Wavelet Based Adaptive Speech Enhancement

By

Essa Jafer Essa

B.Eng, MSc. Eng

A thesis submitted for the degree of Master of Engineering

Department of Electronic and Computer Engineering
University of Limerick
Limerick
Ireland

Submitted to the University of Limerick
September 2003
DEDICATION

In memory of the soul of my late father Dr. Jafer Essa
ACKNOWLEDGMENTS

Firstly I would like to thank my supervisor Dr. A.E. Mahdi for his encouragement and guidance throughout the course of this work.

I wish also to thank University of Limerick/Electronic and Computer Engineering department who gave me the opportunity to complete my research, and special thanks to all my friends and colleagues in the department especially Mr. Dorel Picovici.

Finally and not least, I would like to express my gratitude to Prof K.Arshak and Dr A.Arshak for their moral support and encouragement.
Contents

Abstract.............................................................................................................................................1

List of Symbols and Abbreviations............................................................. 3

Chapter 1: Introduction

1.1 Motivation...............................................................................................................................9
1.2 Speech Characteristics and Modelling.................................................................10
1.3 Speech Enhancement Techniques: An Overview........................................11
   1.3.1 Frequency and time domain based approach...........................................11
   1.3.2 Wavelet based approaches.......................................................................13
1.4 Scope of The Thesis.................................................................................................14

Chapter 2: Speech Enhancement: Review of Existing Approaches

2.1 Introduction.........................................................................................................................15
2.2 Review of Existing Spectral Domain Speech Enhancement Approaches….15
   2.2.1 Spectral subtraction .........................................................................................15
   2.2.2 Short Time Spectral Amplitude (STSA).......................................................18
2.3 Wavelet Domain Speech Enhancement Approaches...................................20
   2.3.1 Wiener filter-based speech enhancement method.................................21
   2.3.2 Adaptive denoising method based on speech classification.................21
   2.3.3 Noise estimation based adaptive denoising method.........................22
2.4 Shortcomings of Conventional Wavelet-Based Speech Signals Denoising..23
2.5 Outline of the Proposed Speech Enhancement System .........................25
2.6 Speech Quality Measures.......................................................................................27
Chapter 3: Wavelet Transforms

3.1 Mathematical Description of The Wavelet Transform
3.2 Second Generation wavelet Transform (SGWT)
   3.2.1 Traditional advantages of lifting implementations
3.3 Wavelet Packet Decomposition (WPD)
3.4 Perceptual / Bark Scaled Wavelet Transform
3.5 Wavelet-Based Denoising Using Thresholding
   3.5.1 Thresholding types
2.8 Summary

Chapter 4: Adaptive Noise Estimation

4.1 Introduction
4.2 Overview of Noise Tracking Techniques For Speech Signals
4.3 Noise Estimation Techniques
   4.3.1 Adaptive smoothing-based noise estimation algorithm
   4.3.2 Quantile based noise spectrum estimation
   4.3.3 Minimum variance tracking-based noise estimation algorithm
   4.3.4 Proposed noise estimation algorithm
4.4 Wavelet Decompositions
   4.4.1 Perceptual wavelet packet decomposition PWPD
4.5 Performance Evaluation and Experimental Results
   4.5.1 Performance of the four noise estimation techniques
   4.5.2 Performance evaluation using speech enhancement
4.6 Discussion Of Results
4.7 Summary
Chapter 5: Voiced/Unvoiced classification algorithm using dyadic wavelet transform

5.1 Overview of Existing Speech Classification Algorithms…………………..76
5.2 V/UV Classification Algorithm……………………………………………77
   5.2.1 Wavelet - based frequency distribution of the average energy………77
   5.2.2 Zero –crossing rate……………………………………………………..80
   5.2.3 Two features V/UV classification ……………………………………81
5.3 Implementation And Performance Evaluation of The Proposed
   Algorithm …………………………………………………………………..85
5.4 Discussion of The Results………………………………………………….88
5.5 Summary…………………………………………………………………...89

Chapter 6: New Adaptive Speech Enhancement System

6.1 Introduction………………………………………………………………...90
6.2 First Implementation Approach Speech Enhancement System…………92
   6.2.1 Time-frequency dependent threshold estimation……………………93
   6.2.2 Soft-thresholding using modified Weiner filtering & magnitude……93
classification using decision directed approach…………………………93
   6.2.3 Implementation and performance evaluation of the first approach….94
   6.2.4 Discussion of the results……………………………………………...95
6.3 Second Implementation Approach Speech Enhancement System………..100
   6.3.1 Structure of the wavelet decomposition…………………………….100
   6.3.2 Voiced/Unvoiced speech classification……………………………..102
   6.3.3 Estimation of noise profile and level………………………………..106
   6.3.4 The proposed threshold setting……………………………………..106
      6.3.4.1 Mean square error of the enhanced signal…………………..107
      6.3.4.2 Setting the threshold gain $F_{SNR}$ .................................108
      6.3.4.3 Setting the threshold gain $F_{V/UV}$ .................................109
   6.3.5 Performance evaluation of the second approach…………………..112
   6.3.6 Discussion of the results…………………………………………….130
6.4 Matlab based Graphical user Interface for speech enhancement systems..132
6.5 Summary…………………………………………………………………..139
Chapter 7: Conclusions and Suggestions for Further Work

7.1 Conclusions........................................................................................................140
7.2 Suggestions For Further Work...........................................................................143

References .................................................................................................................145

Appendix I: Design of Second-generation wavelet filters.................................152

Author Publications.....................................................................................................154
List of Figures

2.1 Basic spectral denoising procedure ........................................................................16
2.2 Speech enhancement system based on Weiner filtering ........................................21
2.3 Speech enhancement system based on speech classification ..............................22
2.4 Speech enhancement system based on adaptive noise estimation-thresholding ………23
2.5 Block diagram of the proposed speech enhancement system ..............................26
2.6 PSQM Testing Process ..........................................................................................32
3.1 3-Stage octave-band wavelet decomposition .....................................................36
3.2 Representation of the SGWT: (a) Forward transform, and (b) Inverse transform ............................................................................................................................38
3.3 Two-channel-filter banks with analysis filters $g$, $h$ and synthesis filters $\tilde{g}$, $\tilde{h}$ .............................................................................................................................38
3.4 Polyphase representation of Wavelet transform ....................................................40
3.5 Left side is the forward wavelet transform using Lifting, right side is the inverse wavelet transform using Lifting .................................................................41
3.6 Discrete wavelet packet decomposition coefficients with 2 decomposition .........43
3.7 Discrete wavelet transform coefficients with 2-level decomposition .................43
3.8 Approximation of the Bark scale (solid line) by BS-WPD (*) ...............................46
3.9 BS-WPD expansion tree .......................................................................................46
4.1 The relation between the frames and the observation window in the quantile based noise estimation algorithm .................................................................57
4.2 PWPD expansion tree ..........................................................................................60
4.3 Exact and approximation of the critical bands using 18-subbands PWPD .........60
4.4 Real and estimated noise using SGWT-based noise estimation with White noise 0dB ...................................................................................................................62
4.5 Real and estimated noise power using SGWT-based noise estimation with Pink noise at 0dB

4.6 Real and estimated noise power using SGWT-based noise estimation with F16 cockpit noise at 0dB

4.7 Real and estimated noise power using PWPD-based noise estimation with White noise at 0dB

4.8 Real and estimated noise power using PWPD-based noise estimation with Pink noise at 0dB

4.9 Real and estimated noise power using PWPD-based noise estimation with F16 cockpit noise at 0dB

4.10 (a) A subband noisy signal at 10 dB with, (b) corresponding updating of values of $\alpha(p)$

4.11 Speech enhancement performance evaluation (PWPD) with AWGN and soft thresholding

4.12 Speech enhancement performance evaluation (PWPD) With AWGN and Weiner soft thresholding

4.13 Spectrograms of:(a) Noisy sentence (AWGN 5dB); (b) The denoised sentence by soft thresholding using adaptive algorithm and PWPD

4.14 Spectrograms of:(a) Noisy sentence (Pink 5dB); (b) The denoised sentence by soft thresholding using adaptive algorithm and PWPD

4.15 Spectrograms of:(a) Speech signal distorted by AWGN 5dB; (b) The denoised sentence by soft thresholding using the proposed algorithm and PWPD

5.1 Block diagram of the proposed V/UV speech classification algorithm

5.2 (a)The waveform of a clean unvoiced-voiced speech signal, and (b) the corresponding wavelet-based frequency distribution of the average energy in each segment

5.3 Histograms of zero-crossing rates for utterances taken from (a),(b) two different female speakers, and (c),(d) two different female speakers

5.4 The per band energy distribution and the zero-crossing rates feature for each segment of the speech signal used in Figure 4.2, when (a) the signal is clean, and (b) the signal is noise-degraded

5.5 The distributions of zero-crossing rates and per band average energy for a
speech signal contaminated with noise at (a) SNR =15dB, and (b) SNR = 5dB

5.6 Voiced-Unvoiced labelled segments of speech signal at three different conditions: (a) Clean signal, (b) Noisy at SNR=15dB and (c) Noisy at SNR=10 dB

6.1 Schematic representation of the proposed speech enhancement system using BS-WPD

6.2 Waveforms of the (a) noisy speech (AWGN) with SNR=5dB, and (b) enhanced speech

6.3 Waveforms of the (a) noisy speech (pink noise) with SNR=5dB, and (b) enhanced speech

6.4 Performance of the proposed speech enhancement algorithm for three different noise types in terms of the improved SSNR

