 5 on $N = 2$. A voiced speech frame with its corresponding impulse signal is illustrated in Figure 3.8. After that, the maximum position of the current speech frame will be identified by tracking the maximum of cross-correlation values.

3.7 Wiener Filter with Adaptive Controller

Wiener filter has been proven to be the optimal filter for the real transform in minimum mean square error (MMSE) sense [21]. During the implementation, it fully depends on the estimation of the a-priori signal to noise ratio (SNR). The a-priori SNR can be computed in many ways among which the decision-directed approach [18] is widely used. From the expression in Equation (2.17) of decision-directed approach, a parameter α is used to set a proportion of contributions from the previous frames to the current estimate. In Fourier transform domain, the value of α is normally equal to 0.98 which is an empirically obtained value and is known to be a good trade-off between noise reduction and speech distortion [18]. The same value of α is also commonly used in DCT speech enhancement schemes [21, 23].

However, this might not be proper for new situation since DCT coefficients may require a different value of α or even an adaptive one.

In DFT domain, there are some work about adapting α for better estimation of the a-priori SNR [72, 73]. Thus it is workable to propose an adaptive α for decision-directed approach in DCT domain which leads to an improved version of Wiener filter. The MMSE criterion is used to derive the optimal expression. Recall the decision-directed approach in Equation (2.17), the a-priori SNR can be expressed in DCT domain as:

$$\hat{\xi}_{m,k} = \alpha_{m,k} \tilde{\xi}_{m-1,k} + (1 - \alpha_{m,k}) \max(\gamma_{m,k} - 1, 0) \quad (3.5)$$

where the time frame index m and frequency index k are moved to the lower right corner to simplify the expression. $\tilde{\xi}_{m-1,k}$ is defined by:

$$\tilde{\xi}_{m-1,k} = \frac{|\hat{S}_{m-1,k}|^2}{\lambda_{NN}} \quad (3.6)$$

where $\gamma_{m,k}$ and λ_{NN} are the corresponding DCT versions of Equation (2.20) and Equation (2.16), respectively. Then the error between the estimated a-priori SNR $\hat{\xi}_{m,k}$ and the real one $\xi_{m,k}$ is:

$$J_\alpha = E \left\{ (\hat{\xi}_{m,k} - \xi_{m,k})^2 \right\} \quad (3.7)$$

If $E \{ \gamma_{m,k} - 1 \}$ is set to $\xi_{m,k}$ which is reasonable based on the same assumption of Equation (2.9), then Equation (3.7) can be rewritten as:

$$J_\alpha = \alpha_{m,k}^2 (\tilde{\xi}_{m-1,k} - \xi_{m,k})^2 - (1 - \alpha_{m,k})^2 \xi_{m,k}^2 + (1 - \alpha_{m,k})^2 E \{ (\gamma_{m,k} - 1)^2 \} \quad (3.8)$$

Based on the assumption that DCT coefficient of speech signal $S_{m,k}$ and noise signal $N_{m,k}$ can be modeled as zero mean random Gaussian variables which are independent of each other [21], $E \{ (\gamma_{m,k} - 1)^2 \}$ can be expressed as:

$$E \{ (\gamma_{m,k} - 1)^2 \} = 3\xi_{m,k}^2 + 4\xi_{m,k} \quad (3.9)$$

$E \left\{ \frac{S_{m,k}^4}{\lambda_{NN}^2} \right\} = 3\xi_{m,k}^2$ is used in Equation (3.9) based on the assumption that the DCT coefficient of speech signal $S(m, k)$ has a Gaussian distribution. Incorporating Equations (3.7), (3.8) and (3.9), the error J_α can be finally obtained by:

$$J_\alpha = \alpha_{m,k}^2 (\tilde{\xi}_{m-1,k} - \xi_{m,k})^2 + (1 - \alpha_{m,k})^2 (2\xi_{m,k}^2 + 4\xi_{m,k}) \quad (3.10)$$

Equating $\partial J_\alpha / \partial \alpha_{m,k}$ to zero, the optimal expression of $\alpha_{m,k}$ can be obtained as:

$$\begin{aligned} \alpha_{m,k} &= \frac{2\xi_{m,k}^2 + 4\xi_{m,k}}{(\tilde{\xi}_{m-1,k} - \xi_{m,k})^2 + 2\xi_{m,k}^2 + 4\xi_{m,k}} \\ &\cong \frac{1}{1 + \frac{1}{2} \left(\frac{\tilde{\xi}_{m-1,k} - \xi_{m,k}}{1 + \xi_{m,k}} \right)^2} \end{aligned} \quad (3.11)$$

The approximation used above is to avoid division by zero. As $\xi_{m,k}$ is unknown, Equation (3.11) cannot be applied directly. An approximate value of $\alpha_{m,k}$ can be obtained substituting $\xi_{m,k}$ by $\bar{\xi}_{m,k}$ which is defined as follow:

$$\bar{\xi}_{m,k} = (\gamma_{m,k} - 1) * H(m) \quad (3.12)$$

where $*$ is the convolution operator, $H(m)$ is a low-pass filter and a Gaussian mask is applied here to realize this smoothing function of $H(m)$. The reason for applying this low-pass filter is that it is able to reduce the variance among different speech frames which is caused by noise. This annoying effect can be further reduced by a running average of $\alpha_{m,k}$:

$$\hat{\alpha}_{m,k} = \beta \hat{\alpha}_{m-1,k} + (1 - \beta) \alpha_{m,k} \quad (3.13)$$

where β is a parameter which is fixed at 0.5 in the experiments. From Equation (3.11), if the signal to noise ratio (SNR) changes slowly, the parameter $\alpha_{m,k}$ will be a value close to one. If the SNR has sharp changes, the parameter $\alpha_{m,k}$ will take a smaller value enabling $\hat{\xi}_{m,k}$ to change adaptively. Thus, the adaptive controller, $\alpha_{m,k}$ is in the range of zero to one.

Figure 3.9: Segmental SNR results of noisy speech, Wiener filtered speech with rectangular window and Wiener filtered speech with Hann window

3.8 Results and Discussion

Ten different segments of speeches, five females and five males, are selected from TIMIT database. They are resampled at 8 kHz and corrupted by four additive noise types including white noise, fan noise, car noise and f16 aircraft noise. The proposed ATSA technique is evaluated using two objective measures, segmental SNR (SegSNR) measure and perceptual evaluation of speech quality (PESQ) measure. Since SegSNR is better correlated with Mean Opinion Score (MOS) than SNR as indicated by [74] and is easy to implement, it has been widely used to qualify

Figure 3.10: PESQ results of noisy speech, Wiener filtered speech with rectangular window and Wiener filtered speech with Hann window

the enhanced speech. PESQ described in ITU-T recommendation P.862 is an objective measurement tool that predicts the results of subjective listening tests on telephony systems. It uses a sensory model to compare the original, unprocessed signals with the enhanced signals. As indicated in [75], the SegSNR is a better evaluation in terms of noise reduction, while the PESQ is more accurate in terms of speech distortion prediction. The latter is also more reliable and highly correlated with Mean Opinion Score (MOS) as compared to other traditional objective measures. In most situations, PESQ is the best objective indicator for overall quality of enhanced speech.

Before evaluating the adaptive time-shift analysis system (ATSA), the effects of window functions should be presented. A standard Wiener filtered described in Equation (3.1) with fixed time-shift analysis is utilized. Two different window functions, rectangular window and Hann window are used to truncate the input signal. SegSNR and PESQ results are shown in Figure 3.9 and Figure 3.10 respectively. From these two figures, it is clear that rectangular window is better for DCT based noise reduction algorithms. The selection of window function is quite an interesting part for DCT based speech enhancement algorithms which should be further investigated in the future.

Another topic to be discussed is the overlap between consecutive speech frames. Normally higher overlap results in better performance, e.g. speech analysis with 75% overlapping frames will lead to better resulting speech compared to one with 50% overlapping frames. The pitch period of voiced sound is usually shorter than the fixed time-shifting interval which is 8ms in this thesis. It thus results in a higher overlap during the speech analysis process. In order to make a fair comparison, the total number of speech frames used in the proposed technique, pitch synchronized Wiener filtering (PSWF) described in Section 3.5, is calculated. After that, a higher overlap can be obtained by dividing the same number of speech frames used in PSWF for the same input noisy speech. In other words, this algorithm is a Wiener filter with a higher fixed overlap which can be denoted as WFHO. The comparison results are shown in Table 3.1 with segmental SNR measure and PESQ measure. This table shows that the PSWF performs the best which means the advantage of PSWF algorithm does not come solely from the higher overlap.

To exhibit the advantages of each part of proposed ATSA system, three speech enhancement schemes are compared: the first one is the pitch synchronized Wiener filtering named as PSWF; the second one is to use maximum alignment to improve

Table 3.1: Comparison Results of SegSNR and PESQ measures

Noisy type	SNR (dB)	SegSNR			PESQ		
		Noisy	WFHO	PSWF	Noisy	WFHO	PSWF
White	0	-4.249	0.741	0.781	1.449	2.133	2.137
	5	-1.530	3.018	3.041	1.743	2.530	2.538
	10	1.802	5.483	5.570	2.077	2.871	2.877
	15	5.522	8.419	8.421	2.419	3.172	3.178
	20	9.564	11.608	11.618	2.764	3.412	3.417
Fan	0	-3.043	5.839	5.895	2.503	3.203	3.204
	5	-0.098	8.721	8.808	2.844	3.454	3.456
	10	3.356	11.667	11.884	3.178	3.688	3.691
	15	7.207	14.688	14.966	3.492	3.890	3.893
	20	11.303	17.755	18.052	3.826	4.090	4.096
Car	0	-2.761	9.011	9.052	3.113	3.553	3.558
	5	0.166	11.666	11.788	3.442	3.757	3.760
	10	3.541	14.539	14.704	3.774	3.980	3.985
	15	7.340	17.495	17.545	4.051	4.175	4.177
	20	11.40	20.223	20.258	4.249	3.309	3.310
F16 aircraft	0	-4.245	0.433	0.494	1.683	2.257	2.263
	5	-1.424	2.693	2.721	2.012	2.656	2.664
	10	1.972	5.301	5.319	2.345	2.991	2.997
	15	5.742	8.195	8.331	2.677	3.252	3.256
	20	9.782	11.468	11.614	3.011	3.491	3.492

pitch synchronization as described in Section 3.6 where the noise reduction filter is still the standard Wiener filter. This scheme is named as PSMax; the last one is the Wiener filter controlled by an adaptive parameter α described in Section 3.7. The overlapping and time-shift are adaptive and can be improved by the maximum alignment. Actually the last one is the whole ATSA speech enhancement system proposed in Figure 3.4. The SegSNR and PESQ results are shown in Table 3.2 and Table 3.3 respectively. From these two tables, it can be seen that ATSA performs

Table 3.2: Comparison of Segmental SNR Results

Noise type	SNR (dB)	SegSNR Results		
		PSWF	PSMax	ATSA
White	0	0.78	0.81	1.07
	5	3.04	3.13	3.39
	10	5.57	5.59	5.96
	15	8.42	8.44	8.91
	20	11.62	11.69	12.21
Fan	0	5.90	5.91	6.25
	5	8.81	8.82	9.18
	10	11.88	11.93	12.26
	15	14.97	14.98	15.34
	20	18.05	18.06	18.36
Car	0	9.05	9.10	9.57
	5	11.79	11.81	12.30
	10	14.70	14.72	15.18
	15	17.55	17.57	18.0
	20	20.26	20.26	20.62
F16 aircraft	-10	0.49	0.54	0.76
	-5	2.72	2.80	3.04
	0	5.32	5.36	5.69
	5	8.33	8.34	8.75
	10	11.61	11.64	12.08

the best, followed by the PSMax, and PSWF. This sequence clearly shows that each technique adds to the improvements in the whole system.

3.9 Conclusion

In traditional DCT based noise reduction algorithms, the observed speech signal is divided into fixed overlapping frames and transformed into DCT domain. This process leads to a variation of DCT coefficients from one frame to another because of

Table 3.3: Comparison of PESQ Scores

Noise type	SNR (dB)	PESQ Results		
		PSWF	PSMax	ATSA
White	0	2.137	2.144	2.175
	5	2.538	2.541	2.567
	10	2.877	2.886	2.89
	15	3.178	3.183	3.185
	20	3.417	3.426	3.427
Fan	0	3.204	3.206	3.210
	5	3.456	3.458	3.465
	10	3.691	3.699	3.701
	15	3.893	3.898	3.90
	20	4.096	4.098	4.106
Car	0	3.558	3.563	3.565
	5	3.760	3.766	3.772
	10	3.985	3.989	3.991
	15	4.177	4.182	4.184
	20	4.310	4.313	4.320
F16 aircraft	0	2.263	2.264	2.267
	5	2.664	2.668	2.679
	10	2.997	2.998	2.999
	15	3.256	3.259	3.265
	20	3.492	3.495	3.503

non-ideal analysis window positions. In order to reduce this variation, an adaptive time-shift analysis technique, ATSA, is proposed. The autocorrelation function is used for detecting the pitch period which is in turn used as the amount of shift for the analysis window. Therefore, a consistent DCT spectrogram is generated for better noise reduction filtering. This technique can be further improved by maximum alignment which results in a much better fit for the DCT basis functions. After the observed signal is truncated into frames, noise reduction algorithms are applied to enhance the speech. An adaptive parameter α is proposed for a better estimate of

the a-priori SNR which affects the optimal DCT filter in mean square error sense. All the proposed techniques can be integrated into a complete speech enhancement system named ATSA which produces good quality enhanced speech. Two objective measures, segmental SNR and PESQ are utilized to evaluate the proposed ATSA system. All the results show that significant improvements can be obtained by the proposed techniques.

Chapter 4

Hybrid Wiener Filter for Spectral Filtering

As mentioned in Chapter 2, Discrete Fourier Transform (DFT) is the most popular transform method in transform based speech enhancement. Since the main focus is the enhancement of speech signals degraded by statistically uncorrelated additive noise which is also a stationary random process, the DFT coefficients offer some properties which are good for separating speech signal from the noise such as spectral subtractive criterion described in Section 2.3.1. Besides, DFT has a fast computational method which is able to highly reduce the computational load. Therefore, a lot of papers have been proposed in the past to investigate speech enhancement algorithms in DFT domain. This chapter will introduce a new spectral filtering method operating on the spectral amplitude of the degraded speech.

4.1 Introduction

If noise is assumed to be additive, stationary and statistically independent of the speech signals, various mathematical tools can be used to derive an optimal statistical operator with respect to mean square error (MSE) for noise reduction in DFT domain. The minimum mean square error (MMSE) estimator by [18] is an important milestone. In these estimation approaches, the transform coefficients are filtered in each short-time frame and attenuated independently of their intra-frame neighboring coefficients as well as several inter-frame neighboring coefficients.

However, some correlation does exist among different time frames and this is exploited by many researchers to some extent [17, 76–78]. This is apparent in many recent works which view speech as a 2D time-frequency signal, especially in the form of a spectrogram. Evans has applied morphological filtering on the spectrogram in [79] using opening operator based on erosion and dilation which is borrowed from digital image processing tools, and has obtained improved results. However, this algorithm emphasizes more on 2D processing without exploiting the characteristics of the speech spectrogram, resulting in the attenuation of the speech content together with the noise. The other popular approach in speech spectrogram filtering is the application of a post-processor to further attenuate the residual noise to get a cleaner result. In such cases, noise will first be removed using some conventional filter such as MMSE [80] or spectral subtraction [81, 82]. The post-processing will focus on targeting the residual musical noise which may appear as isolated peaks. In [80], the enhanced spectrogram is first filtered with a 2D low-pass filter and then classified into speech and non-speech decision regions by simple thresholding. Non-speech regions containing mainly residual noise are then further attenuated. Whipple [81] proposed a time-frequency filter to detect and remove isolated peaks

in the spectrogram domain. The time-frequency filter is implemented using energy difference between the inner and outer rectangular windows. A more advanced post-processing method which takes into consideration the shape of the speech features in the spectrogram has been proposed by [82]. This takes into account that the speech spectrogram consists of many peaks and valleys and they run mainly in a near horizontal direction.