6.5 Performance of the proposed speech enhancement algorithm in a White noise: The enhanced cepstral distance in terms of $\alpha$ taken three different values

6.6 Block diagram of the proposed speech enhancement implementation

6.7 Decomposition tree of the filter bank. H and L respectively represent the low-pass and high –pass filters that cascaded with down 2 perations

6.8 Plot of sigmoid function for $\alpha$ takes three different values

6.9 (a) Clean (red)-Noisy (blue) speech signal (WGN 10dB), (b) Enhanced (red)-Noisy (blue) speech signal, (c) Spectrogram of the noisy signal, (d) Spectrogram of the enhanced signal

6.10 (a) Clean-Noisy speech signal (Car noise 0dB), (b) Enhanced-Noisy speech signal, (c) Spectrogram of the noisy signal, (d) Spectrogram of the enhanced signal (Applying the proposed system using DWT)

6.11 (a) Clean-Noisy speech signal (WGN 10dB), (b) Enhanced-Noisy speech signal, (c) Spectrogram of the noisy signal, (d) Spectrogram of the enhanced signal (Applying the proposed system using DWT)

6.12 (a) Clean-Noisy speech signal (F16 cockpit noise 7dB), (b) Enhanced-Noisy speech signal, (c) Spectrogram of the noisy signal, (d) Spectrogram of the enhanced signal (Applying the proposed system using SGWT)
6.13 Performance of the proposed system using two noise estimation techniques and DWT with White noise .................................................................121
6.14 Performance of the proposed system using two noise estimation techniques and DWT with Pink noise ...............................................................121
6.15 Performance of the proposed system using two noise estimation techniques and SGWT with White noise .........................................................122
6.16 Performance of the proposed system using two noise estimation techniques and SGWT with Pink noise ...........................................................122
6.17 Comparative performance evaluation for the proposed with other three systems using DWT and White noise .........................................................123
6.18 Comparative performance evaluation for the proposed with other three systems using DWT and Pink noise .........................................................124
6.19 Comparative performance evaluation for the proposed with other three systems using SGWT and White noise .....................................................125
6.20 Comparative performance evaluation for the proposed with other three systems using SGWT and Pink noise .........................................................126
6.21 System performance for different decomposition bands, using DWT with White noise .........................................................................................127
6.22 System performance for different decomposition bands, using DWT with F16 cockpit noise ..............................................................................127
6.23 System performance evaluation in term of GSNG using different DWT filters .................................................................................................128
6.24 System performance evaluation in term of Improved ASSNR using different DWT filters .................................................................................128
6.25 System performance evaluation in term of GSNG using different SGWT filters .................................................................................................129
6.26 System performance evaluation in term of Improved ASSNR using different SGWT filters .................................................................................129
6.27 Flow diagram of the computer based graphical user interface GUI ..........133
6.28 Main Interface window ........................................................................134
6.29 Interface window of the evaluation system for the noise estimation techniques .........................................................................................135
6.30 Interface window of the evaluation system for the speech enhancement based speech classification and noise estimation ..........................136
6.31 Interface window for speech classification test system ................................137
6.32 Interface window of the evaluation system for speech enhancement based BS-WPD ..........................................................138
List of Tables

2.1 Comparative performance of objective measures………………………………31
2.2 Description in the mean openion score (MOS)…………………………………32
3.1 BS-WPD approximated and true critical bands………………………………45
4.1 Average relative error ARE in band-2 SGWT for the four noise estimation methods………………………………………………………………………66
4.2 Average relative error ARE in band-2 SGWT for the four noise estimation methods………………………………………………………………………66
4.3 Average relative error ARE in band-3 SGWT for the four noise estimation methods………………………………………………………………………67
4.4 Average relative error ARE in band-3 SGWT for the four noise estimation methods………………………………………………………………………67
4.5 Average relative error ARE in band-7 PWPD for the four noise estimation methods………………………………………………………………………67
4.6 Average relative error ARE in band-15 PWPD for the four noise estimation methods………………………………………………………………………67
4.7 Comparison of cepstral distance for the noise estimation algorithms using SGWT and Soft thresholding (AWGN)…………………………………..69
4.8 Comparison of cepstral distance for the noise estimation algorithms using SGWT and Soft thresholding (Pink noise)………………………………..69
4.9 Comparison of cepstral distance for the noise estimation algorithms using SGWT and Weiner gain thresholding (F16 cockpit noise)………………70
4.10 Comparison of cepstral distance for the noise estimation algorithms using PWPD and Soft thresholding (AWGN)………………………………………70
4.11 Comparison of cepstral distance for the noise estimation algorithms using PWPD and soft thresholding (Pink noise)……………………………………70
4.12 Comparison of cepstral distance for the noise estimation algorithms using PWPD and Weiner gain thresholding (F16 cockpit noise)…………………70
5.1 Performance of the proposed algorithm at different SNRs for three different
male speakers M1, M2 and M3 ................................................................. 88

5.2 Performance of the proposed algorithm at different SNRs for three different
female speakers F1, F2 and F3 ................................................................. 88

6.1 Comparison of Cepstral distances resulting from the application of the Classical
Soft Thresholding and proposed algorithm on a Male speech corrupted by
White noise .......................................................................................... 98

6.2 Comparison of Cepstral distances resulting from the application of the
Classical Soft Thresholding and proposed algorithm on a Female speech
corrupted by White noise ..................................................................... 99

6.3 Comparison of Cepstral distances resulting from the application of the Classical
Soft Thresholding and proposed algorithm on speech corrupted by pink noise.. 99

6.4 Comparison of Cepstral distances resulting from the application of the Classical
Soft Thresholding and proposed algorithm on speech corrupted by F16 cockpit noise ............................................................................. 100

6.5 Per band average energy distribution for Pink noise using 5-levels DWT as
obtained for voiced segments .................................................................. 104

6.6 Per band average energy distribution for Pink noise using 5-levels DWT as
obtained for unvoiced segments ................................................................ 104

6.7 Per band average energy distribution for Car(Volvo) noise using 5-levels
DWT as obtained for voiced segments ..................................................... 104

6.8 Per band average energy distribution for Car(Volvo) noise using 5-levels
DWT as obtained for unvoiced segments ................................................ 104

6.9 Per band average energy distribution for F16 cockpit noise using 5-levels
DWT as obtained for voiced segments ..................................................... 105

6.10 Per band average energy distribution for F16 cockpit noise using 5-levels
DWT as obtained for unvoiced segments ................................................. 105

6.11 Outline of the V/UV speech classification algorithm based wavelet pitch
Detection .................................................................................................. 105

6.12 Performance of the proposed system using DWT .................................. 114

6.13 Performance of the proposed system using SGWT ............................... 115
6.14 Subjective speech quality evaluation with White noise…………………….115
6.15 Subjective speech quality evaluation with Pink noise………………………116
6.16 Subjective speech quality evaluation with Car noise………………………116
6.17 Subjective speech quality evaluation with F16 Cockpit noise………………116
List of Symbols

\( \Gamma \)  
Final indexes set \((l,n,k)\) of wavelet coefficients

\( \Theta \)  
Expansion coefficients \(\{ \theta_{l,n,k} \}\) of the unknown signal \(f(t)\)

\( \alpha(p) \)  
Smoothing subband \(i\) parameter at frame \(p\)

\( \delta \)  
Centre-offset of the sigmoid function

\( \varepsilon \)  
Slop transition factor of the sigmoid function.

\( \gamma_k \)  
Posteriori signal to noise ratio

\( \left( \gamma^{i,n}(k) \right)_{\text{post}} \)  
Posteriori signal-to-noise ratio per segment \(p\)

\( \left( \gamma^{i,n}(k) \right)_{\text{prior}} \)  
Priori signal-to-noise ratio per segment \(p\)

\( \lambda \)  
Threshold value

\( \lambda_V(K) \)  
Variance of the noise \(k\)-th spectral components

\( \lambda_F(K) \)  
Variance of the speech signal \(k\)-th spectral components

\( v_k \)  
Minimum limit of the spectral gain function

\( \theta_{l,n,k} \)  
Expansion coefficients of the unknown clean signal

\( \rho \)  
Correlation coefficient

\( \sigma^2 \)  
Noise variance

\( \sigma_{\lambda}(p) \)  
Subband \(i\) noisy signal power per segment \(p\)

\( \sigma_{\lambda}^2(p) \)  
Subband \(i\) clean signal power per segment \(p\)

\( \hat{\sigma}_{\lambda}^2(p) \)  
Estimated subband \(i\) noise power per segment \(p\)

\( \tilde{\sigma}_{\lambda}^2 \)  
Initial subband \(i\) noise power estimate
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau )</td>
<td>Compensation factor of the sigmoid function</td>
</tr>
<tr>
<td>( \xi_k )</td>
<td>Priori signal to noise ratio</td>
</tr>
<tr>
<td>( \psi(t) )</td>
<td>Mother wavelet function</td>
</tr>
<tr>
<td>( \phi(t) )</td>
<td>Scaling wavelet function</td>
</tr>
<tr>
<td>( a, b )</td>
<td>Translation and shift factors</td>
</tr>
<tr>
<td>( B_{1,n} )</td>
<td>Bandwidth of BS-WPD subband</td>
</tr>
<tr>
<td>( c )</td>
<td>Critical band rate in Bark scale</td>
</tr>
<tr>
<td>( c_f )</td>
<td>LPC cepstral coefficients of the original signal</td>
</tr>
<tr>
<td>( c_y )</td>
<td>LPC cepstral coefficients of the processed signal</td>
</tr>
<tr>
<td>( c_{L,k} )</td>
<td>Scaling coefficient</td>
</tr>
<tr>
<td>( d_{j,k} )</td>
<td>Wavelet coefficient.</td>
</tr>
<tr>
<td>( \hat{E} )</td>
<td>Estimated noise spectrum</td>
</tr>
<tr>
<td>( E_k )</td>
<td>White noise k-th Fourier expansion coefficients</td>
</tr>
<tr>
<td>( E_m )</td>
<td>Set of terminal nodes of an expansion (tree-set)</td>
</tr>
<tr>
<td>( e )</td>
<td>White Gaussian noise</td>
</tr>
<tr>
<td>( e(n) )</td>
<td>Noise data samples</td>
</tr>
<tr>
<td>( E_{1_{L,k}} )</td>
<td>Mean square error of the wavelet coefficients</td>
</tr>
<tr>
<td>( E_{2_{L,k}} )</td>
<td>Mean square error of the scaling coefficients</td>
</tr>
<tr>
<td>( EG_p )</td>
<td>Total energy in the pth speech segment</td>
</tr>
<tr>
<td>( EG_i )</td>
<td>Segment energy in wavelet band I</td>
</tr>
<tr>
<td>( \hat{F} )</td>
<td>Fourier coefficients of the enhanced data</td>
</tr>
<tr>
<td>( f )</td>
<td>Clean data</td>
</tr>
<tr>
<td>( \hat{f} )</td>
<td>Estimate of f</td>
</tr>
</tbody>
</table>
$F_k$      Clean speech $k$-th Fourier expansion coefficients

$F_{SNR}$  Estimated SNR based threshold gain factor

$F_{V/UV}$ Speech classification based threshold gain factor

$f_{1,n}$  Centre frequency of BS-WPD subband

$f(n)$    Clean data samples

$G(Y, \hat{E})$ Spectral gain function as a function of noisy data and estimated noise

$G_i$     Subband denoising gain

$H_0$     Wavelet low pass filter gain

$H_1$     Wavelet high pass filter gain

$h, g$    High and low FIR synthesizing filters

$\tilde{h}, \tilde{g}$ High and low FIR analysing filters

$h_e, h_o$ Even and odd coefficients of the high pass filter.

$I$       Identity matrix

$k$       Time-domain position index for classical wavelet decomposition and (WPD) or frequency domain index for DFT

$L^2(R)$  Square-real function

$\{l, n\}$ Index of tree-node in WPD tree

$L_{frm}^{l,n}$ Decomposed coefficient frame length

$L_{win}^{l,n}$ Finite window observation of a decomposed coefficient

$M(z)$    Modulation matrix for the analysing side

$N$       Set of naturals $\{1, 2, 3\ldots\}$

$n$       Time index for time domain signal signals
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>Segment index</td>
</tr>
<tr>
<td>$p(z)$</td>
<td>Polyphase matrix</td>
</tr>
<tr>
<td>$q$</td>
<td>Quantile noise estimation integer</td>
</tr>
<tr>
<td>$R_p$</td>
<td>Per segment wavelet bands energy ratio</td>
</tr>
<tr>
<td>$r_f$</td>
<td>LP reflection coefficients of the original signal</td>
</tr>
<tr>
<td>$r_y$</td>
<td>LP reflection coefficients of the processed signal</td>
</tr>
<tr>
<td>$s(z^2)$</td>
<td>Laurent polynomial</td>
</tr>
<tr>
<td>$TH(\cdot)$</td>
<td>Non linear operator performs noise subtraction in wavelet domain</td>
</tr>
<tr>
<td>$V_m$</td>
<td>Closed nested subspace of orthonormal wavelet functions</td>
</tr>
<tr>
<td>$w$</td>
<td>Wavelet expansions coefficients of noisy signal</td>
</tr>
<tr>
<td>$W$</td>
<td>Finite orthonormal wavelet transform matrix</td>
</tr>
<tr>
<td>$Y$</td>
<td>Fourier coefficients of the noisy data</td>
</tr>
<tr>
<td>$Y_k$</td>
<td>Noisy speech $k$-th Fourier expansion coefficients</td>
</tr>
<tr>
<td>$y$</td>
<td>Noisy data</td>
</tr>
<tr>
<td>$y(n)$</td>
<td>Noisy data samples</td>
</tr>
<tr>
<td>$y_{l,n}^{l,n}(k)$</td>
<td>Wavelet packet decomposed coefficient with tree node $(l,n)$</td>
</tr>
<tr>
<td>$Z$</td>
<td>Set of integers ${0, \pm 1, \pm 2 \ldots}$</td>
</tr>
<tr>
<td>$Z$</td>
<td>WGN orthogonal transform</td>
</tr>
<tr>
<td>$ZCR_p$</td>
<td>Zero-crossing rate of the $p$th segment</td>
</tr>
<tr>
<td>$z_{l,n,k}$</td>
<td>Expansion coefficients of the unknown noise signal</td>
</tr>
</tbody>
</table>
# List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARE</td>
<td>Average relative error</td>
</tr>
<tr>
<td>ASSNR</td>
<td>Average segmental signal to noise ratio</td>
</tr>
<tr>
<td>BS-WPD</td>
<td>Bark-scaled wavelet packet decomposition</td>
</tr>
<tr>
<td>CEP</td>
<td>Cepstral distance</td>
</tr>
<tr>
<td>CWT</td>
<td>Classical wavelet transforms</td>
</tr>
<tr>
<td>DAM</td>
<td>Diagnostic acceptability measure</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier transform</td>
</tr>
<tr>
<td>DWT</td>
<td>Discrete wavelet transform</td>
</tr>
<tr>
<td>FIR</td>
<td>Finite impulse response</td>
</tr>
<tr>
<td>FWT</td>
<td>Forward wavelet transform</td>
</tr>
<tr>
<td>GC</td>
<td>Grey code</td>
</tr>
<tr>
<td>GSNR</td>
<td>Global signal to noise ratio</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical user interface</td>
</tr>
<tr>
<td>HOS</td>
<td>Higher order statistics</td>
</tr>
<tr>
<td>IDWT</td>
<td>Inverse discrete wavelet transform</td>
</tr>
<tr>
<td>IFWT</td>
<td>Inverse forward wavelet transform</td>
</tr>
<tr>
<td>Lar</td>
<td>Log area ratio distance</td>
</tr>
<tr>
<td>LP</td>
<td>Linear prediction</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear predictive coding</td>
</tr>
<tr>
<td>MAD</td>
<td>Median value</td>
</tr>
<tr>
<td>MCRA</td>
<td>Minimum controlled recursive averaging</td>
</tr>
<tr>
<td>MRA</td>
<td>Multi resolution analysis</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>MOS</td>
<td>Minimum opinion score</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean square error</td>
</tr>
<tr>
<td>PSD</td>
<td>Power spectral density</td>
</tr>
<tr>
<td>PSQM</td>
<td>Perceptual speech quality measure</td>
</tr>
<tr>
<td>PWPD</td>
<td>Perceptual wavelet packet decomposition</td>
</tr>
<tr>
<td>RASTA</td>
<td>Relative spectral processing</td>
</tr>
<tr>
<td>SGWT</td>
<td>Second Generation Wavelet Transform</td>
</tr>
<tr>
<td>SegSNR</td>
<td>Segmental signal to noise ratio</td>
</tr>
<tr>
<td>STFT</td>
<td>Short time Fourier transform</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to noise ratio</td>
</tr>
<tr>
<td>STSA</td>
<td>Short time spectral amplitude</td>
</tr>
<tr>
<td>V/UV</td>
<td>Voiced/Unvoiced segment</td>
</tr>
<tr>
<td>WCT</td>
<td>Wavelet coefficient threshold</td>
</tr>
<tr>
<td>WGN</td>
<td>White Gaussian noise</td>
</tr>
<tr>
<td>WP</td>
<td>Wavelet Packet</td>
</tr>
<tr>
<td>WPD</td>
<td>Wavelet packet decomposition</td>
</tr>
<tr>
<td>WT</td>
<td>Wavelet transform</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Motivation

The problem of enhancing speech degraded by uncorrelated additive noise, when the noisy speech alone is available, has received much attention. This is due to a variety of potential applications in speech enhancement possesses. Furthermore, technologies enabling the implementation of such intricate algorithms are now available. The purpose of the enhancement process, which is performed mainly by denoising, is to improve the quality and comprehension of speech. It is also useful to enhance the speech prior to the implementation of techniques such as coding and recognition.

In general, the need exists for digital voice communications or automatic speech recognition systems to perform reliably in noisy environments. For example in hands-free operation of cellular phones in vehicles, the transmitted speech signal is often contaminated by background noise and reverberation. In many cases, these systems work well in nearly noise-free conditions, however their performance deteriorates rapidly in noisy environments. Therefore, development of appropriate algorithms for reducing noise-related degradation in speech signals is of current interest. With the wide dissemination of mobile communications and the combination of this technology with consumer electronics technologies, the need for practical and powerful speech enhancement systems has grown considerably. Significant progress in this field has been made [1-3], yet the development of a practical speech enhancement system that is robust to adverse conditions and non-stationary noise is still a challenge.
The overall aim of this research is to investigate the development and assessment of a new, practical speech enhancement system that will perform reliably under non-stationary, real-life noise environments, which are encountered in modern digital voice communications systems.