The technique proposed in this chapter is to propose a solution to the spectrogram filtering problem. The technique combines many strong points in previous works to derive an improved filter. A hybrid 1D and 2D Wiener filter [78] is first used to reduce the noise in the speech spectrogram. This hybrid filter is more effective than the conventional spectral subtraction or 1D Wiener filter. Next the multi-blade post-processor is applied on the residual noise in the noise-only regions. However, unlike [80] and [82] which perform the speech/non-speech decision process using heuristics, a more efficient statistical approach is adopted in this thesis. The proposed speech enhancement system is evaluated using objective measures based on segmental SNR and perceptual evaluation of speech quality as well as subjective measures based on listening tests. The proposed method is compared against two previously reported 2D techniques and it is found that significant improvements can be obtained in most cases.

4.2 Structure of Proposed System

The overall block diagram of the proposed system is illustrated in Figure 4.1. The input noisy speech is transformed by the short-time Fourier transform described in Section 2.3.1 and filtered by hybrid Wiener filter incorporating a 1D Wiener filter and 2D Wiener filter. An image processing technique is performed to further reduce

Figure 4.1: Block diagram of the speech enhancement system in DFT domain

the residual noise and is named as multi-blade filtering. The final resulting speech is obtained by an overlap and add process. With reference to Figure 4.1, the main blocks of this speech enhancement system will be described in detail in the following sections.

4.3 1D & 2D Hybrid Wiener Filter

The 1D & 2D hybrid Wiener filter [78] plays a key role in the removal of additive audible noise. In contrast to the spectral magnitude subtraction in time-frequency domain which ignores the 2D properties, the hybrid Wiener filter exploits the correlations by averaging the coefficients within the surrounding frames. The hybrid filter not only exhibits the 2D effects, but also incorporates the advantage of the 1D Wiener Filter. This combination ensures that both independent and correlated speech signals are maximally retained while as much noise as possible is filtered off.

4.3.1 1D Wiener and Two-step Noise Reduction

1D Wiener filter in Equation (2.13) is widely used and its construction requires the a-priori SNR described in Equation (2.14) which is unknown in practice. The common solution is to utilize the decision-directed approach in Equation (2.17) to estimate the a-priori SNR. However, the well known decision-directed approach [18] has the problem of a one-frame delay due to the high value of α ($\alpha = 0.98$) used in Equation (2.17). That means the Wiener filter which is based on the decision-directed approach actually matches the previous frame rather than the current one. Some aberrations with annoying reverberation effect may be generated. A new approach to the estimation of the a-priori SNR which gives better performance is described in [83], to correct these undesirable effects while preserving the advantages of decision-directed approach. In this approach, a second step described in Equation (4.1) is applied after the first step Equation (2.13). The main effect of the second step is to lessen the one-frame delay bias, and to successfully preserve the speech onsets and offsets. This second step plays a critical role in the reduction of musical noise as well as in providing perceptually better quality enhanced speech, and is

given by:

$$\hat{\xi}(m, \omega)' = \beta \frac{|\hat{S}(m, \omega)|^2}{\lambda_{NN}} + (1 - \beta) \max[\gamma(m + 1, \omega) - 1, 0] \quad (4.1)$$

where m is the time frame index, ω is the frequency index, $\hat{S}(m, \omega)$ is obtained by the standard Wiener filtering described in Equation (2.13), variables λ_{NN} and γ are defined in Equation (2.16) and Equation (2.20) which is the information from the next speech frame. Parameter β has the same effect as α with a fixed value. Unlike $\beta = 1$ used in [83], here β is set to a typical value of 0.98, which is able to obtain a good trade-off between residual noise and speech distortion for human listening.

Equation (4.1) is thus the final version that is applied in the 1D Wiener filter and it has been proven in [83] to be more accurate especially in tracking sharp changes in the speech signal without introducing musical noise. The resulting speech, filtered by the 1D Wiener filter is:

$$\hat{S}_{1D}(m, \omega) = \frac{\hat{\xi}(m, \omega)'}{\hat{\xi}(m, \omega)' + 1} Y(m, \omega) \quad (4.2)$$

4.3.2 2D Wiener Filter

The spectrogram is filtered by the 2D Wiener filter which exploits the 2D properties. In this way, the 1D Fourier transform coefficients are not suppressed independently of their neighbors, unlike the traditional approaches. The use of 2D Wiener filter results in a low residual noise, since it removes the noise mean as well as the noise variation [78].

For the purpose of 2D spectrum filtering, the 2D noise model is introduced first. There are two main parts in this model, namely, the DC component and the AC component:

$$|N(m, \omega)| = N_{AC}(m, \omega) + N_{DC}(m, \omega) \quad (4.3)$$

In fact, the DC component could be expressed as $N_{DC}(m, \omega) = E[|N(m, \omega)|]$. The variance of $N_{AC}(m, \omega)$ is denoted as $\sigma_{N_{AC}}(m, \omega)$ and can be obtained by:

$$\sigma_{N_{AC}}(m, \omega)^2 = E[|N(m, \omega)|^2] - N_{DC}(m, \omega)^2 \quad (4.4)$$

Suppose the noisy speech and the additive noise are known, then the 2D Wiener filter based on the maximum likelihood approach is given as follows using the 2D noise model [78]:

$$\begin{aligned} |\hat{S}_{2D}(m, \omega)| &= \frac{\sigma_Y(m, \omega)^2 - \sigma_{N_{AC}}(m, \omega)^2}{\sigma_Y(m, \omega)^2} \times \left(|Y(m, \omega)| - \overline{|Y(m, \omega)|} \right) \\ &+ \overline{|Y(m, \omega)|} - N_{DC}(m, \omega) \end{aligned} \quad (4.5)$$

where σ indicates the local variance and $\overline{|Y(m, \omega)|}$ is the local means of $|Y(m, \omega)|$. In practice, all the local values could be computed over a 3 by 3 window.

4.3.3 Hybrid Wiener Filter

Based on the discussion of 1D and 2D Wiener filters, the noisy speech spectrogram passes through the hybrid Wiener Filter via two channels: the 1D Wiener filter and the 2D Wiener filter. The result of the hybrid Wiener filtering is taken as the minimum of Equation (4.2) and Equation (4.5).

$$\hat{S}(m, \omega) = \min \left[\hat{S}_{1D}(m, \omega), \hat{S}_{2D}(m, \omega) \right] \quad (4.6)$$

The reason for doing so is that a much quieter enhanced speech is obtained based on the listening tests, which is compared with either the 1D Wiener filtered speech or 2D Wiener filtered speech.

4.4 Classification of Speech and Non-speech Components

After the hybrid Wiener filtering, the low but persistent residual noise manifests itself as some isolated speckles and short ridges. A post-processor is thus needed to further suppress the residual noise. This means that a classification of speech and non-speech components should be implemented before the application of post-processor. No operation will be applied on the classified speech components to avoid distorting the speech, while the classified non-speech components can be confidently suppressed by the post-processor. Another reason for adopting the classification is that computational time will be saved since the post-processor will only affect the non-speech components.

A simple and efficient approach for classification is proposed here to dichotomize the filtered spectrum into speech points and non-speech points. It utilizes a probability estimator by considering the two hypotheses, H_0 and H_1 , which represent speech absence and speech presence respectively. This technique is similar to previous work in [19] and [84]. The probability density functions used to discriminate the noisy magnitude $|Y(m, \omega)|$ into two classes are expressed as:

$$P(Y(m, \omega)|H_0) = \frac{2|Y(m, \omega)|}{\lambda_{NN}} \exp\left(-\frac{|Y(m, \omega)|^2}{\lambda_{NN}}\right) \quad (4.7)$$

and

$$\begin{aligned} P(Y(m, \omega)|H_1) &= \frac{2|Y(m, \omega)|}{\lambda_{NN}} \exp\left(-\frac{|Y(m, \omega)|^2 + |S(m, \omega)|^2}{\lambda_{NN}}\right) \\ &\times I_0\left(\frac{2|Y(m, \omega)||S(m, \omega)|}{\lambda_{NN}}\right) \end{aligned} \quad (4.8)$$

where I_0 is the zero order modified Bessel function. Normally the receiving $Y(m, \omega)$ is considered to belong to the speech part H_1 when $P(Y(m, \omega)|H_1) >$

$P(Y(m, \omega)|H_0)$. For simplification and given that the information of the clean speech is unknown, this discriminant can be rewritten as:

$$e^{-\hat{\xi}(m, \omega)} I_0 \left(2\sqrt{\gamma(m, \omega)\hat{\xi}(m, \omega)} \right) > 1 \quad (4.9)$$

where $\hat{\xi}(m, \omega)$ and $\gamma(m, \omega)$ are the estimated a-priori SNR and the a-posteriori SNR which are defined in Equation (2.17) and Equation (2.20) respectively. If the noisy speech amplitude considered satisfies the inequality, then the corresponding point will be classified into the speech part, else it is included in the non-speech part. After the decision estimation, a binary spectrogram is obtained. A noise reduction technique using 2D 3 (in frequency) $\times 5$ (in time) median filter can be used to obtain an over-suppressed spectrum for future processing. This ensures that the speech points are correctly classified and will be left intact, but the non-speech points will be further processed later.

4.5 Multi-Blade method

It is observed from the spectrograms and identified through listening that the musical tones appear randomly in the spectrum as isolated peaks or short ridges. Moreover, the speech content in the spectrograms exhibit distinguishing features. These observations prompted us to use 2D spectrogram for enhancement.

The spectrogram of a speech signal has distinct characteristics, and it should pay attentions to the different directions relative to a point of interest (POI) for non-speech components. A total of 16 blades are used to indicate the different orientations. The first 6 blades focus on different 7-by-7 orientations with the POI in the center. This is exhibited in the left diagram of Figure 4.2. The other blades are shown in the right diagram which includes five left blades and five right blades.

Figure 4.2: Sixteen blades of the different directions

Along these orientations, the properties are needed to be investigated in order to classify the POI into speech components or non-speech components. So, for each blade $\vec{B}_i$ ($i = 1, \dots, 16$), the variance values of these seven spectrum points are:

$$var(\vec{B}_i) = \frac{1}{7} \sum_{m, \omega \in \vec{B}_i} \left[20 \lg \left(\hat{S}(m, \omega) + 1 \right) \right]^2 - \frac{1}{7} \left[\sum_{m, \omega \in \vec{B}_i} 20 \lg \left(\hat{S}(m, \omega) + 1 \right) \right]^2 \quad (4.10)$$

where m and ω represent the frame index and frequency index. The blade with the minimum variance is selected to update the POI, since it is likely that the point within the blade belongs to the same type of region, e.g. within the valleys of the spectrogram.

Then the spectral value of POI is replaced by the median of the seven points from the blade with the minimum variance under the constraint that the latter is not larger than the original spectral magnitude at POI.

$$\hat{X}(m, \omega) = \min \left\{ \hat{X}(m, \omega), \underset{(i, j) \in \vec{B}_{min(var)}}{median} \left[\hat{X}(i, j) \right] \right\} \quad (4.11)$$

where i and j are the time frame index and frequency index, similar to m and ω . In

this way, the points corresponding to speech which are previously wrongly classified as non-speech components will only be slightly altered, while the noise points are replaced by a median value and is hence further suppressed.

After these post-processing step, the spectrogram undergoes inverse transformation using the phase information of the original noisy speech and the enhanced speech is obtained by the standard overlap & add procedure.

4.6 Results and Discussion

Ten different segments of original speeches, five female and five males from the TIMIT database, are used as samples for evaluation. They are resampled at 8kHz and quantized at 16 bits. Noises with SNR ranging from 0dB to 10dB are added to these clean speeches. The noise type includes white noise, car noise, and F16 aircraft noise.

Three methods are compared, namely: 2DFT proposed by Soon in [78], General Spectral Subtraction with Blade post-processor (GSSB) explained in [82] and the proposed system, hybrid Wiener spectrogram filtering (HWSF), since they are all related 2D filtering methods. In order to show the contributions of the individual components of HWSF, two key points, Two-Step Noise Reduction (TSNR) described in Section 4.3.1 and statistical classification of speech and non-speech components described in Section 4.4, will be evaluated as well. As indicated by previous work [18] and [83], the accuracy of the a-priori SNR is a crucial point for the a-priori SNR based methods. The novel second step, TSNR, is introduced after the decision-directed approach for a better estimation of the a-priori SNR. From the perspective of statistical classification of the speech and non-speech components, it helps a lot to preserve the speech content while further suppressing the residual noise, since the

speech component should be left untouched while the non-speech component is to be processed in the multi-blade post-processing. It is also greatly beneficial in saving computational time. The difference between the multi-blade post-processing with classification and the multi-blade post-processing on the whole spectrogram will be shown in the following objective tests, as well as the effects of the TSNR. The HWSF without TSNR is denoted as HWSF-T, and the HWSF without statistical classification is denoted as HWSF-C.

In order to test this speech enhancement system, two objective measurements, namely, average segmental SNR (SegSNR) and perceptual evaluation of speech quality (PESQ), are utilized. According to [75], various methods for objective measure have been compared. Besides PESQ, SegSNR is the best measure in terms of background noise distortion. PESQ is the best measure for overall speech quality prediction both on the speech quality and noise distortion. The final result is obtained by averaging the results from all the sample speech signals. The performance of the proposed speech enhancement system is further evaluated by informal subjective listening tests.

4.6.1 Comparison of Spectrograms

Since spectrogram filtering is used, the resulting spectrograms of the different methods will give a direct visual comparison for evaluation. An example of clean speech is given in Figure 4.3(a) illustrating the distinguishing properties of speech content while the noisy one is shown in Figure 4.3(b). Figure 4.4(d) shows the resulting spectrograms for white noise at SNR of 5dB of the HWSF algorithm. The other three 2D speech enhancement algorithms namely 2DFT, GSSB and HWSF-C are shown in Figure 4.4(a), Figure 4.4(b) and Figure 4.4(c) respectively. From the com-

Figure 4.3: Spectrograms comparison of clean speech and noisy speech with white noise at SNR=5dB

parison of resulting spectrograms, it can be seen that the residual noise which is musical in nature is better attenuated by the proposed algorithm while more speech content is preserved. Similar results can also be observed with other noise types.

Note that HWSF without TSNR (HWSF-T) is not illustrated in Fig. 4.4. This is because, TSNR gives a more accurate estimation of the a-priori SNR by preserving the onset and offset of the speech [83]. The resultant effect of TSNR is therefore difficult to be observed based on the spectrogram above. It will be evaluated by other measurements in the following sections.

4.6.2 Comparison of Segmental SNR Results

Since SegSNR is better correlated with Mean Opinion Score (MOS) than SNR as indicated by [74] and is easy to implement, it has been widely used to qualify the enhanced speech. In this part, the SegSNR has been implemented in [75] where each frame with segmental SNR is thresholded by a -10 dB lower bound and a 35 dB higher bound. From the evaluation results shown in Tables 4.1, 4.2 and 4.3,

Figure 4.4: Resulting spectrograms comparison of three algorithms

the HWSF algorithm yields significant improvements based on SegSNR measure although the HWSF-C has an even better performance. This can be explained by the fact that the SegSNR is more sensitive to the silent part and the more filtering done in HWSF-C therefore results in a higher score. Over filtering will however cause speech quality distortion.

Table 4.1: SegSNR Results with White Noise

SNR	2DFT	GSSB	HWSF	HWSF-T	HWSF-C
0 dB	1.422	0.978	2.543	1.592	1.561
2 dB	2.144	1.842	3.438	2.439	2.477
4 dB	2.937	2.962	4.232	3.300	3.427
6 dB	3.830	4.059	5.151	4.250	4.197
8 dB	4.793	4.910	6.151	5.135	5.197
10 dB	5.718	6.214	7.086	6.159	6.147

Table 4.2: SegSNR Results with Car Noise

SNR	2DFT	GSSB	HWSF	HWSF-T	HWSF-C
0 dB	9.852	10.001	10.605	9.518	10.807
2 dB	10.903	10.982	11.583	10.567	11.714
4 dB	12.012	12.088	12.648	11.666	12.718
6 dB	13.138	13.166	13.743	12.810	13.753
8 dB	14.337	14.171	14.863	14.036	14.858
10 dB	15.598	15.183	16.083	15.343	16.013

4.6.3 Comparison of PESQ Scores

As mentioned previously, Perceptual Evaluation of Speech Quality (PESQ) is an objective measurement tool that predicts the results of subjective listening tests. It has been proven to be more reliable and correlated with the Mean Opinion Score (MOS) in contrast to four widely used objective measures, the segmental SNR measure (SegSNR), weighted-slope spectral (WSS) distance, log likelihood ratio (LLR) and Itakura-Saito (IS) distance [75]. The overall results are shown in Tables 4.4, 4.5 and 4.6. The HWSF scores are higher than the other methods under different types of noise with input SNR ranging from 0dB to 10dB. This observation confirms the efficiency and worthiness of the proposed approach.