1.2 Speech Characteristics and Modelling

Speech is a sound signal, which conveys information in human communication. Linguistic information in speech involves voiced speech, unvoiced speech or plosive sounds. Moreover, there are some parts of the speech that are neither pure voiced nor pure unvoiced, but a mixture of the two. These are transition regions, which involve changes either from voiced to unvoiced or vice versa.

Voiced speech segments are characterized by relatively high-energy content, but more importantly they contain periodicity, which is called the *pitch* of voiced speech. The unvoiced part of speech, on the other hand, looks more like random coloured noise with no periodicity. Unvoiced sounds are generated by forming a constriction at some point in the vocal tract and forcing air through the constriction at a high enough velocity to produce turbulence. Plosive sounds result from making a complete closure, building up pressure behind the closure, and abruptly releasing it [5].

A short time segment of a speech signal can be regarded as a portion of a stationary stochastic process, therefore generally a speech signal is said to be quasi-stationary, i.e. it can be divided into (almost) stationary segments; which are typically 20-50 ms long [4].
1.3 Speech Enhancement Techniques: An Overview

1.3.1 Frequency and time domain based approach

Speech communication under noisy condition is difficult and fatiguing. Speech enhancement is defined as the attempt to improve the performance of speech communication systems when their input or output is corrupted by noise. The main purpose of many enhancement algorithms is to reduce background noise, improve speech quality, or suppress channel or speaker interference. Depending on the specific application, a system may be directed at one or more objectives, such as improving overall quality, increasing intelligibility, or reducing listener fatigue. The objective of achieving higher quality and/or intelligibility of noisy speech may also contribute to improved performance in other speech applications, such as speech compression, speech coding, speech recognition, voice authentication or speaker verification.

Over the last few decades, different approaches to speech enhancement for various speech processing applications have been adopted. The approaches can, generally, be classified into two major categories of single-microphone and multi-microphone methods. Although multi-microphone algorithms have acceptable performance in some applications, limitations to the use of more than one microphone still exist in many practical situations. Among various single-microphone algorithms for speech enhancement, spectral subtraction has been mostly employed mainly due to its relative simplicity [5]. However, despite its proven capability in removing background noise, spectral subtraction produces annoying artefacts known as the musical noise [6]. Several solutions have been proposed to reduce this effect [7-8], however most of these approaches are computationally involved and often not suitable for real-time applications. The robustness of spectral subtraction techniques is dependent on the effectiveness of a pause detection method, which is traditionally performed by a voice activity detector (VAD). When the signal to noise ratio (SNR) is low or the disturbing noise is non-stationary, the performance of the above detection method deteriorates resulting in speech distortion and unnatural sounding or fluctuating residual background noises. In recent years, several approaches such as adaptive iterative Wiener filtering, HMM-based algorithms [9-10] and signal subspace methods [11] have been proposed as alternatives to spectral subtraction. In
most speech enhancement techniques, the corrupting background noise is usually assumed to be stationary, or at least changes rather slowly compared to the modulation of speech. Based on this, a different approach to speech enhancement, known as “Relative Spectral Processing” (RASTA) has been also proposed [12]. RASTA processing based on band-pass filtering is applied to the power spectrum of speech while keeping the phase of the original signal. The filter features of the RASTA are usually derived from the STFTM (Short Time Fourier Transformation Modulation) or LPC (Linear Predictive Coding) envelope [13].

Time-domain nonlinear filtering methods, that utilize data sets where the clean speech is used as a target signal to mostly train neural networks, also exist. For example in [14], a neural network based time-domain method called Dual Extended Kalman Filtering (DEKF) has been explored, based on dual Kalman estimation of both states and weights, for efficient maximum-likelihood optimisation (in the context of robust nonlinear prediction, estimation, and smoothing).

A practical frequency domain speech enhancement system generally consists of two major components: a spectral analysis system and noise power spectral density (PSD) estimation algorithm. The former is well understood [1] and is easily implemented using approaches usually based on the short-time Fourier transform (STFT). Adaptive techniques have been presented based on temporal processing of the signal in adverse environments [15]. The STFT, however, has its limitations when dealing with non-stationary signals and signals with discontinuities, such as speech. These limitations often have adverse effect on the speech enhancement results. On the other hand, the noise estimation is a very important component in the enhancement system, yet, has frequently received less attention. It has a major impact on the overall performance of the speech enhancement system. If the noise estimate is too low, unnatural residual noise will be perceived. If the estimate is too high, speech sounds will be muffled and intelligibility will be lost. The traditional VAD based SNR estimators are difficult to tune and their application to low SNR speech often results in clipped speech [16]. Current research aims therefore at incorporating soft-decision schemes, which are also capable of up-dating the noise PSD during speech activity [17].
1.3.2 Wavelet based approaches

Recently, the wavelet transform (WT) has emerged as an effective tool for processing non-stationary signals, with superior time-frequency localisation and reduced computational load as compared to the STFT. The WT is a multi-scale analysis which involves decomposing signals into elementary building blocks, known as coefficients, that are well localised in both time and frequency. In particular, the WT analysis is very beneficial for analysing non-stationary signals, such as speech. In fact, the analysis is well suited for speech signals due to its similarity to the way human ear processes sound [18]. Thus by using wavelet transform, a speech signal can be analysed at a specific scale corresponding to the range of human speech.

The WT theory, hence, provided a unified framework for a number of efficient techniques for various speech processing applications [19]. In 1995, Donoho [20] introduced wavelet thresholding (shrinkage) as a powerful tool for denoising signals degraded by additive white noise. Subsequently the potential and efficiency of the application of wavelet shrinkage for speech enhancement has been demonstrated in several works. They include the use of Wiener filtering in wavelet domain [21], wavelet filter bank for spectral subtraction [22], and a coherence function where spectral subtraction has been used in conjunction with different non-uniform filter banks [23]. Other techniques based on adaptive denoising wavelet shrinkage [24,25], or Teager energy operator [26] have also been reported. The definition of the WT, along with a detailed description of its various signal decomposition and implementation methods, is given in Chapter 3.

However, as will be discussed in Chapter 2, there are still a number of problems yet to be solved in order to present a successful potential application of the WT thresholding to enhancement of speech degraded by non-stationary and colored noise. This work is an attempt to address these problems via exploitation of certain features of the speech, and the computational and adaptability advantages of new methods of WT, such as the second-generation wavelet transform SGWT [29] and the perceptual WT [30].
1.4 Scope of The Thesis

Following this introduction, Chapter 2 gives a detailed review of the most common and recently reported speech enhancement techniques, with particular focus on WT-based algorithms, which are the basis of this work. It identifies the drawbacks and shortcomings of the reviewed techniques, especially those based on wavelet thresholding. It also gives a detailed overview of standard measures that have been used in this work to evaluate the objective and subjective qualities of the speech. The Chapter concludes by giving an outline of the proposed algorithm for a new WT-based speech enhancement system using two approaches.

In Chapter 3, a detailed definition of the wavelet transform WT and review the various types of the transform and their implementation are provided. In Chapter 4, comparative performance analysis of three different noise estimation techniques implemented using two wavelet decompositions is introduced. Based on the above study, a fourth new noise estimation technique will be proposed and evaluated.

Chapter 5 presents a new multi-feature algorithm to classify short speech segments into voiced/unvoiced. The algorithm has been implemented and evaluated using CWT and white noisy environments. In Chapter 6 results from the proposed implemented speech enhancement system using the two approaches are presented. The performance of the system for each approach was assessed using different noise types and speech quality measures.

Finally, in Chapter 7 the main conclusion points drawn from the work with a summary and discussion on future research directions are given.
Chapter 2

Speech Enhancement: Review of Existing Approaches

2.1 Introduction

With the advent and wide dissemination of digital voice communication and automatic speech recognition, the need has increased for robust speech enhancement techniques that perform reliably under adverse noisy conditions. Over the last few decades various techniques for speech enhancement have been proposed for a variety of applications and their performance evaluated [31,32]. In general, speech enhancement and denoising techniques can be classified according to their domain of implementation. Based on the most commonly used approaches, two main domains can be identified: The spectral domain and the wavelet domain. In this Chapter, we give detailed overview of principles and theory of existing and recently reported speech enhancement approaches using the above classification. We then identify shortcomings of each, before outlining a new robust, wavelet-based speech enhancement algorithm for non-stationary noise environments.

2.2 Review of Existing Spectral Domain Speech Enhancement Approaches.

2.2.1 Spectral subtraction

Suppose we have a noisy data $y = \{ y(n) \}_{n=0}^{N-1}$ where

$$y(n) = f(n) + e(n) \quad n=0, \ldots, N-1$$

(2.1)
The time-domain signal is first broken up into a series of overlapping frames. Each frame is multiplied by a smooth windowed function (like Hamming) in order to reduce the spectral artefacts caused by the discontinuities at the edges of the frame. Then, each windowed segment is transformed into the spectral domain and processed individually. The spectral coefficients are multiplied by an appropriate gain. The modified spectral components are then transformed back into the time domain, and the frames are assembled to get the enhanced signal.

Transformations into spectral domain can be performed using Discrete Fourier Transform (DFT).

All frequency domain algorithms need to estimate the noise spectrum prior to denoising. If one synthesizes a noisy speech signal in order to verify performance of a given denoising algorithm, it is useful to estimate the spectrum from known noise signal. In Fig 2.1 \( G(Y, \hat{E}) \) is a gain function. The generalised spectral subtraction gain function [6] is given by:

\[
\hat{f} = \frac{Y}{\hat{F}} = \frac{Y}{G(Y, \hat{E})Y}
\]
\[ G_k = G(Y_k, \hat{E}_k) = \begin{cases} 
1 - \alpha \left( \frac{|\hat{E}_k|}{|Y_k|} \right)^{\gamma_1}, & \text{if } \left( \frac{|\hat{E}_k|}{|Y_k|} \right)^{\gamma_1} < \frac{1}{\alpha + \beta} \\
\beta \left( \frac{|\hat{E}_k|}{|Y_k|} \right)^{\gamma_2}, & \text{otherwise} 
\end{cases} \]  

(2.2)

where \( Y_k \) is the \( k \)-th noise speech spectral component, \( \hat{E}_k \) is the estimated \( k \)-th noise spectral component, \( G_k \) is the \( k \)-th spectral component’s gain, \( \alpha \) is the\textit{ oversubtraction factor} (\( \alpha > 1 \) leads to the reduction of residual noise but also to increased speech distortion), \( \beta \) is \textit{spectral flooring factor} (\( 0 \leq \beta \leq 1 \) permits a certain level of background noise), \( \gamma_1 \) and \( \gamma_2 \) are exponent parameters which determine the sharpness of the transition. A number of classical gain functions can be derived from (2.2) by appropriate choice of the parameters:

\textit{a. Amplitude spectral subtraction or magnitude subtraction} \([6]\):

\[ G_k = G(Y_k, \hat{E}_k) = \begin{cases} 
1 - \alpha \left( \frac{|\hat{E}_k|}{|Y_k|} \right), & \text{if } \left( \frac{|\hat{E}_k|}{|Y_k|} \right) < 1 \\
0, & \text{otherwise} 
\end{cases} \]  

(2.3)

\textit{b. Power spectral subtraction}:

\[ G_k = G(Y_k, \hat{E}_k) = \begin{cases} 
1 - \alpha \left( \frac{|\hat{E}_k|^2}{|Y_k|^2} \right)^{1/2}, & \text{if } \left( \frac{|\hat{E}_k|^2}{|Y_k|^2} \right) < 1 \\
0, & \text{otherwise} 
\end{cases} \]  

(2.4)
c. Weiner estimator:

\[
G_k = G(Y_k, \hat{E}_k) = \begin{cases} 
1 - \alpha \frac{\left| \hat{E}_k \right|^2}{\left| Y_k \right|^2}, & \left| \frac{\hat{E}_k}{Y_k} \right|^2 < 1 \\
0, & \text{otherwise}
\end{cases}
\]  

The above spectral subtraction gain functions can also be thought of as thresholding algorithms.

2.2.2 Short Time Spectral Amplitude (STSA)

Ephraim – Malah [1] proposed an algorithm that capitalizes on the major importance of STSA of the speech signal to its perception. It is well known that a distortion measure, which is based on the mean squared error of the log-spectra, is more suitable for speech processing. Thus the proposed algorithm for speech enhancement is the STSA estimator, which minimizes the mean squared error of the log-spectra [33]. The estimation problem of the STSA is formulated in a similar fashion to that of estimating the amplitude of each Fourier expansion coefficient of the speech signal.

Let \( F_k = A_k e^{j\alpha_k} \) and \( Y_k = R_k e^{j\beta_k} \), denote the \( k \)-th Fourier expansion coefficients of the clean speech signal and the noisy speech in the analysis, respectively. Also the \( k \)-th Fourier expansion coefficient of the noise signal is denoted by \( E_k \). The estimator \( \hat{A}_k \), which minimizes the distortion measure: 

\[
E \left\{ \left| \log A_k - \log \hat{A}_k \right|^2 \right\}
\]

can be calculated as:

\[
\hat{A}_k = \frac{\bar{\xi}_k}{1 + \bar{\xi}_k} \exp \left\{ \frac{1}{2} \int_{v_c}^{T} \frac{e^{-t}}{t} dt \right\} R_k
\]

(2.6)

Here, \( \bar{\xi}_k \) is the so called a priori signal to noise ratio (SNR):
\[ \xi_k = \frac{\lambda_F(k)}{\lambda_E(k)} \]  

(2.7)

\( \gamma_k \) is the so called a posteriori SNR:

\[ \gamma_k = \frac{R^2_k}{\lambda_E(k)} \]  

(2.8)

and \( v_k \) is defined by

\[ v_k = \frac{\xi_k}{1 + \xi_k} \gamma_k \]  

(2.9)

where \( \lambda_E(k) = E\{\|E_k\|^2\} \) and \( \lambda_F(k) = E\{\|F_k\|^2\} \) are the variances of the noise and the enhanced signal \( k \)-th spectral components. From (2.6), it can be seen that \( \hat{A}_k \) is obtained from \( R_k \) by a multiplicative nonlinear gain function which depends only on the a priori and the posteriori SNR’s. The gain function is defined by:

\[ G_k(\xi_k, \gamma_k) = \frac{\hat{A}_k}{R_k} = \frac{\xi_k}{1 + \xi_k} \exp \left\{ \frac{1}{2} \int_{-\infty}^{\infty} e^{-t} \, dt \right\} \]  

(2.10)

There are two main approaches, which can be used to estimate \( \xi_k \) of \( \xi \) [33]:

a. Maximum Likelihood Estimation Approach:

\[ \hat{\xi}_k = \begin{cases} \hat{\gamma}_k(p) - 1, & \hat{\gamma}_k(p) - 1 \geq 0 \\ 0, & \text{otherwise} \end{cases} \]  

(2.11)

where

\[ \hat{\gamma}_k(p) = \alpha \hat{\gamma}_k(p - 1) + (1 - \alpha) \frac{\gamma_k(p)}{\beta}, \quad 0 \leq \alpha < 1, \quad \beta \geq 1 \]  

(2.12)

Here \( \beta \) is the correction factor, \( \alpha \) is the smoothing parameter and \( p \) is the segment index.
b. Decision Directed Estimation Approach:

\[
\hat{z}_k = \alpha \frac{\hat{A}_k(p-1)}{\lambda_z(k, p-1)} + (1 - \alpha) \min[(\gamma_k(p) - 1), 0], \quad 0 \leq \alpha < 1
\]  

(2.13)

Most of the spectral speech enhancement approaches are implemented using discrete Fourier transform (DFT) since a good separation between noise and signal in the frequency domain can be obtained. A major drawback of the basic spectral subtraction is the residual tonal noise, which occurs because of estimation errors of the actual magnitude of the noise spectrum. In general such approach is appropriate when both noise and signal are stationary. Such assumption is definitely not true for human speech, which is non-stationary. To some extent the problem could be reduced through the use of short time Fourier transform (STFT). The speech here is divided into short frames during which the signal is assumed to be near stationary. The STSA developed by Ephraim and Malah [1,33] is based on short time spectral framing and has been proved to be efficient in reducing the musical noise for speech signals. Unfortunately the non-optimal SNR quantities affect the multiplicative gain and deteriorate the performance of the speech enhancement system.

### 2.3 Wavelet Domain Speech Enhancement Approaches

Compared to the DFT, the STFT provides a better time resolution but poorer frequency resolution. On the other hand Wavelet Transform (WT) has the advantage of using variable window size for different frequency components. This allows the use of long time intervals to obtain more precise low frequency information and shorter intervals for high frequency information. This often results in better handling of non-stationary data like speech. Unlike the STFT, where the processing interval is fixed once the frame length is decided upon, there is much more flexibility in the implementation of the WT. Over the last few years, various wavelet domain based approaches have been reported. These approaches are generally based on the following three methods:
2.3.1 Wiener filter-based speech enhancement method

This method is shown diagrammatically described in Fig 2.2. It involves first processing the noisy speech signal using fast wavelet transform decomposition (FWT) to obtain the wavelet coefficients. Then instead of thresholding according to Donoho aspect [20], the WT coefficients are compressed using Weiner filtering in the wavelet domain [21,34]. The noise power is estimated using a pause or voice activity detector (VAD), and used to calculate the Weiner gain for each band. The wavelet coefficients are then scaled by the corresponding gain value in each band.

![Diagram of speech enhancement system based on Weiner filtering](image)

*Figure 2.2: Speech enhancement system based on Weiner filtering*

2.3.2 Adaptive denoising method based on speech classification

This method is outlined by the diagram shown in Fig 2.3. Speech signal segments are first classified into voiced/unvoiced categories. An adaptation for the wavelet thresholding process is then applied, motivated by the fact that unvoiced speech segments should be treated differently from the voiced [25,35]. Applying the same thresholding to speech segments affect the intelligibility in the reconstructed signal, since unvoiced sound contains mainly noise-like high frequency components and eliminating them can cause severe degradation of the composed speech.
2.3.3 Noise estimation based adaptive denoising method

In this method, the noise spectrum is estimated adaptively in each band using either a VAD indicating any silence [25] or directly from the coefficient using an efficient technique [36]. A per-band wavelet threshold, which is dependent mainly on the noise power $\sigma$, is then calculated as shown in Fig 2.4. Some proposed techniques require a pre-processing stage for the input before estimating the noise power spectrum [37]. In other techniques, a time/frequency threshold adaption was introduced for each coefficient using different wavelet decompositions [25,39].

In [24], the time-adapted threshold consumption is introduced, where there is no need for any priori knowledge of the SNR or pause detector. This may enhance the suitability of any speech enhancement technique for real time implementation.
2.4 Shortcomings of Conventional Wavelet-Based Speech Signals Denoising

Some major problems arise when the basic wavelet thresholding method is used to enhance speech signals degraded by real-life noises:

1. The basic method assumes that the noise spectrum is white. However, in most practical situations we are faced with coloured and non-stationary noises. As a result, the basic form of wavelet shrinkage does not provide a satisfactory speech quality. In addition, the method encounters problems in removing non-stationary noises since no time adaptation process is provided in the basic algorithm.