Table 4.3: SegSNR Results with F16 Aircraft Noise

SNR (dB)	2DFT	GSSB	HWSF	HWSF-T	HWSF-C
0 dB	0.850	0.290	1.534	0.720	1.798
2 dB	1.645	1.127	2.372	1.502	2.518
4 dB	2.471	2.162	3.265	2.332	3.385
6 dB	3.360	3.172	4.227	3.319	4.310
8 dB	4.347	4.256	5.287	4.392	5.345
10 dB	5.433	5.483	6.420	5.574	6.449

Table 4.4: PESQ Results with White Noise

SNR	2DFT	GSSB	HWSF	HWSF-T	HWSF-C
0 dB	2.066	2.080	2.132	2.066	2.087
2 dB	2.217	2.236	2.293	2.234	2.193
4 dB	2.412	2.392	2.445	2.372	2.362
6 dB	2.551	2.533	2.615	2.466	2.510
8 dB	2.686	2.687	2.741	2.647	2.611
10 dB	2.791	2.817	2.893	2.756	2.691

4.6.4 Comparison of Subjective Measure Results

To further evaluate the quality of the enhanced speech signals, and to reaffirm the objective evaluation results, informal subjective tests are also conducted. Ten speech utterances which are corrupted by white noise with SNR of 2 dB are used in the subjective test. Ten listeners, five females and five males, are asked to give a score between one and five to evaluate the enhanced speech quality [1=bad, 2=poor, 3=fair, 4=good, 5=excellent]. The experiment is conducted in the following manner for each of the ten utterances:

- The clean speech and the noisy speech are both played twice to form a reference for the listener.

Table 4.5: PESQ Results with Car Noise

SNR	2DFT	GSSB	HWSF	HWSF-T	HWSF-C
0 dB	3.492	3.547	3.621	3.547	3.531
2 dB	3.600	3.655	3.719	3.654	3.632
4 dB	3.705	3.739	3.811	3.765	3.730
6 dB	3.804	3.814	3.896	3.873	3.828
8 dB	3.903	3.882	3.987	3.967	3.942
10 dB	3.992	3.972	4.077	4.062	4.043

Table 4.6: PESQ Results with F16 Aircraft Noise

SNR (dB)	2DFT	GSSB	HWSF	HWSF-T	HWSF-C
0 dB	2.064	2.047	2.129	1.906	1.921
2 dB	2.247	2.235	2.317	2.106	2.103
4 dB	2.411	2.407	2.498	2.270	2.281
6 dB	2.537	2.569	2.668	2.462	2.492
8 dB	2.687	2.724	2.815	2.646	2.671
10 dB	2.824	2.889	2.957	2.820	2.840

- Each of the enhanced speech by each of the three different algorithms is repeated thrice randomly. In other words, a total of nine (3 repeats $\times$ 3 algorithms) enhanced utterances are being played out in a random order for each speech utterance.
- For each payout, the listener will listen to a repeat before giving a score.
- The earlier steps are repeated for all the ten speech utterances.

The overall MOS result, shown in Table 4.7, is therefore the average of 300 scores (10 listeners $\times$ 10 speech utterances $\times$ 3 random payouts). From the reports of the volunteer listeners, the enhanced speech signals obtained by using the proposed algorithm yield a better perceptual quality since the musical tones in the background have been suppressed significantly. The subjective listening test results vindicate

Table 4.7: Results of Subjective Measure and Objective Measure with White Noise at SNR=2dB

Algorithms	MOS	SegSNR	PESQ
2DFT	2.68	2.14 dB	2.22
GSSB	2.50	1.84 dB	2.24
HWSF	3.05	3.44 dB	2.29

the SegSNR and PESQ evaluation results that the HWSF performs the best. The marked reduction in musical tone is considered as the major improvement in garnering higher MOS from the listeners. This shows the effectiveness of the two-step noise reduction process introduced in the hybrid Wiener filter.

4.7 Conclusion

In this chapter, a speech enhancement system is proposed based on hybrid Wiener filter spectrogram filtering. The algorithm essentially comprises two main parts, namely, the hybrid Wiener filter and the multi-blade post-processor. The hybrid Wiener filter not only functions as a 1D Wiener filter, but also considers the correlations around the 2D neighborhoods. It is thus more effective in noise removal. The hybrid Wiener filter, incorporating a novel two-step noise reduction algorithm is especially effective in musical noise suppression. Nevertheless, despite the low residual noise, post-processing is further performed for an even better outcome. The features of speech and residual noise in the 2D spectrogram are exploited to discriminate speech from noise components to facilitate effective noise removal without compromising the intelligibility of the speech. Furthermore, a multi-blade approach is applied to recover those speech content which has been erroneously classified as noise, to further preserve the integrity of the speech signals.

The proposed approach is compared to other related methods using objective SegSNR and PESQ measures as well as informal subjective listening tests. Its SegSNR and PESQ results are superior to other 2D techniques, namely, 2DFT and GSSB. All the objective measure results are further vindicated by the subjective listening test results which put the proposed algorithm above the others in terms of overall perceptual quality of speech and musical noise reduction. It also presents a more pleasant listening experience to the testers.

Chapter 5

A Spectral Post-processing Technique

Despite the quality improvement of the speech signal with most traditional noise reduction (TNR) algorithms in Discrete Fourier Transform (DFT) domain, the output is always distorted to some extent due to the over-attenuation of speech components. Weak speech components are usually regarded as noise in noise reduction processing and are therefore highly suppressed. In this chapter, a novel post-processing technique will be introduced to reduce this problem.

5.1 Introduction

The goal of speech enhancement is to remove noise while preserving the clean speech as much as possible. Many algorithms have been proposed for that, such as spectral subtraction (SS) [17], minimum-mean square error (MMSE) estimator [18, 85] and Wiener filter based algorithms [33, 78, 86]. These algorithms are all designed as an optimal filter to balance the speech distortion and noise reduction. They are

denoted as traditional noise reduction (TNR) filters since they share the common shortcoming in that some distortions in speech components are unavoidable. The weak components, especially those at higher frequencies, are often eliminated. To address this inherent problem in TNR filter, two algorithms have been proposed recently, with both focusing on harmonic component regeneration to reduce the speech distortion caused by the TNR. Harmonic regeneration noise reduction filter (HRNR), is proposed in [87] and [83]. Its gain varies according to the speech harmonics and thus gives a richer harmonics output as compared to the TNR filter. With the regenerated harmonic component, HRNR can noticeably improve the voiced speech. However, there is no attempt to recover the unvoiced speech in HRNR, since the standard procedure is applied to the unvoiced part. Hence the distortion remains, especially in speech with a large portion of unvoiced components. Another recently proposed algorithm, a post-processing technique for high-frequency regeneration (PHFR) [88], uses a blind approach to synthesize the missing harmonics. It is applied as a post-processing step which contributes to the high frequency portion of the recovered speech. This work is motivated by the bandwidth recovery schemes proposed to transform narrow-band speech into wide-band speech since there is a similar relationship between TNR filtered speech versus clean speech and narrow-band speech versus wide-band speech. The bandwidth recovery scheme attempts to transform narrow-band speech to wide-band speech by adding synthetic high frequency speech signal to 4kHz sampled signal to make it sound like an 8kHz signal. The high frequency content is synthesized by extending the spectral envelop and generating an excitation signal. Although the PHFR algorithm deals with both the voiced and unvoiced speech components, the truncated low frequency speech incorporating artificially regenerated high frequency component sounds less natural.

Despite the drawback of the variable speech components attenuation in the TNR

filters, these filters are capable of reducing the background noise. Hence it is natural to design a post-processing technique that is able to regenerate the lost speech components and add them back to the TNR filter output. This combination can maintain the advantage of the TNR filters while addressing its shortcoming efficiently, yielding a better result. In this chapter, an over-attenuated speech components regeneration (OSCR) algorithm is proposed for the recovery of the lost speech components. Unlike the methods in [87] and [83] which only regenerate harmonic components for voiced speech, the proposed algorithm can recover both the voiced and unvoiced speech in the entire frequency domain. The quality of the resulting speech will be evaluated by some objective measures as well as informal subjective test.

5.2 Structure of Proposed System

The main idea of proposed OSCR algorithm is based on the well-known source-filter structure of speech generation model. It is assumed that the human vocal tract can be modeled by an auto-regressive filter, which is excited by a spectrally flat excitation signal. Hence the proposed algorithm is made up of two parts, spectral envelop estimation and the excitation signal generation. This proposed procedure can be viewed from the perspective of Hilbert transform. Unlike Fourier transform which decomposes signals into a sum of sinusoids of different frequencies, Hilbert transform splits the signal into a product of a rapidly varying fine structure and a slowly varying envelop. The first signal can be considered as the "carrier" signal and the latter as the "modulating" signal. A classic example is for the modulator to be a voiced signal and the carrier to be a spectrally rich sound. Similar to this idea, the two key components of the regeneration process, namely, the excitation signal

Figure 5.1: Block diagram of the proposed speech enhancement system

and the estimated spectral envelop can be developed independently.

The noisy speech is first filtered by a TNR filter such as the Wiener filter which is adopted in this chapter. After that, the proposed post-processing technique is launched: a non-linear transformation is first applied to the Wiener filtered speech and the transformed signal is multiplied by a pre-estimated spectral envelop to form the regenerated speech. The resulting synthesized speech is then obtained using a weighted combination of the regenerated speech components and the Wiener filtered speech. The whole block diagram is shown in Figure 5.1.

5.3 Traditional Noise Reduction Filter

In the additive noise model, the noisy speech $y(t)$ is the result of clean speech $s(t)$ corrupted by uncorrelated noise signal $n(t)$. It is given by Equation (2.1) in the time

domain. Noisy speech is the only available information in a single channel speech enhancement problem. In order to obtain the clean speech from the noisy one, the TNR filter $g(t)$ is designed to output an enhanced speech $\hat{s}(t)$ with minimum bias from the clean speech. The enhanced speech is expressed as

$$\hat{s}(t) = y(t) * g(t) \quad (5.1)$$

where $*$ is the convolution operator. Then the residual noise signal $\epsilon(t)$ is defined as the difference between the clean and the enhanced speech. Incorporating Equations (2.1) and (5.1), the residual noise signal can be expressed as

$$\epsilon(t) = \hat{s}(t) - s(t) = [g(t) - 1] * s(t) + g(t) * n(t) \quad (5.2)$$

The convolution process in the time domain can be converted to a multiplication process in the frequency domain. Then Equations (5.1) and (5.2) are converted into frequency expressions in Equations (5.3) and (5.4).

$$\hat{S}(\omega) = Y(\omega) \times G(\omega) \quad (5.3)$$

$$\epsilon(\omega) = \hat{S}(\omega) - S(\omega) = \epsilon_S(\omega) + \epsilon_N(\omega) \quad (5.4)$$

where ω is the frequency index. $\epsilon(\omega)$, $G(\omega)$, $S(\omega)$, $N(\omega)$ and $Y(\omega)$ are the frequency expressions of $\epsilon(t)$, $g(t)$, $s(t)$, $n(t)$ and $y(t)$, respectively. Spectral speech distortion $\epsilon_S(\omega)$ and spectral noise distortion $\epsilon_N(\omega)$, which have been defined in previous work [28, 89], are obtained from Equations (5.2) and (5.4) as

$$\epsilon_S(\omega) = [G(\omega) - 1]S(\omega) \quad (5.5)$$

$$\epsilon_N(\omega) = G(\omega)N(\omega) \quad (5.6)$$

From Equations (5.4), (5.5) and (5.6), it can be seen that filter function G has a significant impact on speech components with signal energy similar to the noise

energy. This is because G affects such speech components and noise equally. The higher the degree of noise suppression, the higher is the amount of speech distortion. Unfortunately most of the TNR filters share this common trait since they can all be represented as a gain function, for example, from Equations (2.7), (2.18) and (2.13), the gain functions of SS, MMSE and Wiener filter are given as

$$G_{SS} = \sqrt{\max\left(0, 1 - \frac{1}{\gamma}\right)} \quad (5.7)$$

$$G_{MMSE} = \frac{\sqrt{\pi}}{2} \sqrt{\frac{\hat{\xi}}{(1 + \hat{\xi})\gamma}} M\left[-0.5; 1; -\frac{\hat{\xi}}{1 + \hat{\xi}}\gamma\right] \quad (5.8)$$

$$G_{Wiener} = \frac{\hat{\xi}}{1 + \hat{\xi}} \quad (5.9)$$

where frequency index ω is omitted for convenience, $M[\cdot]$ is the confluent hypergeometric function which can be computed by Equation (2.21). γ and $\hat{\xi}$ are the a-posteriori signal-to-noise ratio (SNR) and the estimated a-priori SNR defined in Equation (2.20) and Equation (2.17) (by decision-directed approach [18]) respectively.

Since the HRNR method utilizes a gain controller and Wiener filter and both are based on a two-step noise reduction (TSNR) technique [83] which has been described in Section 4.3.1, the same technique is also adopted here so that a fair comparison can be made. The TSNR technique is an improved means to estimate the a-priori SNR which is used to correct the one-frame delay problem. It exhibits a better performance than decision-directed approach for a-priori SNR based speech enhancement algorithms [83]. This advanced second step can be obtained by Equation (4.1) which gives a better estimate $\hat{\xi}(m, \omega)'$ where m is the time frame index. Then the enhanced speech $\hat{S}(m, \omega)$ can be obtained by Equation (4.2).

It should be noted that the TNR filter possesses some shortcomings. From Equations (5.7) to (5.9), it can be seen that the gain will reach the asymptotic

values of 0 or 1 when the SNR is 0 or infinity (Note that SNR discussed here is without the log operation). However it is extremely unlikely that the gain will ever reach the value of 1. Therefore, even if noise is not present in a particular frequency coefficient at a particular time frame, it will still be attenuated. This is especially true for speech components that are well below the noise level. They are suppressed along with the noise by the TNR filtering. For the rest of the chapter, the widely used Wiener filter is selected as the TNR filter as it is simple and yet reasonably effective. However the proposed post-processor can be applied to any TNR filter since it is an independent step.

5.4 Excitation Signal

There are many ways to obtain the excitation signal in the literature, such as non-linear transform, over-sampling excitation, spectral duplication and Gaussian modulation, etc. They are all able to produce an abundance of harmonics from one with less harmonics, i.e. the TNR filtered speech. Owing to the phase coherence and simplicity consideration, a time domain method is chosen for the generation of the excitation signal in this chapter. According to [90], the non-linear transform is preferred over over-sampling excitation since the latter causes spectrum folding by aliasing resulting in some artificial tones being generated. The waveform rectifier is the most popular method of obtaining the excitation signal amongst all the non-linear mapping methods. However it should be noted that in [87] a half-wave rectifier is used instead. The output signal of the waveform rectifier has a full spectrum and exhibits phase coherency with the input signal. All these properties are favorable and the process can be mathematically represented by the following expression with

input TNR filtered speech $\hat{s}(t)$ and output speech $x(t)$:

$$x(t) = \frac{1}{2} [(1 + \beta)|\hat{s}(t)| + (1 - \beta)\hat{s}(t)] \quad (5.10)$$

where $\beta = 0$ corresponds to the half-wave rectifier and $\beta = 1$ corresponds to the full-wave rectifier. $\beta = 0.5$ has also been used in the past [91]. The value of β does not have much effect on the output speech as reported in [92].

The full-wave rectifier is preferred notwithstanding the fact that Makhoul and Berouti have reported that performances of the various types of rectifiers do not differ very much [92]. A full-wave rectifier converts the entire input waveform to one of constant polarity, while a half-wave rectifier only allows either the positive or negative half of the AC wave to pass through. In the formant frequency sense, the regenerated harmonics are completely determined by $|\hat{s}(t)|$. From Equation (5.10), full-wave rectified signal $x_{full}(t)$ equals to $|\hat{s}(t)|$ and the half-wave rectified signal $x_{half}(t)$ will be expressed as

$$x_{half}(t) = \frac{1}{2} [x_{full}(t) + \hat{s}(t)] \quad (5.11)$$

This relationship still holds in the frequency domain. In [83], the half-wave rectified signal is proven to be a convolution between the spectrum of $\hat{s}(t)$ and an harmonic comb which has the same fundamental frequency and desired harmonics. Thus it is clear from Equation (5.11) that the full-wave rectified signal contains more harmonics while the half-wave rectified signal includes the fundamental frequency which is contributed by the part, $\hat{s}(t)$. As the regenerated harmonics are the main interest and there is no necessity to further emphasize the fundamental frequency, the full-wave rectifier is thus more suitable for usage, and it gives a slightly better performance over the half-wave rectifier in the experiments.

In conjunction, a whitening filter is implemented to flatten the spectrum of excitation signal so that the regenerated speech can be made to closely follow the

shape of the estimated spectral envelop signal. In the Linear Predictive Coding (LPC) analysis, the signal can be divided into a spectrally flat excitation signal and an all-pole synthesis filter comprising its LPC coefficients. Therefore, the excitation signal $x(t)$ is fed into the synthesis filter to obtain the whitened excitation signal $\hat{X}(m, \omega)$ in the time-frequency domain. The LPC order is set to ten in this chapter.

5.5 Spectral Envelop Estimation

The SS (Spectral Subtraction) filtered speech is selected to implement the spectral envelop estimation since it is simple and relatively straight-forward. On the other hand, the a-priori SNR based methods including the Wiener filter and MMSE estimator are not accurate since they have a one-frame-delay problem due to the decision-directed approach which has been reported in [93]. To extract an approximate coarse spectral envelop which bears some resemblance to that of the clean speech, the estimated spectral envelop $C(m, \omega)$ of time frame m is given as

$$C(m, \omega) = \sqrt{\max[0, |Y(m, \omega)|^2 - \lambda_{NN}]} * H(\omega) \quad (5.12)$$

where $H(\omega)$ is a low-pass finite impulse response (FIR) filter and the stop frequency is around 150Hz for optimal performance. This is determined empirically.

The estimated spectral envelop is actually a highly smoothed version of the speech filtered by the SS method. That means the regeneration process for over-attenuated speech components is mainly contributed by the low-pass filter $H(\omega)$. The effect of this low-pass filtering is to reduce the chances of occasional mismatches, to avoid overly rapid changes and to regenerate some weak portions whose values have dropped to zero in the SS filtered spectrum. Among all these advantages, the regeneration effect is the most important for the proposed algorithm. This is clearly

Figure 5.2: Voiced spectrum of (a) Clean speech (b) Wiener filtered speech (c) SS filtered speech (the latter two are with input white noise at 5dB); Spectral Envelop of (d) Clean speech smoothed by $H(\omega)$ (e) Wiener filtered speech smoothed by $H(\omega)$ (f) Proposed method $C(m, \omega)$

Figure 5.3: Unvoiced spectrum of (a) Clean speech (b) Wiener filtered speech (c) SS filtered speech (the latter two are with input white noise at 5dB); Spectral Envelop of (d) Clean speech smoothed by $H(\omega)$ (e) Wiener filtered speech smoothed by $H(\omega)$ (f) Proposed method $C(m, \omega)$

seen by the comparisons shown in Figures 5.2 and 5.3, the former is for the voiced speech and the latter is for the unvoiced component. In plots (b) of Figures 5.2 and 5.3, the TNR filtered speech which is Wiener filtered here obviously experienced an energy loss and is seriously suppressed in the high frequency portion in contrast to the SS filtered speech (see plots (c)), as well as the clean speech (see plots (a)). From plots (f) of Figures 5.2 and 5.3, the respective estimated spectral envelopes are thus able to successfully recover some of the over-attenuated components from both the voiced and unvoiced parts. The estimated spectral envelopes are quite close to their clean speech counterparts (see plots (d)); the latter being an ideal realization, serving as a reference. Note that the main focus is just to strengthen the over-attenuated components which in this case are the high frequency peaks captured within the envelop. Of course, the significant recovery as seen from plots (b) to (f) in Figures 5.2 and 5.3 cannot be achieved just by a smoothing process $H(\omega)$. It is also contributed by the SS filtered speech (see plots (c)) which has been less suppressed speech components as compared to the Wiener filtered ones shown in plots (b) of Figures 5.2 and 5.3. Thus, the proposed algorithm is good for Wiener filter which is efficient at noise reduction but at the same time, incurs severe distortion of the speech signal. As for the SS filtered speech, the effects of the proposed method will be less significant.

A smoother and more continuous envelope is generated after passing through a low-pass filter. In summary, envelop estimation is certainly feasible for the regeneration of the voiced and unvoiced speech.

5.6 OSCR Filtered Speech

From the conclusions of the previous two sections, the regenerated speech components are the product of the excitation and the estimated envelop, and can be expressed as

$$R(m, \omega) = \hat{X}(m, \omega) \times C(m, \omega) \quad (5.13)$$

The shortcoming of the TNR filtered speech can be partially reversed by incorporating the adaptively weighted portion of the artificially regenerated speech components into the TNR filtered speech. A gain function $\nu(m)$ should be present to control the amount of the various speech components to be added back in each time frame. The gain function is estimated by

$$\nu(m) = \frac{\overline{|Y(m, \omega)|^2} - \overline{\lambda_{NN}} - \overline{|\hat{S}(m, \omega)|^2}}{\overline{|R(m, \omega)|^2}} \quad (5.14)$$

where λ_{NN} is the one defined in Equation (2.16), $\overline{(\cdot)}$ is the mean function and it takes the mean value along the frequency direction in Equation (5.14) to get the mean energy in each speech frame. It is clear that the gain function $\nu(m)$ is time frame dependent and can either be positive or negative. With stationary background noise corruption, two situations of $\nu(m)$ appear:

- $\nu(m) > 0$, which means $\overline{|Y(m, \omega)|^2} > \overline{\lambda_{NN}} + \overline{|\hat{S}(m, \omega)|^2}$, and mean energy of estimated speech frame is over-suppressed. $\implies$ Energy compensation is needed.
- $\nu(m) < 0$, which means $\overline{|Y(m, \omega)|^2} < \overline{\lambda_{NN}} + \overline{|\hat{S}(m, \omega)|^2}$, and mean energy of estimated speech frame is higher. $\implies$ Further suppression is needed.

By this gain function, the regenerated speech components are adjusted adaptively to the needs of the TNR filtered speech in each frame, and is suitable for the "adding" operation for the final synthesized speech.

Therefore, the OSCR filtered speech of frame m is given by the following expression

$$\tilde{S}(m, \omega) = e^{j\theta_Y(m, \omega)} \sqrt{|\hat{S}(m, \omega)|^2 + \nu(m)|R(m, \omega)|^2} \quad (5.15)$$

where $\theta_Y(m, \omega)$ represents the phase information of noisy speech $Y(m, \omega)$. From the above equation, it can be seen that the regeneration process operates on the entire frequency spectrum unlike the method in [88] which combines the low-passed filtered speech with synthesized high frequency components. The fully restored speech harmonics are produced by the excitation signal, which has been shown to achieve good performance in [87]. Furthermore, some weak components, especially the unvoiced speech, can be recovered by the estimated envelop $C(m, \omega)$ since a flattened and continuous envelop is extracted by Equation (5.12). Thus the synthesized speech $\tilde{S}(m, \omega)$ obtained from Equation (5.15) enjoys good noise reduction with lower speech distortion both on the voiced and the unvoiced speech components.

Inverse Fourier transform and the overlap & add process should be applied on the synthesized speech $\tilde{S}(m, \omega)$ to produce the restored speech $\tilde{s}(t)$ as the final output signal.

5.7 Results and Discussion

Ten different segments of speeches, five females and five males are randomly selected from the TIMIT database. They are resampled at 8 kHz and quantized at 16 bits, and are corrupted by four types of noises, namely, white noise, fan noise, car noise and F16 aircraft noise. All these speech segments have silence periods at the beginning and at the end. Four methods are compared in this section: Wiener filter (WF) using Equation (4.2) as the representative of the TNR filter, PHFR [88] which

is a post-processing technique for high-frequency regeneration, HRNR [83] which is a method for harmonic regeneration noise reduction and the proposed OSCR algorithm. Except for the WF which is the traditional type, the last three methods focus on addressing the over-attenuation problem. In the following sections, a deep insight of the performance of all these speech enhancement methods will be provided through various objective and subjective evaluations.

5.7.1 Voiced and Unvoiced Components Regeneration

Three similar enhancement methods, PHFR, HRNR and OSCR, are all devoted to regenerating the over-attenuated speech components in conjunction with TNR filtering. The regeneration process can be divided into two aspects, for voiced components and for unvoiced components, of which the latter is quite difficult since the unvoiced components normally occur in the high frequency band and is usually severely suppressed by TNR filters. HRNR method yields an impressive performance on harmonic restoration in contrast to WF filtered speech in [83] while the unvoiced parts are not addressed. It just ensures that the unvoiced parts of speech are not being degraded by the harmonic regeneration process. PHFR is able to recover both the voiced and unvoiced sounds in principle although it has not been explicitly mentioned in the article. It still suffers from low frequency distortion and possibly unfavorable artifacts on the high frequency band. The aim of OSCR is to perform the recovery on the entire frequency domain including the voiced and unvoiced components. The ability of regeneration can be tested and measured using the average log-spectral distance (LSD).

LSD has been used as the objective measure in previous work [88, 94]. It is a distance measure between two spectral envelopes which is expressed by the following

equation

$$LSD = \frac{1}{M} \sum_{m=1}^M \sqrt{\frac{1}{\omega_s} \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} \left[\ln \frac{|S(m, \omega)|}{|\tilde{S}(m, \omega)|} \right]^2 d\omega} \quad (5.16)$$

where ω_s is the sampling frequency. Although the cepstral distance (CD) measure which is adopted in [87] can be viewed as an accurate replacement for the direct root-mean-square measure between two log spectra but without Fast Fourier Transform (FFT) or logarithm operations, the LSD and CD are actually the same indeed. LSD is chosen for convenience and also because it is more correlated with the spectrum. A smaller distance indicates a better correlation between the two candidates in comparison.

LSD is also chosen for voiced/unvoiced performance validation. For the silence period, the data amplitudes are very small and the LSD results will be meaningless. Fortunately the silence speech frames only occupy 7.1% in the testing speech fragments and can be omitted. A pitch detection technique based on the auto-correlation method is introduced to give a unvoiced/voiced decision for each spectral frame. Therefore the LSD results are presented for both the unvoiced and voiced classes shown in Tables 5.1 and 5.2 respectively and the overall LSD results are shown in Table 5.3. For the unvoiced parts, OSCR improves the quality with flying colors, while PHFR results show slight variations from the WF results which means PHFR is more effective for filters with severe suppression on the high frequency portions. HRNR yields an even worse result which is confined to just the peak values in the CD results shown in paper [87, 95].

For the voiced parts, OSCR also exhibits good ability in improving the noisy speech quality. HRNR is as good and is superior under the car noise scenario due to its good control on introducing additional harmonics. In contrast, the OSCR emphasizes relatively more on the high frequency portion although it acts on the entire frequency band. Hence its advantages are less obvious in low frequency noise

Table 5.1: Mean log-spectral distance comparison for unvoiced parts

Noise type	SNR (dB)	Mean LSD Value			
		WF	PHFR	HRNR	OSCR
White	-10	2.63	2.67	4.15	2.16
	-5	2.64	2.70	3.82	1.96
	0	2.61	2.66	3.47	1.80
	5	2.52	2.57	3.06	1.67
	10	2.39	2.44	2.66	1.57
Fan	-10	2.27	2.34	2.62	1.72
	-5	2.11	2.18	2.27	1.63
	0	1.93	2.0	1.94	1.52
	5	1.68	1.76	1.62	1.35
	10	1.41	1.52	1.33	1.16
Car	-10	1.88	1.95	1.87	1.46
	-5	1.64	1.74	1.54	1.31
	0	1.42	1.54	1.28	1.17
	5	1.18	1.33	1.06	1.01
	10	0.98	1.18	0.88	0.88
F16 aircraft	-10	2.81	2.87	3.92	2.12
	-5	2.71	2.76	3.46	1.99
	0	2.54	2.57	2.98	1.88
	5	2.31	2.35	2.48	1.74
	10	2.08	2.11	2.06	1.58

situations. In summary, OSCR does an excellent job on the entire speech spectrum with the unvoiced portions being well operated upon and the voiced portions well maintained.

The outstanding ability of OSCR algorithm to recover the unvoiced components can be further vindicated by evaluating the effectiveness of the envelop estimation process. The reason is that PHFR, HRNR and OSCR all utilize the non-linear transform for excitation signal while OSCR adopts an extra smoothing envelop to guide the excitation signal. The smoothing process is effective as it restores the

Table 5.2: Mean log-spectral distance comparison for voiced parts

Noise type	SNR (dB)	Mean LSD Value			
		WF	PHFR	HRNR	OSCR
White	-10	2.57	2.60	3.48	1.93
	-5	2.37	2.38	2.83	1.73
	0	2.19	2.19	2.30	1.55
	5	2.0	1.99	1.92	1.39
	10	1.79	1.79	1.61	1.25
Fan	-10	1.88	1.91	1.76	1.55
	-5	1.62	1.65	1.46	1.39
	0	1.35	1.41	1.21	1.19
	5	1.08	1.20	1.0	0.98
	10	0.84	1.03	0.82	0.78
Car	-10	1.24	1.36	1.09	1.12
	-5	1.02	1.18	0.90	0.94
	0	0.82	1.05	0.77	0.78
	5	0.65	0.94	0.67	0.63
	10	0.50	0.86	0.59	0.50
F16 aircraft	-10	2.83	2.91	3.80	1.88
	-5	2.45	2.49	2.86	1.73
	0	2.07	2.09	2.13	1.56
	5	1.77	1.78	1.62	1.37
	10	1.46	1.51	1.28	1.18

over-attenuated parts shown in Figure 5.2 and 5.3, thus resulting in an estimated spectral envelop that is closer to the clean one. The second advantage of OSCR can be shown by studying the correlation factors between the spectra of the clean speech, the noisy speech and the SS filtered speech, against those of their respective smoothing envelopes. Figure 5.4 shows the results of the correlation factors which are calculated from the cross-correlation between the clean spectrum and certain testing spectrum. The index *Noisy* indicates the cross-correlation between the clean spectrum and noisy spectrum. *SS* indicates the cross-correlation between the clean

Table 5.3: Overall mean log-spectral distance comparison

Noise type	SNR (dB)	Mean LSD Value			
		WF	PHFR	HRNR	OSCR/CE _{Env}
White	-10	2.50	2.51	3.59	2.06/1.76
	-5	2.41	2.43	3.12	1.85/1.67
	0	2.32	2.33	2.74	1.66/1.59
	5	2.20	2.21	2.40	1.52/1.51
	10	2.06	2.08	2.08	1.41/1.39
Fan	-10	2.0	2.03	2.06	1.65/1.64
	-5	1.81	1.83	1.77	1.52/1.51
	0	1.61	1.65	1.53	1.363/1.359
	5	1.38	1.46	1.31	1.17/1.168
	10	1.14	1.28	1.11	0.982/0.980
Car	-10	1.52	1.60	1.43	1.30/1.29
	-5	1.30	1.43	1.22	1.12/1.11
	0	1.10	1.28	1.06	0.96/0.95
	5	0.91	1.14	0.92	0.812/0.807
	10	0.75	1.04	0.81	0.69/0.686
F16 aircraft	-10	2.70	2.75	3.68	1.98/1.88
	-5	2.47	2.50	3.0	1.83/1.76
	0	2.21	2.23	2.42	1.67/1.61
	5	1.97	1.99	1.96	1.52/1.47
	10	1.74	1.77	1.63	1.35/1.32

spectrum and the SS filtered speech. All these spectra are not smoothed by the low-pass filter $H(\omega)$. $EnvN$ indicates cross-correlation between the clean envelop and noisy envelop. $EnvSS$ indicates cross-correlation between the clean envelop and estimated envelop given by $C(m, \omega)$ in Equation (5.12). All these envelops mentioned above with Env are obtained by the low-pass filtering process. Before the low-pass filtering, the noisy speech spectrum is more correlated with the clean speech than the SS filtered spectrum. However the smoothed spectrum of SS filtered speech, i.e. the estimated envelop $C(m, \omega)$, is highly correlated with the clean

Figure 5.4: Correlation factors of *Noisy*, *SS*, *EnvN* and *EnvSS* under different noise types.

spectrum smoothed by $H(m, \omega)$ and it approaches the upper bound value (i.e. the ideal case represented by that of the clean speech). Therefore the envelop acts as an accurate guide for the excitation signal to result in good regenerated speech components.