2. Another problem is related to the application of wavelet shrinkage to unvoiced (UV) segments of the speech. Since the UV parts of the speech usually contain large number of noise-like high frequency components, eliminating them in the wavelet domain can cause severe degradation to the intelligibility of the reconstructed speech. In fact, the use of the same threshold for all wavelet bands is another disadvantage for speech applications.
3) Thresholding by setting some wavelet coefficients to zero often results in discontinuities in the enhanced signal. This leads to unwanted artefacts and distortion of the output speech.

Regarding the performance of the methods presented in section 2.3, the following shortcomings can be noted:

a) The estimation of the noise power depends on the performance of the voice activity detector VAD. This may sometimes cause a number of errors if the pause silence detection is not correct.

b) The wavelet coefficients are denoised by the Weiner filter gain, which is obtained for each band based on the estimated noise power. Hence, the reliability of this approach depends on the accuracy of the noise power estimation. Also the same coefficients scaling is applied irrespective of whether the process segments are voiced or invoiced.

Regarding the technique described in Section 2.3.2, the performance depends mainly on the accuracy of the V/UV classification algorithm. For example, in [25] the value of threshold for the V/UV classification decision was selected based on pause detection. Such classification process is often not reliable and may increase classification errors for high-level noise cases. In [35], the classification has been done according to the energy distribution of the wavelet coefficients, since they carry useful information about the voiced/invoiced regions. However, for low SNR and real life non-stationary noise cases, such feature V/UV classification method yields poor performance.

For the technique presented in Section 2.3.3, the performance could be greatly enhanced if a robust and accurate noise power density estimator is used in the algorithm. Different noise power estimation algorithms with varying degrees of performance regarding their noise tracking abilities, computational complexity and real-time implementation, have been recently reported. To select an appropriate noise power estimator, these algorithms can be implemented and evaluated in the wavelet domain; as will be reported in Chapter 4.
2.5 Outline of the Proposed Speech Enhancement System

Based on the reviews and discussions presented in preceding sections, this Section proposes a novel, practical speech enhancement system for non-stationary, real-life noise environments, which are encountered in modern digital voice communications systems.

The system is depicted in the diagram given in Fig 2.5. An outline description of the main stages of the system is given here:

A. First, the sampled input speech signal $y(n)$, where $n$ is the discrete time index, is pre-emphasised and segmented using an appropriate window.

B. The segmented speech, $y'(n)$, will then be transformed into the wavelet domain using three different wavelet decompositions, Bark-Scaled wavelet packet decomposition BS-WPD, discrete wavelet transform DWT and second generation wavelet transform SGWT. The principles and definitions of the wavelet transform, as well as description of the various decomposition procedures and their implementation will be presented in Chapter 3.

C. Threshold setting: a threshold setting process based mainly on estimation of the noise power will be used here. This will be performed by computing the standard deviation of the noise for energy wavelet band $i$ and time segment $p$, $\sigma_i(p)$. Using the estimated values of $\sigma_i(p)$, an appropriate wavelet threshold $TH_i(p)$, will be calculated for each band of each frame. In our work, this processing stage will be implemented by first investigating the implementation and performance of a number of recently reported noise estimation algorithms, using two different wavelet decompositions as will be presented in Chapter 4. An appropriate noise estimation algorithm will then be selected for the threshold setting using criteria based on accuracy, robustness and computational burden.

However, the problem of over-filtering of high frequency components of the unvoiced (UV) speech segments has also to be also addressed here. Accordingly, we will also investigate employing a voiced/unvoiced (U/UV) classification algorithm, using a new multi-feature algorithm developed during
the course of this work-see Chapter 5, for the adaptation of $TH_i (p)$ in order to alleviate this problem.

D. In this stage, the wavelet shrinkage will be performed on the BS-WPD, DWT or SGWT coefficients using the adapted threshold values from step (C) above. The work will investigate different types of classical and developed thresholding techniques. Specifically, instead of setting certain wavelet coefficients to zero, which may cause noticeable sharp time-frequency discontinuities in the reconstructed speech, we will investigate attenuating the coefficients that are smaller than the threshold in an appropriate manner that would not create abrupt changes.

E. Finally in this stage, the speech signal is reconstructed by applying an inverse BS-WPD, IDWT or inverse SGWT.

Figure 2.5: Block diagram of the proposed speech enhancement system

Full detailed description of the full implementation of the proposed speech enhancement system will be presented in Chapter 6.
2.6 Speech Quality Measures

The evaluation of speech quality is of critical importance in the field of speech enhancement. The purpose of the evaluation is to quantify limitations and to identify the limitations responsible for the loss in intelligibility. For assessment of speech communication systems mainly three major evaluation methods are used:

a. Objective measures based on physical properties of the speech and the transmission system. These measures, which measure the difference between the input (original) and output (degraded) signals of speech systems are widely used.

b. Subjective listening measures based on scores for correct recall by a group of listeners to sentences, words or phonemes. These measures are used to assess the intelligibility and perception of auditory system.

2.6.1 Objective speech measures

There have been numerous objective measures suggested and used for evaluation of speech coding systems [38]. Not only is it necessary to have consistent subjective assessments, via subjective tests, for the comparative assessments of alternative coders, it is also essential to have an objective distortion measure which, during the development phase, can give the designer an immediate and reliable estimate of the anticipated perceptual quality of a particular coding algorithm. Such measures include signal-to-noise ratio (SNR), arithmetic and geometric spectral distance measures, cepstral distance measures, various parametric distance measures such as pseudo area functions and log area ratios using linear prediction analysis, as well as many more. To measure the speech quality of a distorted sentence as compared with the original one, we assume that the waveforms of the two sentences are synchronised. We divide the two sentences into frames. For each measure $s$, a local function $f(n)$ is computed for each frame $n$ and the speech measure is defined as the time average of the local function over all the frames.
Objective methods rely on a mathematically based measure between the original and degraded speech signals. The success of these measures rests with their correlation with subjective quality. Definitions of the various objective measures that have been used in this work [39,40] are given in the following sections:

a. **Global signal-to-noise ratio (GSNR):**

\[
\text{GSNR} = 10 \log_{10} \frac{\sum_{n} f^2(n)}{\sum_{n} (f(n) - y(n))^2}
\]  

(2.15)

where \( f(n) \) is sampled enhanced speech signal, \( y(n) \) is the noisy speech signal, and \( n \) is the sample index.

b. **Segmental signal-to-noise ratio (SegSNR):**

\[
\text{SegSNR} = \frac{10}{P} \sum_{p=0}^{P-1} \log_{10} \frac{1}{N} \sum_{n=1}^{N} \left( \frac{f^2(n)}{[y(n) - f(n)]^2} \right)
\]  

(2.16)

where \( N \) is the segment length and \( P \) is the number of segments in the speech signal.

c. **Log-area-ratio (lar):**

The lar measure is based on the dissimilarity of the linear prediction LP coefficients between the original and the processed speech signals. The log-area-ratio parameters are obtained from the \( P^{th} \) order LP reflection coefficients of the original \( r_f(p) \) and processed \( r_y(p) \) signals for frame \( p \). The objective measure can be formed as follows:
\[
lar = \left[ \frac{1}{|F|} \sum_{p=1}^{F} \left( \log \frac{1 + r_f(p)}{1 - r_f(p)} - \log \frac{1 + r_d(p)}{1 - r_d(p)} \right)^2 \right]^{1/2}
\]

(2.17)

d. **Cepstral distance (Cep)**

Linear prediction coefficients (LPC) also can be used to compute a distance measure based on cepstral coefficients. Unlike the cepstrum computed directly from speech waveform, one computed from the predictor coefficients provides an estimate of the smoothed speech spectrum. A measure based on cepstral coefficients for frame \( p \) can be computed as:

\[
Cep(p) = \left[ (c_f(0) - c_y(0))^2 + 2 \sum_{k=1}^{L+1} [c_f(k) - c_y(k)]^2 \right]^{1/2}
\]

(2.18)

where \( c_f(k) \) and \( c_y(k) \) are the LPC cepstral coefficients for the original and the distorted speech, respectively. In (2.19) only the sums over the first \( L+1 \) coefficients differences are used. In practise, the value of \( L \) can be small as the order of the LPC analysis.

Table 2.1 lists the performance of different speech quality objective measures in terms of correlation coefficient \( \rho \). Such a coefficient is based on statistical analysis performed between the objective measures and the subjective mean opinion score (MOS) on a designed speech database [41].

### 2.6.2 Subjective speech measures

In order to develop an objective measure that correlates well with subjective quality assessments, a perceptually oriented subjective measure is required. Two commonly used measures are the mean opinion score (MOS) and the diagnostic acceptability measure (DAM) [40].
MOS scores require lengthy subjective testing, but are widely accepted as a norm for comparative rating of different systems. The automatic prediction of MOS scores directly from the speech signals and without human subjects could, therefore, be of great practical value. The rating scale employed in MOS testing is illustrated in Table 2.2 along with a general description of the levels of distortion typically associated with each numerical score. An MOS score is mapping of perceived levels of distortion into either the descriptive terms “excellent, good, fair, poor, unsatisfactory,” or into equivalent numerical ratings in the range 5-1.