A scheme, named CEnv, is also tested which utilizes the clean envelop instead of the estimated envelop $C(m, \omega)$ in the OSCR system. The results are shown in Table 5.3 and Figure 5.5. The results show that OSCR with the estimated envelop is quite close to OSCR with the clean envelop for narrow band noise scenarios (i.e. fan and car noises). As for the wide band noise scenarios, the relatively large gap

Figure 5.5: Overall PESQ results comparison

between the OSCR and CEnv plots heralds the potential further work needed to improve the envelop estimation beyond that used in this algorithm so as to further improve the performance of OSCR in such noise scenarios.

5.7.2 Overall Speech Quality

According to [96], the perceptual evaluation of speech quality (PESQ) measure has been proven to be more reliable than some traditional objective measures, such as segmental SNR measure (SegSNR), log likelihood ratio (LLR) and Itakura-Saito (IS) distance measure, etc. It has been tested in speech distortion aspect where LLR

performs the best followed by PESQ measure, while in the noise distortion aspect, PESQ measure performs the best. Therefore, PESQ measure is considered to be an excellent objective measure tool for predicting the overall quality of enhanced speech.

The PESQ results of different speech enhancement algorithms are plotted in Figure 5.5. The result for CEnv (using clean envelope) is also included as an effective indicator of the upper bound for OSCR. It can be seen from Figure 5.5 that, OSCR scores are higher than the other three methods under the different noise types. The improvements of PHFR are not significant and it might be due to the noise type. When the noise is mainly concentrated in the high frequencies such as F16 aircraft noise, PHFR works well since the effectiveness of its high frequency regeneration improves. HRNR does not work as well when the input SNR is very low which is similar to previous results while it is quite good at low frequency noise conditions. This can be explained by the fact that the output signal is mainly dependent on the 'artificial signal' when the SNR is very low. This 'artificial signal' is obtained nonlinearly without any other constraints imposed. This may result in some arbitrary parts in the output signal yielding poor results. HRNR will work better when the input SNR is high. The obvious improvements in low input SNR conditions indicate that OSCR works very well under severe noise disturbance. In conclusion, the PESQ results reaffirm the previous observations that, the OSCR is superior to the other three methods.

Informal subjective tests are conducted as well and the same conclusion is obtained. The well-known residual noise which is musical in nature is rarely heard. The main reason is due to the masking effect provided by the regenerated speech components in OSCR. Therefore the resulting speech of OSCR sounds far richer since more high frequency contents have been reinstated.

5.8 Conclusion

Despite the success of TNR filters, the filtered speech still suffer from undesirable speech distortion when the weak speech components are removed together with noise. In this chapter, a speech enhancement algorithm OSCR has been proposed to correct for the over-attenuation problem to yield high quality speech signal based on the regeneration of both voiced and unvoiced speech components. For the voiced sounds, a full-wave rectifier is applied to efficiently restore the degraded harmonics. For the unvoiced parts, it is mainly contributed by the envelop estimation which actually acts on all the frequency bins and it is not just confined to the unvoiced components. The experimental results confirmed that OSCR works well in comparison with WF, PHFR and HRNR methods. To further demonstrate the effectiveness of proposed algorithm, it is evaluated on two common objective measures, LSD measure for testing voiced and unvoiced components regeneration, and PESQ measure for the overall quality of enhanced speech. The results clearly show the excellent performance of the proposed algorithm. Informal subjective listening tests also show that OSCR filtered speech sounds far more natural and richer since residual musical noise has been masked during the regeneration process and unvoiced components have been partially restored.

Chapter 6

Speech and Noise Distortion Measures

So far some proposed speech enhancement systems have been introduced together with their objective evaluation results. Although the evaluating methods are complementary as opposed to the aim of showcasing the performance of the speech enhancement system, detailed discussions are still needed so that they can be characterized for different applications or requirements.

6.1 Introduction

Although historically the majority of assessment methods have been developed to evaluate speech coding algorithms, these methods have also been used to quantify the performance of speech enhancement algorithms [97].

Quality assessment can be done using subjective listening tests or objective quality measures. Subjective evaluation involves comparisons of original and processed

speech signals by a group of listeners who are asked to rate the quality of speech along a predetermined scale. Objective evaluation involves a mathematical comparison of the original and processed speech signals. Objective measures quantify quality by measuring the numerical "distance" between the original and processed signals. Clearly, for the objective measure to be valid, it needs to correlate well with subjective listening tests, and for that reason much research has been focused on developing objective measures that model various aspects of the auditory system.

All the traditional objective measures tend to give one score for one resulting speech to show how similar it is to the original speech. They may be highly correlated with certain aspects of speech enhancement such as speech distortion and noise reduction [75, 96], but they cannot present a clear and specific judgement on either aspect. On the other hand, most speech enhancement algorithms focus on obtaining a good trade-off between speech distortion and noise reduction. If an objective measure is able to predict the performance on speech distortion or noise reduction, then the speech enhancement algorithm can be selected for suitable applications. Moreover, speech enhancement algorithms can be improved correspondingly based on their scores of speech distortion and noise reduction. In this chapter, two novel objective measures are proposed to separately measure signal distortion and noise distortion. Actually this idea is an inspiration from Equation (5.4) where the residual signal $\varepsilon(\omega)$ consisted of the signal distortion $\varepsilon_S(\omega)$ and the noise distortion $\varepsilon_N(\omega)$. To facilitate understanding, Equation (5.4) is presented here again with time frame index m and frequency index ω :

$$\varepsilon(m, \omega) = \hat{S}(m, \omega) - S(m, \omega) = \varepsilon_S(m, \omega) + \varepsilon_N(m, \omega) \quad (6.1)$$

where $\varepsilon_S(m, \omega)$ and $\varepsilon_N(m, \omega)$ are the speech distortion and the noise distortion

respectively, and be defined as follows:

$$\varepsilon_S(m, \omega) = [G(m, \omega) - 1]S(m, \omega) \quad (6.2)$$

$$\varepsilon_N(m, \omega) = G(m, \omega)N(m, \omega) \quad (6.3)$$

6.2 Analysis for Speech Distortion

For speech distortion, a good speech enhancement filter should not filter off any speech components. The difference between the clean speech and the enhanced speech, $\varepsilon(\omega)$ defined in Equation (6.1), therefore needs to be analyzed. The ideal case is when the difference equals zero which is unattainable in practice. A low $\varepsilon_S(\omega)$ defined in Equation (6.2) implies the residual noise contains less speech components, which means the speech enhancement filter does not reduce the speech signal severely.

In order to test the performance of each speech enhancement filter on the speech distortion aspect, a scheme is proposed to calculate the cross-correlation between clean speech and the residual noise. Then the amount of clean speech which is contained in the residual noise can be evaluated. The cross-correlation can be calculated either in time domain or frequency domain. In this chapter, the frequency domain approach is preferred as the human hearing is more sensitive to the frequency magnitudes but is much less sensitive to the phase information. Moreover, the accuracy of the spectral envelop estimation incorporating the excitation signal can be further evaluated by this step. The normalized spectral cross-correlation $R_{S\varepsilon}$ can be expressed as a convolution of two spectra, clean speech, $S(m, \omega)$, and the residual signal, $\varepsilon(m, \omega)$, in the following equation for a certain speech frame m :

$$R_{S\varepsilon}(m, \omega) = \frac{S(m, \omega) * \varepsilon^*(m, -\omega)}{\sqrt{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |S(m, \omega)|^2 d\omega \times \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |\varepsilon(m, \omega)|^2 d\omega}} \quad (6.4)$$

where $*$ is the convolutional operator, ω_s is the sampling frequency, and $\varepsilon^*(m, \omega)$ is the conjugate complex of speech signal $\varepsilon(m, \omega)$. The cross-correlation is calculated between the spectral vectors on a frame by frame rule. If $S(m, \omega)$ and $\varepsilon(m, \omega)$ are of the same size M by N where M is the number of frames and N is the number of points in the frequency domain, then $R_{S\varepsilon}$ is a M by $2N - 1$ matrix in the Matlab function "xcorr". The degree of speech distortion in frame m , λ_m , is defined as

$$\lambda_m = \max_{\omega} [|R_{S\varepsilon}(m, \omega)|] \quad (6.5)$$

where function $\max_{\omega}(\cdot)$ is to find the maximum value along the frequency direction. Although the maximum is supposed to be in the center point, there is still a chance that it appears in other positions. The average value of λ_m computed over the entire speech utterance can be obtained by

$$\lambda = \frac{1}{M} \sum_{m=1}^M \lambda_m \quad (6.6)$$

A lower value for λ indicates a better performance of the speech enhancement filter in terms of reducing speech distortion.

6.3 Analysis for Noise Distortion

The next interesting issue is the noise distortion aspect. After analyzing the speech enhancement filter from the speech distortion perspective, the percentage of clean speech contained in the residual noise can be further exploited in the noise distortion analysis. Before that, the parameter λ_m defined in Equation (6.5) should be analyzed first. Combining Equation (6.1) and Equation (6.4), Equation (6.5) can be written as

$$\lambda_m = \frac{\max_{\omega} [|S(m, \omega) * \varepsilon_S^*(m, -\omega)|]}{\sqrt{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |S(m, \omega)|^2 d\omega \times \int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |\varepsilon(m, \omega)|^2 d\omega}} \quad (6.7)$$

where $S(m, \omega) * \varepsilon_N^*(m, -\omega)$ is eliminated since the clean speech is assumed to be uncorrelated with noise signal. For a certain speech and speech enhancement filter, the denominator of Equation (6.7) is a constant in each frame. Then the amount of clean speech contained in the residual noise is

$$\begin{aligned}
 \lambda_m \sqrt{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |\varepsilon(m, \omega)|^2 d\omega} &= \max_{\omega} \left\{ \left| \frac{S(m, \omega)}{\sqrt{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |S(m, \omega)|^2 d\omega}} * \varepsilon_S^*(m, -\omega) \right| \right\} \\
 &= \int \left\{ \left| \frac{S(m, \tau + \tau_{max})}{\sqrt{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |S(m, \omega)|^2 d\omega}} \varepsilon_S^*(m, \tau) \right| \right\} d\tau \\
 &= \int \mu(m, \tau + \tau_{max}) \cdot |\varepsilon_S(m, \tau)| d\tau \tag{6.8}
 \end{aligned}$$

where

$$\mu(m, \tau) = \frac{|S(m, \tau)|}{\sqrt{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |S(m, \omega)|^2 d\omega}} \geq 0 \tag{6.9}$$

τ_{max} is the position to obtain the maximum of $R_{S\varepsilon}$ and normally it equals to zero, $\mu(m, \tau)$ can be considered as a weighting function and $\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} \mu(m, \omega)^2 d\omega = 1$. Equation (6.8) is thus a weighted sum of the speech components contained in the residual signal, $\varepsilon_S(m, \omega)$. For a certain speech frame, all the weighting coefficients are the same and positive. Therefore, the parameter λ_m can quantitatively reflect the percentage of speech components embedded in the residual signal $\varepsilon(m, \omega)$.

Based on the above analysis, the removed clean speech during the speech enhancement process can be extracted by the residual noise $\varepsilon(m, \omega)$ multiplied with λ_m . The product $\lambda\varepsilon$ does not exactly represent the clean speech components contained in the residual noise. However, it is still able to provide a good insight into the statistical relationship between the removed clean speech and the residual noise. After removing the clean speech components from the residual noise, the remaining components are the so-called pure noise. The pure noise can be evaluated by its

spectral energy in the following equation

$$E_{PureNoise} = \frac{1}{M} \sum_{m=1}^M (1 - \lambda_m) \sqrt{\int_{-\frac{\omega_s}{2}}^{\frac{\omega_s}{2}} |\varepsilon(m, \omega)|^2 d\omega} \quad (6.10)$$

where M is the number of speech frames, λ_m is the degree of speech distortion in each speech frame defined in Equation (6.5), ω_s is the sampling frequency, and $\varepsilon(m, \omega)$ is residual noise signal defined in Equation (6.1). A lower value for $E_{PureNoise}$ indicates that a larger amount of noise has been removed, which means a better performance by the speech enhancement filter on noise reduction. It should be noted that the $E_{PureNoise}$ here is a square root of the energy. The reason for doing so is to be consistent with Equation (6.8).

6.4 Results and Discussion

Ten different segments of speeches, five females and five males are randomly selected from the TIMIT database. They are resampled at 8 kHz and quantized at 16 bits, and are corrupted by four types of noises, namely, white noise, fan noise, car noise and F16 aircraft noise. All these speech segments have silence periods at the beginning and at the end. All the proposed speech enhancement systems proposed in previous chapters can be evaluated by the proposed two novel objective measures as well. The results are discussed in the following sections.

6.4.1 DCT based Speech Enhancement Filtering

The proposed DCT based speech enhancement filtering described in Chapter 3 is an adaptive time-shift analysis system (ATSA). Four speech enhancement schemes are compared. The first one is the standard Wiener filter with fixed overlapping and time-shift analysis named as WF. The second one is pitch synchronized Wiener

Table 6.1: Normalized cross-correlation λ between clean speech and residual noise

Noise type	SNR (dB)	Normalized Cross-correlation			
		WF	PSWF	PSMax	ATSA
White	0	0.692	0.702	0.714	0.674
	5	0.667	0.674	0.687	0.638
	10	0.618	0.623	0.637	0.580
	15	0.563	0.565	0.577	0.520
	20	0.504	0.507	0.516	0.465
Fan	0	0.563	0.565	0.577	0.535
	5	0.525	0.528	0.540	0.494
	10	0.487	0.489	0.502	0.457
	15	0.448	0.450	0.460	0.420
	20	0.413	0.414	0.424	0.390
Car	0	0.459	0.465	0.472	0.442
	5	0.428	0.430	0.438	0.411
	10	0.402	0.403	0.409	0.388
	15	0.382	0.383	0.386	0.370
	20	0.369	0.368	0.369	0.358
F16 aircraft	0	0.739	0.749	0.759	0.722
	5	0.706	0.716	0.728	0.676
	10	0.657	0.663	0.676	0.616
	15	0.597	0.599	0.612	0.549
	20	0.528	0.532	0.544	0.480

filtering described in Section 3.5 which is named as PSWF. The third uses maximum alignment to improve pitch synchronization as described in Section 3.6 where the noise reduction filter is still the standard Wiener filter. This scheme is named as PSMax. The last one is the ATSA which has a Wiener filter controlled by an adaptive parameter α described in Section 3.7. The overlapping and time-shift is adaptive and can be improved by the maximum alignment.

To test the amount of speech distortion and noise reduction, residual noise signal defined in Equation (6.1) is obtained first from each type of speech enhancement

Figure 6.1: Spectral energies of pure noise from the different speech enhancement algorithms: WF, PSWF, PSMAX and ATSA

filters. The normalized cross-correlation factor λ is calculated by Equation (6.6) and the comparison results are shown in Table 6.1. From this table, ATSA performs the best since it causes the lowest speech distortion which is mainly contributed by the adaptive parameter α . For the noise distortion aspect, pure noise energy obtained from the different algorithms under different noise types and conditions are calculated by Equation (6.10). The results are shown in Figure 6.1. Relatively, the ATSA algorithm performs the worst, although the performances of the four speech enhancement schemes are actually very close. This result notwithstanding, the overall objective measures presented in Section 3.8 show that the ATSA algorithm

Table 6.2: Normalized cross-correlation λ between clean speech and residual noise

Noise type	SNR (dB)	Normalized Cross-correlation		
		2DFT	GSSB	HWSF
White	0	0.757	0.793	0.849
	2	0.737	0.779	0.823
	4	0.714	0.762	0.806
	6	0.687	0.742	0.776
	8	0.661	0.724	0.757
Fan	0	0.589	0.655	0.648
	2	0.570	0.640	0.632
	4	0.554	0.624	0.615
	6	0.537	0.610	0.596
	8	0.519	0.593	0.578
Car	0	0.507	0.553	0.554
	2	0.495	0.539	0.537
	4	0.487	0.525	0.523
	6	0.482	0.514	0.510
	8	0.479	0.502	0.496
F16 aircraft	0	0.788	0.812	0.839
	2	0.762	0.796	0.819
	4	0.735	0.778	0.799
	6	0.703	0.759	0.776
	8	0.669	0.742	0.754

is superior to the others.

6.4.2 Hybrid Wiener Spectrogram Filtering

The proposed hybrid Wiener spectrogram filtering (HWSF) method is described in Chapter 4. This algorithm places more emphasis on the noise reduction aspect. Three methods are compared, namely: 2DFT proposed by Soon in [78], General Spectral Subtraction with Blade post-processor (GSSB) explained in [82] and the

Figure 6.2: Spectral energies of pure noise from the different speech enhancement algorithms: 2DFT, GSSB and HWSF

proposed system, HWSF, which has been evaluated by other objective measures in Chapter 4.

The normalized cross-correlation factor λ is calculated by Equation (6.6). The values of λ under different noise types and input SNR conditions are shown in Table 6.2. This table shows that 2DFT performs the best in terms of speech distortion. The HWSF produces less speech distortion in narrow-band noise type and GSSB produces less distortion in wide-band noise type when compared to each other. These results mean that the blade post-processor incorporated in both HWSF and GSSB produces higher speech distortion which is an inevitable cost of its capability

Table 6.3: Normalized cross-correlation λ between clean speech and residual noise

Noise type	SNR (dB)	Normalized Cross-correlation			
		WF	PHFR	HRNR	OSCR
White	-10	0.817	0.826	0.934	0.676
	-5	0.785	0.789	0.880	0.672
	0	0.737	0.741	0.812	0.653
	5	0.680	0.698	0.739	0.627
	10	0.625	0.669	0.666	0.597
Fan	-10	0.680	0.693	0.736	0.649
	-5	0.633	0.653	0.676	0.612
	0	0.585	0.619	0.616	0.578
	5	0.538	0.592	0.561	0.552
	10	0.491	0.574	0.508	0.540
Car	-10	0.591	0.605	0.635	0.586
	-5	0.540	0.575	0.574	0.553
	0	0.496	0.555	0.520	0.533
	5	0.458	0.547	0.475	0.524
	10	0.427	0.545	0.438	0.525
F16 aircraft	-10	0.825	0.845	0.903	0.737
	-5	0.803	0.819	0.867	0.729
	0	0.751	0.764	0.803	0.703
	5	0.686	0.700	0.729	0.652
	10	0.623	0.648	0.657	0.603

in noise reduction. On the other hand, the comparison results of pure noise energy defined in Equation (6.10) are shown in Figure 6.2. The HWSF performs the best followed by the 2DFT and GSSB. This indicates that the hybrid Wiener filter is superior to the spectral subtraction algorithm in the noise reduction aspect.

6.4.3 Spectral Post-processing

The post-processing technique described in Chapter 5, named over-attenuated speech components regeneration (OSCR), aims to regenerate the over-attenuated speech components during the noise reduction process. This is one of the most suitable cases for utilizing the evaluation of speech distortion and noise distortion to illustrate the performance of the proposed technique. The speech distortion results will objectively determine if the proposed scheme meets its objective.

Four methods which have also been adopted in Section 5.7 are compared in this section: Wiener filter (WF) using Equation (4.2) as the representative of the TNR filter, PHFR [88] which is a post-processing technique for high-frequency regeneration, HRNR [83] which is a method for harmonic regeneration noise reduction and the proposed OSCR algorithm. To test the amount of speech distortion, residual noise from each type of speech enhancement filters is first obtained. Then the normalized cross-correlation factor λ is calculated by Equation (6.6). The values of λ under different noise types and input SNR conditions are shown in Table 6.3.

From the speech distortion aspect, among the four techniques, OSCR performs the best under wide-band noise types (white noise and F16 aircraft noise) followed by WF, PHFR and HRNR. For narrow-band noise types (car noise and fan noise), WF performs the best followed by OSCR, HRNR and PHFR. WF outperforms the rest in the narrow-band (i.e. low frequency band) noise scenarios in terms of speech distortion because it does not over-suppress the high frequency parts. Usually WF and some other TNR filters exhibit severe distortion in high frequency bands. However in the case of the narrow-band noise corruption, the high frequency portions are actually rather low in noise energy and are hence not severely suppressed thus resulting in the best performance for WF. The same explanation also holds for

Figure 6.3: Spectral energies of pure noise from the different speech enhancement algorithms: WF, PHFR, HRNR and OSCR

HRNR since HRNR still performs mainly as a noise reduction filter, although it introduces some compensatory harmonics at the same time.

To evaluate the performance of the four algorithms in terms of noise distortion which is introduced in Section 6.3, pure noises obtained from the different algorithms under different noise types and conditions are calculated by Equation (6.10). The comparison results are showed in Figure 6.3, where the lower the pure noise energy, the better the algorithm performs in noise reduction. It can be seen that, the proposed OSCR consistently gives the best results on the speeches with narrow-band noise corruptions (fan noise and car noise), demonstrating its excel-

lent capability in narrow-band noise reduction. On the other hand, for the noisy speeches contaminated by wide-band noises (white noise and F16 noise), the proposed algorithm performs well under low level noise circumstance (high SNR) but not so good under severe noise, although OSCR can better maintain the speech component in this situation as demonstrated in Table 6.3. This clearly demonstrates that the noise reduction capabilities of the OSCR is comparable with other techniques and the reduction in speech distortion does not come at the expense of noise reduction. It is also worth noting that HRNR does not perform as well as expected due to the more challenging cases and fewer voiced components appearing in the tested speech fragments (around 48.2% which is much lower than 80% used in [83]). The informal listening tests further validate the previous evaluation analysis on the four algorithms, where OSCR results for wide-band noise corrupted speeches sound more natural and richer in harmonics with little residual noise while the results for narrow-band noise corrupted speeches sound clear of noise.

Another interesting observation in Figure 6.3 is that pure noise energy in narrow-band noise scenarios is almost monotonically decreasing, while wide-band noise has a bell-shape like performance. The reason is that, compared to the wide-band noise, it is easier to separate the speech from the narrow-band noise, even under severe conditions. Besides, the pure noise energy defined in Equation (6.10) is a relative value and not an absolute value. A lower pure noise energy hence may also indicate that a higher amount of speech has been removed from the noisy speech, resulting in higher speech distortion. With reference to the wide-band noise scenarios in Table 6.3, there is a high correlation between clean speech and residual noise signal under low SNR conditions, which results in abnormally low pure noise energy in Figure 6.3.

6.5 Conclusion

Presently, there is no single objective measure that predicts the speech distortion directly. Most measures have been validated for evaluating a specific type of distortion, e.g. the Log Likelihood Ratio (LLR) is good for evaluating speech distortion while the segmental SNR is good for evaluating noise reduction [75,96]. They are all trying their best to predict the overall quality of speech through the use of a single score. A novel evaluation tool for filter performance is introduced in this chapter to test the impact of the filter on speech distortion and noise distortion. The experimental results have shown the performance of different speech enhancement systems in these two aspects. Among them, the results of over-attenuated speech components regeneration (OSCR) have been more thoroughly discussed, since the objective of this algorithm is to lessen the speech distortion by regenerating the over-attenuated speech components. The two proposed objective measures especially the measure for speech distortion are well placed to verify if OSCR meets its objective. The experimental results confirmed that OSCR works well and is superior to other related works. This conclusion is the same as the one obtained by other traditional objective measures discussed in Section 5.7.

Chapter 7

Evaluation of Speech Enhancement Algorithms

7.1 Introduction

This chapter compares the performances of the various speech enhancement algorithms proposed in the previous chapters. Three speech enhancement systems, namely adaptive time-shift analysis (ATSA), hybrid Wiener spectrogram filtering (HWSF) and over-attenuated speech components regeneration (OSCR) have been introduced in Chapters 3, 4 and 5, respectively. In this chapter, a detailed comparison among the three methods is presented in order to highlight their relative strengths and weaknesses, so that suitable methods for particular applications can be chosen depending on the specific characteristics.

The scope of these evaluations is limited to speech quality while intelligibility is not covered as it is more complicated to test especially by objective measures. Speech quality can be evaluated both by objective measures and subjective mea-

asures. Subjective measure is not favorable since it is very time consuming and may be less reliable when large amount of speech utterances have to be tested which will result in listening fatigue. On the other hand, objective measure which aims to predict the subjective quality of speech, Mean Opinion Score (MOS), has been proven to be reliable and is widely utilized by many researchers. A large number of test speeches can be evaluated by the computer in a consistent way.

The characteristics of different proposed speech enhancement algorithms will be reviewed. Several factors are important in the choice of a speech enhancement algorithm. The first will be the computational complexity of the enhancement algorithm. It will assess the cost for achieving a good quality speech. It also affects the feasibility of real time processing. The next one is the inherent time delay due to the algorithm. It determines if a speech enhancement algorithm is suitable for realtime implementation. The last but most important factor is the output quality of processed speech which determines whether the effort is worthwhile.

7.2 Computational Complexity

For comparison, the computational complexity will be discussed by presenting the amount of time it takes to complete all the Matlab operations performed by the speech enhancement scheme. Ten speech utterances under seven input SNR conditions ranging from -10 dB to 50 dB with a step size of 5 dB are tested. The running time is recorded by a computer with Intel(R) Core(TM)2 Duo CPU E8400 @ 3.0 GHz, 3.0 GB of RAM. The average computational time of the three proposed methods, ATSA, HWSF and OSCAR under different noise corruptions are shown in Table 7.1. From this table, it can be seen that OSCAR has the least computational complexity. Neither the noise type nor the input SNR affects the computational

Table 7.1: Comparison Results of Computational Time

Noise type	SNR (dB)	Computational Time (s)		
		ATSA	HWSF	OSCR
White	-10	5.03	10.14	0.23
	-5	5.85	10.08	0.22
	0	6.02	10.08	0.23
	5	5.52	10.04	0.23
	10	5.02	10.03	0.22
	15	4.55	9.97	0.22
	20	4.26	9.91	0.23
	Average	5.18	10.04	0.23
Fan	-10	10.18	10.06	0.23
	-5	9.17	10.01	0.22
	0	7.66	9.96	0.22
	5	6.40	9.87	0.22
	10	5.57	9.78	0.22
	15	5.13	9.66	0.23
	20	4.93	9.53	0.23
	Average	7.01	9.83	0.23
Car	-10	9.51	9.93	0.22
	-5	8.13	9.82	0.22
	0	7.02	9.70	0.23
	5	6.53	9.57	0.22
	10	6.0	9.48	0.23
	15	5.51	9.39	0.23
	20	5.27	9.30	0.23
	Average	6.85	9.60	0.23
F16 aircraft	-10	4.89	10.10	0.22
	-5	5.64	10.09	0.22
	0	5.82	10.06	0.22
	5	5.38	10.05	0.22
	10	4.94	9.98	0.22
	15	4.51	9.94	0.23
	20	4.33	9.84	0.23
	Average	5.07	10.01	0.22

complexity of OSCR in contrast to the other two methods. The reason is that the pitch synchronization in ATSA and multi-blade processing in HWSF are affected by the noise. Between these two methods, HWSF is less sensitive to the noise type and input SNR than ATSA, since multi-blade processing is a post-processor which is applied after the noise reduction process.

7.3 Time Delay

Table 7.2: Delay time for different frame size at sampling frequency of 8 kHz and 16 kHz

Frame Size	8 kHz	16 kHz
128	16 ms	8 ms
256	32 ms	16 ms
512	64 ms	32 ms
1024	128 ms	64 ms

The algorithmic time delay of an enhancement system is also an important factor when a system is deployed for real-time applications. If speech enhancement is to be incorporated into a speech coder or other realtime communication systems, the additional delay may introduce distortions to normal conversations. If the total delay is more than 100 ms, there is a high chance of double talk.

Normally the total delay time consists of the time taken for the framing process as well as the processing time. The processing time has been discussed in the previous section and it is becoming less important since it can be shorten drastically with the advent of high speech digital signal processors. Moreover, the computational complexity can often be reduced by using table lookup as the cost of large memory is low. Therefore the total delay time mainly depends on the framing process which is decided by the sampling frequency and the number of samples in a frame. The

delay time for the different frame size and sampling frequency is shown in Table 7.2.

Table 7.3: Delay time for different number of frames with different amount of overlap

No. of Frames	50 %	75 %
3	48 ms	40 ms
5	64 ms	48 ms
7	80 ms	56 ms
9	96 ms	64 ms
11	112 ms	72 ms
13	128 ms	80 ms
15	144 ms	88 ms

In this thesis, the frame size is 256 and the sampling frequency is 8 kHz. The delay time given in Table 7.2 is applicable to the proposed algorithms, namely, adaptive time-shift analysis (ATSA) in Chapter 3, and over-attenuated speech components regeneration (OSCR) in Chapter 5. The algorithm, hybrid Wiener spectrogram filtering (HWSF), proposed in Chapter 4 may not be suitable for real time processing since multiple frames are involved. It needs the "future" information of the signal since the hybrid Wiener filter needs 3 frames to calculate the local characteristics and the multi-blade post-processor needs a maximum of 13 frames to calculate the local variance. Thus the HWSF algorithm is more suitable to be applied on recorded speeches. However, the delay time can still be discussed. It is a function of the number of frames and the amount of overlap (50% and 75%) for a certain sampling frequency and frame size. In the case of 8 kHz sampling frequency and a frame size of 256, the delay time can be found in Table 7.3.

7.4 Clean Speech and Noise Type

The clean speech data are obtained from the TIMIT acoustic-phonetic speech database which contains over 6000 read sentences by different speakers. The data is originally sampled at 16 kHz and quantized to 16 bits. In this thesis, ten sentences consisting of five female and five male speakers were randomly selected from the database. They were appropriately filtered and down-sampled to 8 kHz to obtain the narrow-band speech which are used in this investigation.

Various noise types are used for the evaluation of the proposed speech enhancement algorithms. The white noise is the most commonly used and can be easily generated by Gaussian distributed random numbers. This type of noise has a flat frequency spectrum which means that it will overlap with all the speech spectral components. Fan noise and car noise have their energy concentrated in the lower frequency spectrum and can be considered to be narrow-band noise types. They are relatively easy to be removed as there is less overlap with the speech spectrum. In contrast, F16 aircraft noise is a kind of wide-band noise which is similar to white noise. The level of difficulty in reducing such kind of noise is moderate as compared to other noise types.

The three proposed algorithms can deal with both white noise and colored noise. They work reasonably well with stationary or slow varying additive noise such as the examples given above. However, in the case of highly non-stationary noise, such as interfering voice or music, as well as echo, these enhancement algorithms are unlikely to be effective in reducing the noise and improving the speech quality.

7.5 Performance Comparison

In computer simulations of algorithms, the original clean speech is available and the quality of the processed speech can be judged by its closeness to the original clean speech. Towards this end, mathematical comparisons can be implemented by computers objectively. In Chapters 3, 4 and 5, three different speech enhancement algorithms, namely, ATSA, HWSF and OSCR, are proposed to estimate the embedded speech in the different domains. In order to compare these algorithms, several objective measures will be employed in the following comparison.

7.5.1 Objective Measures

In [75, 96], several widely used objective speech quality measures were evaluated: the segmental SNR (SegSNR) measure, the weighted-slope spectral (WSS) distance, the perceptual evaluation of speech quality (PESQ) measure, the log-likelihood ratio (LLR) measure, the Itakura-Saito (IS) distance and etc. Composite measures combining a subset of the aforementioned measures are proposed and evaluated as well. PESQ measure has been reported to perform the best in terms of overall speech quality among all these compared common objective measures. Three composite measures, SIG, BAK and OVL proposed in [75] have been reported to be more correlated with mean opinion score (MOS) in terms of speech distortion, noise distortion and overall quality, respectively. Therefore, in this chapter, these four advanced objective measures will be used to compare the different speech enhancement algorithms. The SegSNR measure is also adopted since it is one of the simplest objective measures which is easy to be implemented and has been widely used for evaluating speech enhancement algorithms.

Segmental SNR Measure

Signal to noise ratio (SNR) is a common indicator of signal statistics which is defined as signal energy divided by noise energy and is normally expressed in decibels (dB). It has been reported to be a poor estimator of speech quality for a broad range of speech distortions [98,99]. The segmental SNR measure is thus proposed to improve SNR measure by taking the average of SNR values calculated for segments of speech data. It is defined as:

$$SegSNR = \frac{1}{M} \sum_{j=0}^{M-1} 10 \log_{10} \left[\frac{\sum_{m_j}^{m_j+N-1} s(n)^2}{\sum_{m_j}^{m_j+N-1} (\hat{s}(n) - s(n))^2} \right], \quad (7.1)$$

where $m_0, m_1, \dots, m_{M-1}$ are the end-times for the M frames, each of which is of length N . SegSNR is better correlated with Mean Opinion Score (MOS) than SNR as indicated by [74], however, the signal energy during intervals of silence in the speech signal will be very small, resulting in large negative values, which will bias the overall measure. This can be remedied by comparing short-time energy measurements against a threshold or by flooring the SegSNR values to a small value. In this thesis, the SegSNR is limited in the range from -10 dB to 35 dB. The segmental SNR can also be evaluated in the frequency domain.