2.6.2.1 Perceptual speech quality measurement (PSQM)

PSQM is a mathematical process that provides a measurement of the subjective quality of speech. The objective of PSQM is to produce scores that reliably predict the results of subjective tests. PSQM scores reflect the amount of divergence from a clean signal that a distorted signal exhibits. The PSQM testing process is shown in Fig 2.6. PSQM is approved by the ITU-T and published by the ITU as a standard recommendation [42].

To perform PSQM measurement, a sample of recorded human speech is fed into a speech system and is processed by whatever communication system is using. The output signal is recorded as it is received. It is then time-synchronized with the input signal, and compared to the input using. This comparison is performed on individual time segments (or frames) acting on parameters derived from spectral power densities of the input and output time-frequency components. The comparison is based on factors of human perception, such as frequency and loudness sensitivities, rather than on simple power spectral densities (PSD). The resulting PSQM score range from 0 to infinity, representing the perceptual distance between the input and output signals. At the final stage, the PSQM scale is mapped from its objective scale to the 5-1 subjective MOS scale.
Table 2.1. Comparative performance of objective measures

<table>
<thead>
<tr>
<th>Objective speech measure</th>
<th>Correlation factor $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNR-based measure</strong></td>
<td></td>
</tr>
<tr>
<td>SNR</td>
<td>0.24</td>
</tr>
<tr>
<td>Segmental SNR</td>
<td>0.77</td>
</tr>
<tr>
<td>Frequency variant SNR</td>
<td>0.43</td>
</tr>
<tr>
<td>Frequency variant Segmental SNR</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>LPC-based measure</strong></td>
<td></td>
</tr>
<tr>
<td>LPC</td>
<td>0.41</td>
</tr>
<tr>
<td>Log LPC</td>
<td>0.34</td>
</tr>
<tr>
<td>Linear reflection coefficient</td>
<td>0.46</td>
</tr>
<tr>
<td>Log reflection coefficient</td>
<td>0.11</td>
</tr>
<tr>
<td>Linear area ratio</td>
<td>0.24</td>
</tr>
<tr>
<td>Log area ratio</td>
<td>0.62</td>
</tr>
<tr>
<td>Itakura energy distance</td>
<td>0.59</td>
</tr>
<tr>
<td>Cepstrum distance</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>Linear spectral distance</strong></td>
<td></td>
</tr>
<tr>
<td>Inverse linear spectral distance</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Log spectral distance</strong></td>
<td></td>
</tr>
<tr>
<td>Log spectral distance</td>
<td>0.60</td>
</tr>
<tr>
<td><strong>Nonlinear spectral distance</strong></td>
<td></td>
</tr>
<tr>
<td>Nonlinear spectral distance</td>
<td>0.61</td>
</tr>
<tr>
<td><strong>Weighted-slope spectral distance</strong></td>
<td></td>
</tr>
<tr>
<td>Weighted-slope spectral distance</td>
<td>0.74</td>
</tr>
</tbody>
</table>
Table 2.2 Description in the mean opinion score (MOS)

<table>
<thead>
<tr>
<th>Rating</th>
<th>Speech quality</th>
<th>Level of distortion</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Excellent</td>
<td>Imperceptible</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
<td>Just perceptible but not annoying</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
<td>Perceptible and slightly annoying</td>
</tr>
<tr>
<td>2</td>
<td>Poor</td>
<td>Annoying but not objectionable</td>
</tr>
<tr>
<td>1</td>
<td>Unsatisfactory</td>
<td>Very annoying and objectionable</td>
</tr>
</tbody>
</table>

Figure 2.6: PSQM Testing Process
2.7 Summary

At the beginning of this Chapter, an overview of the basic speech enhancement systems in different domains is presented. The main spectral speech enhancement approaches are described in section 2.2. In section 2.3, the different wavelet based speech enhancement approaches are studied and their performance discussed in section 2.4.

Objective and subjective speech quality measures have been explained in section 2.5, with more details about the PSQM are given.

The following Chapter deals with the basic aspects of WT and gives descriptions for the recent wavelet decompositions. Chapter 3 deals also with the principle of wavelet thresholding and present different thresholding techniques.
Chapter 3

Wavelet Transforms

3.1 Mathematical Description of The Wavelet Transform

Wavelet Transform is a relatively recent development from functional analysis that has attracted great attention from the signal processing community. Its fundamental property is that it allows for efficient time-frequency representations of signals and data sets, which can be computed fast. In other words wavelets are capable of quickly capturing the essence of a data set using only a small set of coefficients. This is based on the fact that most data sets have correlation both in time and frequency. The continuous wavelet transform (CWT) is based on the correlation of a given continuous-time signal with a set of filters defined by a family of basis wavelet functions. The two-dimensional parameterisation is achieved by generating a set of basis wavelets by dilation and translation of a single function $\psi(t)$, known as the 'mother wavelet', using the following relation:

$$
\psi_{a,b}(t) = \left| a \right|^{1/2} \psi\left(\frac{t-b}{a}\right) 
$$

(3.1)

where $a$ and $b$ are real and continuous parameters known as the scale or dilation factor and the translation or shift factor, respectively, and $\psi_{a,b}(t)$ is the set of wavelet basis functions [27]. A set of wavelet functions in discretized parameters can be obtained by sampling the parameters $a$ and $b$ on a two-dimensional grid. Using a sampling grid defined by $a = a_0^j$ and $b = kb_0a_0^j$, (3.1) yields

$$
\psi_{j,k}(t) = a_0^{-j/2} \psi(a_0^{-j/2}t - kb_0) 
$$

(3.2)
where \( \psi_{j,k}(t) \) is referred to as the basis functions of the discrete wavelet transform (DWT), \( a_0 \) and \( b_0 \) are the sampling intervals for \( a \) and \( b \), and \( j, k \in Z \), where \( Z \) is the set of all integers. Accordingly, we may express any continuous-time function \( f(t) \in L^2(R) \) as a series using a dyadic wavelet grid for example, by taking \( a_0 = 2 \) and \( b_0 = 1 \) such that:

\[
f(t) = \sum_{j,k} d_{j,k} 2^{-j/2} \psi(2^{-j} t - k)
\]

(3.3)

where \( d_{j,k} \) is the wavelet coefficient or the discrete wavelet transform (DWT) of \( f(t) \) defined as:

\[
d_{j,k} = 2^{-j/2} \int f(t) \psi(2^{-j} t - k) \, dt
\]

(3.4)

The above parameterisation of the time or space location by \( k \) and the frequency or scale by \( j \) has been proven to be very effective in the analysis of non-stationary signals [28].

The multi-resolution analysis (MRA) property of the DWT provides an approach for the construction of orthonormal wavelets of compact form and, hence, reduces the computational time required to obtain the dyadic expansion of signals. This analysis is represented by a nested sequence of closed subspaces \( \{V_m \mid m \in Z \} \) of \( L^2(R) \) such that

\[
\ldots 
\begin{array}{c}
\ldots V_2 
V_1 
V_0 
V_{-1} 
V_{-2} 
\end{array}
\]

(3.5)

\begin{tabular}{cc}
\text{Coarser} & \text{Finer}
\end{tabular}

or \( V_{j+1} \subset V_j \) for all \( j \in Z \). Assuming a finite number of decomposition stages, a scaling function \( \phi(t) \) can be introduced, and the reconstructed function \( f(t) \) can thus be represented by:
\[ f(t) = \sum_k c_{L,k} \phi_{L,k}(t) + \sum_k \sum_{j=1}^L d_{j,k} \psi_{j,k}(t) \]  

(3.6)

where \( c_{L,k} \) are the scaling coefficients and \( L \) is the number of resolution stages [27,28]. For example a three stages octave-band decomposition can be represented as shown in Fig 3.1, where \( H_0 \) and \( H_1 \) are the low and high pass filters respectively, each followed by a down sampling process. The outputs of the high pass filters represent the scaling coefficients.

![Figure 3.1: 3-Stage octave-band wavelet decomposition](image)

**3.2 Second Generation Wavelet Transform (SGWT)**

The basis functions of classical wavelet basis functions are generated by translation and dilation of a single mother wavelet function. This restriction makes it possible to employ Fourier transform techniques in constructing wavelets. The construction, however, requires a regular mesh and unbounded or periodic domain. The classical WT, therefore, works well for infinite or periodic signals, but special adaptations of the basis functions near the boundaries are required in order to handle non-periodic boundary conditions, which are often encountered in natural speech. In 1996, Daubechies [43], introduced the second generation wavelet transforms (SGWT) which are based on a more general class of adaptive wavelet basis functions that are
not necessarily translates and dilates of each other and that can exist on bounded and irregular meshes. Besides being adaptive, the SGWTs enjoy all the powerful properties such as time-frequency localisation, multi-resolution and fast implementation algorithms. Daubechies also introduced an entirely spatial construction technique for SGWT known as the lifting scheme [43]. The lifting scheme has been shown to speed up the implementation of the SGWT by a factor of two compared to classical WT [29]. It is for these advantageous properties that we propose the use of the SGWT in this work.

The basic idea behind the second-generation wavelet is to first split a signal, \( x(n) \), into an even set, \( \{ x(n) : n \text{ even} \} \), and an odd set, \( \{ x(n) : n \text{ odd} \} \), by predicting the odd signal from the even part. What is missed by the prediction is called the detail. The even samples are then adjusted to serve the coarse version of the original signal. The adjustment is needed to maintain the same average for the fine and coarse versions of the same signal. The above process can be summarized as follows:

a) Split data: even and odd.

b) Predict odd using even: detail = odd – P (even).

c) Update even using detail: Coarse=even + U (detail).