PESQ Measure

Most of the preceding objective measures have been found unsuitable for assessing distortions resulting from a speech signal which has traversed a telecommunication network. Signal delays and codec distortions may cause inaccurate predictions of speech quality. A number of objective measures were proposed in the 1990s focusing on this type of distortion as well as filtering effects and variable signal delays. A competition was held in 2000 by the ITU-T study group 12 and the perceptual evaluation of speech quality (PESQ) measure described in [100], which is capable of

Figure 7.1: Block diagram of the PESQ measure

performing reliably across a wide range of codec and network conditions, was selected as the ITU-T recommendation P.862 which to replace the old P.861 recommendation.

The structure of the PESQ measure is shown in Figure 7.1. The original and degraded signals are first level-equalized to a standard listening level and then aligned in time to correct for time delays. After that, the signals will be processed through an auditory transform to obtain the loudness spectra. The difference, termed the disturbance, between the loudness spectra is computed and averaged over time and frequency to produce the prediction of subjective MOS score. High correlations (above 0.92 for codec and network conditions) with subjective listening tests were reported in [100] using the preceding PESQ measure for a large number of testing conditions. PESQ score only reflects the effects of one-way speech or noise distortion perceived by the end user while effects such as loudness loss and talker echo are not reflected.

Composite Measures

In addition to the existing measures, one can form the so-called composite measures by combining multiple objective measures. Different objective measures capture different characteristics of the distorted signal, and therefore combining them in a linear or nonlinear fashion can potentially yield significant gains in correlations. In [75], multiple linear regression analysis is used to form the following composite measures:

- A measure, SIG, for signal distortion formed by linear combination of the LLR, PESQ and WSS measures.
- A measure, BAK, for noise distortion formed by linear combination of the SegSNR, PESQ and WSS measures.
- A measure, OVL, for overall quality formed by linear combination of the PESQ, LLR and WSS measures.

The three new composite measures can be obtained by:

$$SIG = 3.093 - 1.029 \cdot LLR + 0.603 \cdot PESQ - 0.009 \cdot WSS \quad (7.2)$$

$$BAK = 1.634 + 0.478 \cdot PESQ - 0.007 \cdot WSS + 0.063 \cdot SegSNR \quad (7.3)$$

$$OVL = 1.594 + 0.805 \cdot PESQ - 0.512 \cdot LLR - 0.007 \cdot WSS \quad (7.4)$$

For speech enhancement algorithms, correlations of 0.7, 0.58 and 0.73 with subjective listening test are obtained by SIG, BAK and OVL, respectively [75].

7.5.2 Results Discussion

In this comparison study, 10 speech utterances from the TIMIT speech database with half from male speakers and another half from female speakers are used. Four

Figure 7.2: Comparison results of the SegSNR measure

types of noise from NOISEX database are used for the performance evaluation. They include white noise, fan noise, car noise and F16 aircraft noise (Section 7.4). The SNRs used are from -10 dB and 10 dB in steps of five. Therefore a total of 200 sentences are tested. Five objective measures are chosen, namely, SegSNR, PESQ, SIG, BAK and OVL which have been introduced in Section 7.5.1. Three proposed speech enhancement system, namely, ATSA, HWSF and OSCR, will be evaluated by these objective measures. The results for the four noise types and five objective measures are presented graphically in the following for easier interpretation.

Figure 7.2 shows the segmental SNR performance of ATSA, HWSF and OSCR

Figure 7.3: Comparison results of the PESQ measure

under different noise types, namely, white noise, fan noise, car noise and F16 aircraft noise. HWSF is the best in the case of wide-band noise except F16 aircraft noise at high SNR input, while OSCR is the best in the narrow-band noise with a little drop at 10 dB. ATSA has the lowest scores in most cases among these three algorithms in SegSNR results. According to [96], SegSNR is more correlated to the noise reduction aspect which means the HWSF and OSCR are better at noise reduction than ATSA.

Figure 7.3 shows the PESQ scores of three algorithms under different noise types. OSCR has the overall best PESQ performance in the cases of car noise and fan noise, while ATSA shows an excellent performance on speech enhancement under white

Figure 7.4: Comparison results of the SIG measure

noise and F16 aircraft noise. High energy compaction capability of DCT transform allows ATSA to suppress wide-band noise type effectively and efficiently. HWSF performs the worst due to the distortion of speech signal although it has a better performance for noise reduction.

In terms of composite objective measures, the comparison results of SIG measure is shown in Figure 7.4. ATSA performs the best among the three algorithms which has been presented in Section 6.4.1 showing good performance in speech distortion. In terms of speech distortion, the three speech enhancement algorithms perform even worse than the noisy speech at -10 dB under the wide-band noise types. This

Figure 7.5: Comparison results of the BAK measure

is because that the severe noise has been suppressed at the expense of some speech distortions. Noisy speech definitely contains more speech components than the processed speech which make it more correlated with the embedded speech despite the presence of more noise. HWSF has the lowest scores among these three speech enhancement algorithms.

In terms of noise distortion, OSCR performs the best with a little drop at 10 dB in narrow-band noise type. HWSF does not show as good a performance as in SegSNR comparison, and ATSA is not as superior as in SegSNR comparison for wide-band noise types either. The comparison results are shown in Figure 7.5 where

Figure 7.6: Comparison results of the OVL measure

OSCR performs the best, followed by ATSA and HWSF.

Figure 7.6 shows the overall score of the resulting speech quality. OSCR and ATSA have competitive performances over four noise types. It can be considered that ATSA is the best for the wide-band noise type while OSCR is the best for narrow-band noise types. This conclusion is the same as the one derived from PESQ results, and according to [96, 100], PESQ is highly correlated with subjective listening test and presents the best prediction on overall speech quality among several traditional objective speech enhancement measures.

7.6 Conclusion

The performances of three speech enhancement algorithms proposed in the previous chapters have been compared. They are ATSA, adaptive time-shift analysis algorithm, HWSF, hybrid Wiener spectrogram filtering, and OSCR, over-attenuated speech components regeneration. All the three algorithms are aimed at enhancing the noise corrupted speeches.

As presented, none of the three speech enhancement algorithms turns up consistently the best in all the performance measures. The different performance measures highlight the relative strengths and weaknesses of the various methods when dealing with the different types of noise. Basically, OSCR yields the best performance over the other two algorithms based on the overall performance and computational cost. It has the same time delay as ATSA. ATSA gives an excellent performance in suppressing wide-band noise types and exhibits good capability in lowering speech distortion. However, it may be sensitive to noise types and different noise corruptions can lead to different processing time. HWSF has the highest computational cost and time delay compared to the other two algorithms. Although HWSF is relatively not as good in comparison, it is still able to provide quite satisfactory performance and is highly effective in reducing background noise, as indicated by the SegSNR comparison results. With no time delay requirements and for a situation that demands more aggressive noise reduction, HWSF may be the most suitable choice.

Chapter 8

Recommendations and Conclusion

8.1 Conclusion

The main objective of speech enhancement is to suppress the noise components while keeping distortion of the speech signals to a minimum. All the algorithms do improve the perceptual quality of the speech after noise suppression. The noise problem can never be completely solved, but it can be reduced to a tolerable level. On the other hand, it is observed that the remaining artifacts can be disturbing despite the reduction in noise level. For high SNR situations, the output quality can be quite good. For very low SNR situations, the output quality improves but it still suffers from either residual noise or distortion to speech or both. Therefore, how to reduce the noise to the minimum and how to optimally achieve a trade-off between the amount of artifacts and the level of reduction of the noise have been the main focus of most speech enhancement algorithms.

Various techniques for single channel speech enhancement have been investigated in this thesis in the context of the scope mentioned above. Amongst that, Discrete

Cosine Transform (DCT) based speech enhancement algorithm is introduced first to show the feasibility of utilizing the transform. Other than the most commonly used Fourier transform, many researchers have also investigated other transforms such as the Karhunen-Loeve Transform (KLT) and Wavelet transform. DCT is known to be a good approximation to the KLT and has features similar to the Discrete Fourier Transform (DFT). Thus it is applicable to speech enhancement as well. Work in this area was initiated by Soon et al. [21]. The better energy compaction property of DCT has been shown to improve the performance of simple speech enhancement filters which are workable for both white noise and colored noise. However, due to the lack of phase even for a perfectly stationary signal, the same transform coefficient in different frames exhibits significant variation. This is a disadvantage when inter-frame information is used, i.e. the decision-directed approach for the a-priori SNR estimate. An adaptive time-shift analysis (ATSA) speech enhancement system is thus introduced in Chapter 3 to reduce this variation based on a periodic shifting scheme. The pitch period of voiced sound is used as the amount of shift for the analysis window while a fixed amount of shift is used for the unvoiced sound. A more consistent DCT spectrogram is thus generated for better noise reduction filtering. An adaptive parameter is also proposed for a better estimate of the a-priori SNR which affects the optimal DCT filter, called Wiener filter. By these proposed techniques, the DCT properties have been further exploited and the DCT based speech enhancement algorithm presents a good performance for pursuing good quality speech. In terms of several objective measures, the resultant enhanced speech has the lowest speech distortions amongst all the proposed algorithms.

In this thesis, a spectral filtering method, named hybrid Wiener spectrogram filtering (HWSF), is developed in Chapter 4, which involves the use of hybrid Wiener filter and multi-blade post-processing technique. Extending from those works us-

ing inter-frame information such as the a-priori SNR estimation using the previous frame, hybrid Wiener filter uses multiple frames for speech enhancement by considering a block area of spectrum. A two-step noise reduction (TSNR) technique is also incorporated to correct the one-frame-delay problem for the a-priori SNR estimate during the implementation of Wiener filter. It therefore results in a more accurate estimate of the embedded speech but still suffers from some residual noises. To further reduce this residual noise, an image processing technique, named multi-blade post-processing technique, is utilized to locate and suppress the isolate peaks after hybrid Wiener filtering. The probability of speech presence and absence, so-called hard decision classification, is used to determine which non-speech components should be filtered again. This classification prevents the speech signal from undergoing further suppression and distortion. Actually if the enhancement algorithm is near perfect, it will not make any sense to apply the classification as it will lead to further distortion in weak speech signals. In Chapter 4, the objective measures results have shown the necessity of utilizing this hard decision classification, as well as the TSNR technique. Through computer simulations, the proposed HWSF algorithm has been shown to outperform some other related speech enhancement algorithms and has the lowest residual noise amongst all proposed algorithms in terms of segmental SNR objective measure.

Most speech enhancement algorithms tend to over-suppress the background noise which in turn tends to disturb the underlying speech signal. An over-attenuated speech components regeneration (OSCR) technique is introduced in Chapter 5 to compensate for this problem. For the voiced sounds, a full-wave rectifier is applied to efficiently restore the degraded harmonics. For the unvoiced parts, it is mainly contributed by the envelop estimation which actually acts on all the frequency bins and it is not just confined to the unvoiced components. Therefore, either voiced

sound or unvoiced sound can be regenerated along the entire frequency domain by this proposed technique which is quite novel compared to some previous work. The proposed OSCR technique not only addresses the shortcoming of noise reduction filter, but also maintains the advantage and yields a better result. Through a number of objective comparison results, the OSCR speech enhancement system outperforms many existing methods. It has achieved a good trade-off between minimizing speech distortion and maximizing residual noise reduction. However, the OSCR technique still can be improved by more accurate envelop estimation since the upper bound is given by the clean speech envelop which shows there is room for improvement. This has been discussed in Section 5.7.1 and can be considered for the future work.

The author noticed that all currently utilized objective measures seek to evaluate the enhanced speech by representing the degree of similarity to the original clean speech. Although such a practice may give some preference to different kinds of measures, the specific circumstances, such as the degree of speech distortion and noise distortion, have not been clear. A good resultant speech is one who has minimum residual noise while maintaining the underlying speech signal as much as possible. Therefore, speech distortion and noise distortion are the two key metrics for evaluating a speech enhancement system. A novel evaluation tool for filter performance is introduced in Chapter 6 to test the impact of the filter on speech distortion and noise distortion. The experimental results have shown the performance of different speech enhancement systems based on these two aspects. The proposed ATSA and OSCR show a better performance for speech distortion and HWSF shows a better performance for noise distortion. This has been verified by other objective measures and the results have been discussed in Chapter 7. However, these two proposed objective measures still have shortcomings in that the predicted degrees of speech distortion and noise distortion are not very accurate. The cross-correlation factor

calculated for speech distortion is disturbed by the noise variance which can be seen from its expression in Equation (6.4). Besides, the pure noise energy calculated for noise distortion is a relative value and is not a perfectly accurate value as well due to the inaccurate computation of clean speech components contained in the residual noise. Although the two proposed objective measures are working well, they can be further improved in the future.

Finally, the various proposed algorithms are also compared with each other in terms of their computational cost, inherent delay times and their output speech qualities. The output speech quality is measured using several objective tests. Objective tests are necessary for automated evaluation but it may be of little use if the results do not correlate well with subjective listening tests. Six objective tests, segmental signal to noise ratio (SegSNR), perceptual evaluation of speech quality (PESQ) and composite measures, namely SIG for signal distortion, BAK for noise distortion and OVL for overall quality, have been used for evaluation of the speech enhancement algorithms. All the measures except SegSNR have a high correlation with mean opinion score (MOS) which has been reported in previous work [96]. Nevertheless, SegSNR is one of the easiest and most widely utilized objective measures which make it easier to compare with results reported by other researchers.

8.2 Recommendations for Further Research

The DCT based speech enhancement scheme proposed in Chapter 3 seems very promising and there remains much to be explored especially on the DCT properties. Owing to the difference between DCT and DFT, some advanced DFT based speech enhancement filter can be rederived for the DCT domain. For example, the 2D technique adopted in Chapter 4 can be applied to DCT domain. Since there is less

speech distortion due to the high energy compaction capability of DCT, such kind of filter which is good at noise reduction should also be favorable. Combining these two strong points, a good DCT based speech enhancement filter can be produced. The background noise can be suppressed as much as possible without overly affecting the speech signal.

Another good area for future research will be the development of improved OSCR speech enhancement system. OSCR is really a novel system which placed much emphasis on the speech distortion aspect. As mentioned previously, an accurate estimate of speech envelop will result in good performance for both voiced sound and unvoiced sound regeneration. One attempt is to add some pre-information such as clean speech code book to get a "cleaner" speech envelop. After improving the speech envelop, more speech components can be restored and the speech distortion caused by the noise reduction filter can be further reduced.

For a more accurate prediction of resultant speech, the two proposed novel objective measures should be further improved as mentioned previously. The cross-correlation factor for speech distortion prediction is obtained by normalization of speech variance and noise variance. This may be disturbed by the noise and results in an inaccurate prediction which cannot truly reflect the speech components contained in the residual noise. On the other hand, the computation of pure noise energy for noise distortion is based on the results of speech distortion, leading to a cumulative result. It is also a relative value but not an absolute value and it can only be used for comparison. In the future, the speech distortion and noise distortion may be derived anew to find a more accurate level of prediction results for better objective evaluations.

Author's Publications

Journal Papers

1. H. Ding and I. Y. Soon and S. N. Koh and C. K. Yeo, A spectral filtering method based on hybrid wiener filters for speech enhancement. *Speech Communication*, Vol. 51, pp. 259–267, March 2009.
2. H. Ding and I. Y. Soon and C. K. Yeo, Over-attenuated components regeneration for speech enhancement. *IEEE Trans. Audio, Speech and Language Processing*, Vol. 18, pp. 2004–2014, 2010.
3. H. Ding and I. Y. Soon, A DCT based speech enhancement system with an adaptive time-shift analysis. *IEEE Trans. Audio, Speech and Language Processing*, [submitted].

Conference Papers

1. H. Ding and I. Y. Soon and C. K. Yeo and S. N. Koh, 2D spectrogram filter for single channel speech enhancement. *The 7th WSEAS International Conference on Signal, Speech and Image Processing (SSIP 07)*, Vol. 51, pp. 259–267, September 2007.
2. H. Ding and I. Y. Soon and S. N. Koh and C. K. Yeo, A post-processing technique for regeneration of over-attenuated speech components. *Proc. Internal Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3889–3892, April 2009.
3. H. Ding and I. Y. Soon, An adaptive time-shift analysis for DCT based speech enhancement. *Proc. Internal Conference on Information, Communications and Signal Processing (ICICS)*, pp. 1–4, December 2009.

Bibliography

- [1] M. R. Sambur and N. S. Jayant. LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise. *IEEE Trans. Acoust., Speech, Signal Processing*, 24:488–494, 1976.
- [2] B. H. Juang. Recent developments in speech recognition under adverse conditions. In *Proc. Int. Conf. Spoken Language Process*, pages 1113–1116, Kobe, Japan, November 1990.
- [3] J. H. L. Hansen and M. A. Clements. Constrained iterative speech enhancement with application to speech recognition. *IEEE Trans. Signal Processing*, 39:795–805, 1991.
- [4] T. Ramabadran, J. Ashley, and M. McLaughlin. Background noise suppression for speech enhancement and coding. In *IEEE Workshop on Speech Coding For Telecommunications Proceeding*, pages 43–44, 1997.
- [5] B. Milner, J. Darch, and S. Vaseghi. Applying noise compensation methods to robustly predict acoustic speech features from MFCC vectors in noise. In *Proceedings ICASSP*, pages 3945–3948, 2008.
- [6] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero. A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition. In *Proceedings ICASSP*, pages 4041–4044, 2008.
- [7] P. C. Loizou. *Speech Enhancement Theory and Practice*. CRC Press, 2007.
- [8] T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd edition*. Wiley-Interscience, 2006.
- [9] L. B. Thomas and A. Ravindran. Intelligibility enhancement of already noisy speech signals. *J. Audio Eng. Soc.*, 22:234–236, 1974.
- [10] J. S. Lim and A. V. Oppenheim. All-pole modeling of degraded speech. *IEEE Trans. on Acoust., Speech, Signal Processing*, 26:197–210, 1978.

-
- [11] K. K. Paliwal and A. Basu. A speech enhancement method based on kalman filtering. In *Proceedings ICASSP*, pages 177–180, 1987.
 - [12] P. Sorqvist, P. Handel, and B. Ottersten. Kalman filtering for low distortion speech enhancement in mobile communication. In *ICASSP*, volume 2, pages 1219–1222, 1997.
 - [13] Z. Goh, K. C. Tan, and B. T. G. Tan. Kalman filtering speech enhancement method based on a voiced-unvoiced speech model. *IEEE Trans. Speech and Audio Processing*, 7:510–524, 1999.
 - [14] Y. Ephraim. Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, 80:1526–1555, 1992.
 - [15] Stefan Oberle and August Kaelin. HMM-based speech enhancement using pitch period information in voiced speech segments. In *Proceedings IEEE International Symposium on Circuits and Systems*, volume 4, pages 2645–2648, 1997.
 - [16] M. E. Deisher and A. S. Spanias. HMM-based speech enhancement using harmonic modeling. In *Proceedings ICASSP*, volume 2, pages 1175–1178, 1997.
 - [17] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, 27:113–120, 1979.
 - [18] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, 32:1109–1121, 1984.
 - [19] R. J. McAulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing*, 28:137–145, 1980.
 - [20] P. Scalart and J. Vieira Filho. Speech enhancement based on a a priori signal to noise estimation. In *Proceedings ICASSP*, volume 2, pages 629–632, 1996.
 - [21] I. Y. Soon, S. N. Koh, and C. K. Yeo. Noisy speech enhancement using discrete cosine transform. *Speech Communication*, 24:249–257, 1998.
 - [22] I. Y. Soon, S. N. Koh, and C. K. Yeo. Speech enhancement using discrete cosine transform. In *IASTED International Conference on Signal and Image Processing*, volume 1, pages 3–7, 1996.
 - [23] H. Ding and I. Y. Soon. An adaptive time-shift analysis for DCT based speech enhancement. In *Proceedings ICICS*, pages 1–4, 2009.

-
- [24] Jong Won Seok and Keun Sung Bae. Speech enhancement with reduction of noise components in the wavelet domain. In *Proceedings ICASSP*, volume 2, pages 1323–1326, 1997.
 - [25] I. Y. Soon, S. N. Koh, and C. K. Yeo. Wavelet for speech denoising. In *TENCON Proceedings*, volume 2, pages 479–482, 1997.
 - [26] I. Y. Soon, K. J. Tan, and C. K. Yeo. Speech enhancement using wavelet packet analysis. In *IEEE Asia Pacific Conference on Communications/ Singapore International Conference on Communication Systems*, volume 1, pages 117–121, 1998.
 - [27] M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise: A regenerative approach. *Speech Communication*, 10:45–57, 1991.
 - [28] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech and Audio Processing*, 3:251–266, 1995.
 - [29] A. Rezayee and S. Gazor. An adaptive KLT approach for speech enhancement. *IEEE Trans. Speech and Audio Processing*, 9(2):87–95, 2001.
 - [30] J. C. R. Licklider and I. Pollack. Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. *The Journal of the Acoustical Society of America*, 20:42, 1948.
 - [31] M. R. Schroeder. Apparatus for suppressing noise and distortion in communication signals, 1965. U.S. Patent No. 3,180,936.
 - [32] M. R. Weiss, E. Aschkenasy, T. W. Parsons, and NICOLET SCIENTIFIC CORP NORTHVALE NJ. Study and development of the INTEL technique for improving speech intelligibility. *Technical Report NSC-FR/4023, Nicolet Scientific Corporation*, 1975.
 - [33] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of IEEE*, 67:1586–1604, 1979.
 - [34] Y. Ephraim and I. Cohen. Recent advancements in speech enhancement. *The Electrical Engineering Handbook*, 3, 2004.
 - [35] P. T. Brady. A technique for investigation on-off patterns of speech. *The Bell System Technical Journal*, XLIV:1–22, 1965.
 - [36] J. C. Junqua, B. Reaves, and B. Mak. A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer.

-
- In *Second European Conference on Speech Communication and Technology*, pages 1371–1374, 1991.
- [37] L. Rabiner and M. Sambur. Voiced-unvoiced-silence detection using the Itakura LPC distance measure. In *Processings ICASSP*, volume 2, 1977.
 - [38] R. Tucker. Voice activity detection using a periodicity measure. *IEE Proceedings I Communications, Speech and Vision*, 139:377–380, 1992.
 - [39] J. A. Haigh and J. S. Mason. Robust voice activity detection using cepstral features. In *1993 IEEE Region 10 Conference on TENCN'93. Proceedings. Computer, Communication, Control and Power Engineering*, pages 321–324, 1994.
 - [40] A. M. Kondo. *Digital speech: coding for low bit rate communication systems, 2nd edition*. Wiley, 2004.
 - [41] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.
 - [42] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5):504–512, 2001.
 - [43] L. Lin, W. H. Holmes, and E. Ambikairajah. Subband noise estimation for speech enhancement using a perceptual Wiener filter. In *Proceedings ICASSP*, pages 80–83, 2003.
 - [44] I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11:466–475, 2003.
 - [45] H. G. Hirsch and C. Ehrlicher. Noise estimation techniques for robust speech recognition. In *Proceedings ICASSP*, volume 1, pages 153–156, 1995.
 - [46] R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim. Enhancement of speech by adaptive filtering. In *Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing*, pages 251–253, Kobe, Japan, April 1976.
 - [47] J. S. Lim, A. V. Oppenheim, and L. D. Braida. Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition. *IEEE Trans. on Acoust. , Speech, Signal Processing*, 26:354–358, 1978.

-
- [48] J. D. Gibson, B. Koo, and S. D. Gray. Filtering of colored noise for speech enhancement and coding. *IEEE Transactions on Signal Processing*, 39:1732–1742, 1991.
 - [49] D. Popescu and I. Zeljkovic. Kalman filtering of colored noise for speech enhancement. In *Proceedings ICASSP*, volume 2, pages 997–1000, 1998.
 - [50] Y. Wang, J. An, V. Sethu, and E. Ambikairajah. Perceptually motivated pre-filter for speech enhancement using Kalman filtering. In *Information, Communications & Signal Processing, 2007 6th International Conference on*, pages 1–4, 2007.
 - [51] W. R. Wu, P. C. Chen, H. T. Chang, and C. H. Kuo. Frame-based subband Kalman filtering for speech enhancement. In *Proceedings Fourth International Conference on Signal Processing*, volume 1, pages 682–685, 1998.
 - [52] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.
 - [53] D. B. Dean, P. J. Lucey, S. Sridharan, and T. J. Wark. Fused HMM-adaptation of multi-stream HMMs for audio-visual speech recognition. *Interspeech, Antwerp, August*, pages 666–669, 2007.
 - [54] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan. HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. *IEEE Trans. Speech and Audio Processing*, 6:445–455, 1998.
 - [55] T. T. Le and J. S. Mason. Artificial neural networks for nonlinear time-domain filtering of speech. *IEE Proceedings Vision, Image and Signal Processing*, 143:149–154, 1996.
 - [56] I. Potamitis, N. D. Fakotakis, and G. Kokkinakis. Impulsive noise suppression using neural networks. In *Proceedings ICASSP*, volume 3, pages 1871–1874, 2000.
 - [57] A. T. Nelson and E. A. Wan. Neural speech enhancement using dual extended Kalman filtering. In *Proceedings of IEEE International Conference on Neural Networks*, volume 4, pages 2171–2175, 1997.
 - [58] K. Y. Lee, S. McLaughlin, and K. Shirai. Speech enhancement based on neural predictive hidden Markov model. *Signal Processing*, 65:373–381, 1998.
 - [59] R. E. Crochiere. A weighted overlap-add method of short-time Fourier analysis /synthesis. *IEEE Trans. Acoust., Speech, Signal Processing*, 28:99–102, 1980.

-
- [60] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *IEEE Proc. ICASSP*, pages 208–211, April 1979.
 - [61] P. Lockwood and J. Boudy. Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. *Speech Communication*, 11(2-3):215–228, 1992.
 - [62] J. Lim. Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(5):471–472, 1978.
 - [63] Olivier Cappe. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Trans. Speech and Audio Processing*, 2:345–349, 1994.
 - [64] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Wiley, New York, 1980.
 - [65] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, 33:443–445, 1985.
 - [66] M.S.A. Zilany M.K. Hasan and M.R. Khan. DCT speech enhancement with hard and soft thresholding criteria. *Electronics Letters*, 38(13):669–670, 2002.
 - [67] U. Mittal and N. Phamdo. Signal/noise KLT based approach for enhancing speech degraded by colored noise. *IEEE Trans. Speech and Audio Processing*, 8:159–167, 2000.
 - [68] S. Gazor and A. Rezayee. An adaptive subspace approach for speech enhancement. In *Proceedings ICASSP*, volume 3, pages 1839–1842, 2000.
 - [69] Joon-Hyuk Chang. Warped discrete cosine transform-based noisy speech enhancement. *IEEE Trans. Circuits and Systems II: Express Briefs*, 52(9):535–539, 2005.
 - [70] S. Ou, X. Zhao, and J. Dong. Combining DCT and Adaptive KLT for Noisy Speech Enhancement. In *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, pages 2857–2860, 2007.
 - [71] N. L. Phu, E. Ambikairajah, and E. Choi. An improved soft threshold method for DCT speech enhancement. In *2nd International Conference on Communications and Electronics*, pages 268–271, 2008.

-
- [72] I. Y. Soon and S. N. Koh. Low distortion speech enhancement. *IEE Proceedings Vision, Image and Signal Processing*, 147:247–253, 2000.
- [73] M. K. Hasan, S. Salahuddin, and M. R. Khan. A modified a priori SNR for speech enhancement using spectral subtraction rules. *IEEE Signal Processing Letters*, 11(4):450–453, 2004.
- [74] Shihua Wang, Andrew Sekey, and Allen Gersho. An objective measure for predicting subjective quality of Speech Coders. *IEEE Journal on Selected Areas in Communications*, 10:819–829, 1992.
- [75] Y. Hu and P. C. Loizou. Evaluation of objective measures for speech enhancement. In *Proceedings of INTERSPEECH-2006, Philadelphia, PA*, 2006.
- [76] Yan Ming Cheng and Douglas O’Shaughnessy. Speech enhancement based conceptually on auditory evidence. *IEEE Trans. Signal Processing*, 39:1943–1954, 1991.
- [77] J. Jensen and H. L. Hansen. Speech enhancement using a constrained iterative sinusoidal model. *IEEE Trans. Speech and Audio Processing*, 9:731–740, 2001.
- [78] I. Y. Soon and S. N. Koh. Speech enhancement using 2-D Fourier transform. *IEEE Trans. Speech and Audio Processing*, 11:717–724, 2003.
- [79] N. W. D. Evans, J. S. Mason, and M. J. Roach. Noise compensation using spectrogram morphological filtering. In *ProcEEDINGS 4th IASTED International Con. on Signal and Image Processing*, pages 157–161, 2002.
- [80] Z. Lin and R. Goubran. Musical noise reduction in speech using two-dimensional spectrogram enhancement. In *Proceedings HAVE*, pages 61–64, 2003.
- [81] G. Whipple. Low residual noise speech enhancement utilizing time-frequency filtering. In *Proc. ICASSP*, volume 1, pages I/5–I/8, 1994.
- [82] Zenton Goh, Kah-Chye Tan, and B. T. G. Tan. Postprocessing method for suppressing musical noise generated by spectral subtraction. *IEEE Trans. Speech Audio Processing*, 6:287–292, 1998.
- [83] C. Plapous, C. Marro, and P. Scalart. Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Trans. Audio, Speech and Language Processing*, 14:2098–2108, 2006.

-
- [84] I. Y. Soon, S. N. Koh, and C. K. Yeo. Improved noise suppression filter using self adaptive estimator of probability of speech absence. *Signal Processing*, 75:151–159, 1999.
 - [85] C. H. You, S. N. Koh, and R. Susanto. β -order MMSE spectral amplitude estimation for speech enhancement. *IEEE Trans. Speech and Audio Processing*, 13:457–486, 2005.
 - [86] H. Ding, I. Y. Soon, S. N. Koh, and C. K. Yeo. A spectral filtering method based on hybrid wiener filters for speech enhancement. *Speech Communication*, 51:259–267, 2009.
 - [87] C. Plapous, C. Marro, and P. Scalart. Speech enhancement using harmonic regeneration. In *Proceedings ICASSP*, pages 157–160, 2005.
 - [88] X. Zhang, S.N. Koh, I.Y. Soon, and C. H. You. Post-processing in masking-based β -order MMSE speech enhancement. *Applied Acoustics*, 69:354–357, 2008.
 - [89] Y. Hu and PC Loizou. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Trans. Speech and Audio Processing*, 11(4):334–341, 2003.
 - [90] J. M. Valin and R. Lefebvre. Bandwidth extension of narrowband speech for low bit-rate wideband coding. In *Proceedings IEEE Workshop on Speech Coding*, pages 130–132, 2000.
 - [91] C.J. Weinstein. A linear prediction vocoder with voice excitation. In *Proceedings EASCON*, pages 30A–30G, 1975.
 - [92] J. Makhoul and M. Berouti. High-frequency regeneration in speech coding systems. In *Proceedings ICASSP*, volume 4, pages 428–431, 1979.
 - [93] C. Plapous, C. Marro, P. Scalart, and L. Mauuary. A two-step noise reduction technique. In *Proceedings ICASSP*, pages 289–292, 2004.
 - [94] I. Y. Soon, S. N. Koh, and W. H. Ngo. Transformation of narrowband speech into wideband speech with aid of zero crossings rate. *IEE Electronics Letters*, 38(24):1607–1608, 2002.
 - [95] H. Ding, I. Y. Soon, S. N. Koh, and C. K. Yeo. A post-processing technique for regeneration of over-attenuated speech components. In *Proceedings ICASSP*, pages 3889–3892, 2009.

- [96] Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio, Speech and Language Processing*, 16:229–238, 2008.
- [97] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements. *Objective Measures of Speech Quality*. Prentice Hall, 1988.
- [98] B. J. Mcdermott, C. Scagliola, and D. J. Goodman. Perceptual and objective evaluation of speech processed by adaptive differential PCM. In *Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing*, pages 581–585, April 1978.
- [99] J. M. Tribolet, P. Noll, and B. J. Mcdermott. A study of complexity and quality of speech waveform coders. In *Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing*, pages 586–590, April 1978.
- [100] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proceedings ICASSP*, volume 2, pages 749–752, 2001.