The process of computing a prediction and recording the detail is called a lifting step as shown in Fig 3.2. In general, the lifting scheme speeds up the implementation as compared to the case of classical WT. All operations within one lifting step can be done in parallel while the only the sequential part is the order of the lifting operations, resulting in an adaptive wavelet transform. Lifting provides a framework that allows the construction of certain biorthogonal wavelets and can be generalized for the setting of the SGWT. First generation families can also be built with lifting framework. Wavelet filters can be decomposed into lifting step, which leads to expressing the WT in a polyphase form. The lifting can be made using matrices, referred to as the modulation matrices, with Laurent polynomial elements [43]. As illustrated in Fig 3.3, the CWT is usually performed using FIR filters. In this figure, \( \tilde{h}, \tilde{g} \) are the analyzing filters and \( h, g \) are the synthesizing filters. For the filter bank in Fig 3.3, the conditions for perfect reconstruction are given by:
\[ h(z)\tilde{h}(z^{-1}) + g(z)\tilde{g}(z^{-1}) = 2 \]
\[ h(z)\tilde{h}(-z^{-1}) + g(z)\tilde{g}(-z^{-1}) = 0 \]

\[ (3.7) \]

Figure 3.2: Representation of the SGWT: (a) Forward transform, and (b) Inverse transform.

Figure 3.3: Two-channel-filter banks with analysis filters \( \tilde{g}, \tilde{h} \) and synthesis filters \( \tilde{g}, \tilde{h} \).
The modulation matrix $M(z)$ can be defined as:

$$M(z) = \begin{bmatrix} h(z) & h(-z) \\ g(z) & g(-z) \end{bmatrix}$$

(3.8)

In a similar way, the dual modulation matrix $\tilde{M}(z)$ for the analyzing side can be defined. The perfect reconstruction condition can now be written as:

$$\tilde{M}(z^{-1}) M(z) = 2I$$

(3.9)

where $I$ is the identity matrix. The polyphase representation [43] is a particularly convenient tool to express the special structure of the modulation matrix. The polyphase representation of a filter $h$ is given by:

$$h(z) = h_e(z^2) + z^{-1} h_o(z^2)$$

(3.10)

where $h_e$ contains the even coefficients, and $h_o$ contains the odd coefficients:

$$h_e(z) = \sum_k h_{2k} z^{-k} \quad \text{and} \quad h_o(z) = \sum_k h_{2k+1} z^{-k}$$

(3.11)

The polyphase matrix for the reconstruction side filters, can be assembled as:

$$\tilde{p}(z) = \begin{bmatrix} \tilde{h}_e(z) & \tilde{h}_o(z) \\ \tilde{g}_e(z) & \tilde{h}_o(z) \end{bmatrix}$$

(3.12)

where $p(z)$ can be defined in the similar way as that for the decomposition side filters. Based on the above, the wavelet transform can be represented schematically as in Fig 3.4. As can be seen, the condition for perfect reconstruction is now given by:
From (3.13) and Cramer’s rule [43] it follows that:

\[
\begin{align*}
\tilde{h}_c(z) &= g_o(z^{-1}) \\
\tilde{h}_s(z) &= -g_e(z^{-1}) \\
\tilde{g}_e(z) &= -h_o(z^{-1}) \\
\tilde{g}_o(z) &= h_e(z^{-1})
\end{align*}
\]  

(3.14)

Figure 3.4: Polyphase representation of Wavelet transform.

The lifting theorem indicates that any other finite filter \( g^\text{new}(z) \) complementary to \( h \) is of the form:

\[
g^\text{new}(z) = g(z) + h(z)s(z^2)
\]  

(3.15)

where \( s(z^2) \) is a Laurent polynomial [43]. Conversely any filter of this form is complementary to \( h \). If \( g^\text{new}(z) \) is written in polyphase form then the new polyphase matrix reads out as follows:

\[
p^\text{new}(z) = p(z) \begin{bmatrix} 1 & s(z) \\ 0 & 1 \end{bmatrix}
\]  

(3.16)

Similarly, we can use the lifting theorem to create the filter \( \tilde{h}^\text{new}(z) \) complementary to \( \tilde{g}(z) \):

\[
\tilde{h}^\text{new}(z) = \tilde{h}(z) - \tilde{g}(z)s(z^2)
\]  

(3.17)
The dual polyphase matrix is given by

$$\tilde{p}^{rew}(z) = \tilde{p}(z)\begin{bmatrix} 1 & 0 \\ -s(z^{-1}) & 1 \end{bmatrix}$$  \hspace{1cm} (3.18)$$

Any two-band FIR filter bank can be factored in a set of lifting steps using Euclidean algorithm [43]. Polyphase matrix is factored in a cascade of triangular sub-matrices where each sub-matrix corresponds to a lifting or dual lifting step.

As indicated in Fig 3.5, a polyphase matrix $\tilde{p}(z)$ of a filter bank, can be factored in triangular sub-matrices:

$$\tilde{p}(z) = \prod_{i=1}^{m} \begin{bmatrix} 1 & 0 \\ -s_i(z^{-1}) & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/k & 0 \\ 0 & k \end{bmatrix}$$  \hspace{1cm} (3.19)$$

$$p(z) = \prod_{i=1}^{m} \begin{bmatrix} 1 & s_i(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t_i(z) & 1 \end{bmatrix} \begin{bmatrix} k & 0 \\ 0 & 1/k \end{bmatrix}$$  \hspace{1cm} (3.20)$$

Figure 3.5: left side is the forward wavelet transform using lifting, right side is the inverse wavelet transform using lifting.

The derivation of wavelet filters such as Haar, Daubechies(9-7), and Cubic B-splines using lifting steps are shown in Appendix-I.
3.2.1 Traditional advantages of lifting implementations

The following points present some basic comments on the SGWT decomposition and its usefulness.

a. Lifting leads to speed-up when compared to the standard implementation.

b. Lifting allows for an in-place implementation of the fast wavelet transform, a feature similar to the Fast Fourier Transform FFT. This means the wavelet transform can be calculated without allocating auxiliary memory.

c. All operations within one lifting step can be done entirely in parallel while the only sequential part is the order of the lifting operations.

d. Using Lifting can be considered an easy way to build non-linear wavelet transforms.

e. Lifting allows for adaptive wavelet transforms. This means one can start the analysis of a function from coarsest levels and then build the finer levels by refining only in the areas of interest.

3.3 Wavelet Packet Decomposition (WPD)

Wavelet packets are basically particular linear combinations of wavelets. The coefficients in the linear combinations are computed by a factored algorithm, whose details are given in [44]. With the result, the use of this algorithm results in the wavelet packet expansion having a low computational complexity. The construction of an orthonormal basis $L^2(\mathbb{R})$ for any wavelet packet tree can be done as follows [45]:

Let $E_m$ be the tree-set of indices $\{l, n\}$ that corresponds to the terminal nodes of some binary tree, where the index $l$ indicates tree level and $n$ the node sequence in each level. If a set of disjoint intervals defined by:

$$\cap\bigcup_{m=0}^{2^n} E_m$$

forms a complete cover of $[0,1)$, then the set

$$\psi_{j,n,k} = 2^{j/2} \psi_n(2^j t - k) : (l,n) \in E_m , k \in \mathbb{Z}$$

(3.22)
represents a wavelet packet basis. A particular wavelet packet case of WPD (or so-called Subband Decomposition) of a signal \( x = \{x(i)\}_{i=0}^7 \) is given in the Fig 3.6. Here the tree-set \( E_m \) is the subset of the set of all terminal nodes in the binary tree \( E_m \subset \{(-1,0), (-1,1), (-2,0), (-2,1), (-2,2), (-2,3)\} \) and has to satisfy the condition defined by the following (3.21).

\[
\begin{align*}
\mathcal{H}_0 & \quad H_0 x_0, H_0 x_1, H_0 x_2, H_0 x_3, \\
\mathcal{H}_1 & \quad H_1 x_0, H_1 x_1, H_1 x_2, H_1 x_3, \\
\mathcal{H}_2 & \quad H_2 x_0, H_2 x_1, H_2 x_2, H_2 x_3
\end{align*}
\]

**Figure 3.6:** Discrete wavelet packet decomposition coefficients with 2 decomposition

\[
\begin{align*}
&x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \\
&\mathcal{H}_0 \quad H_0 x_0, H_0 x_1, H_0 x_2, H_0 x_3, \\
&\mathcal{H}_1 \quad H_1 x_0, H_1 x_1, H_1 x_2, H_1 x_3, \\
&\mathcal{H}_2 \quad H_2 x_0, H_2 x_1
\end{align*}
\]

**Figure 3.7:** Discrete wavelet transform coefficients with 2-level decomposition

\[
\begin{align*}
&x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \\
&\mathcal{H}_0 \quad H_0 x_0, H_0 x_1, H_0 x_2, H_0 x_3, \\
&\mathcal{H}_1 \quad H_1 x_0, H_1 x_1, H_1 x_2, H_1 x_3, \\
&\mathcal{H}_2 \quad H_2 x_0
\end{align*}
\]
The discrete wavelet transform (DWT) is a particular case of WPD, where the coefficients obtained by high-pass filtering are not transformed further as shown in Fig 3.7.

Wavelet packet analysis is perfectly recursive, such that each newly computed wavelet packet coefficient sequence becomes the root of its own analysis tree. The expansion coefficients associated with a prescribed signal \( x \) can be computed efficiently; using the low and high pass filters \( H_1 \) and \( H_0 \) (Fig 3.6), according to the well-known decomposition into subband signals [44].

### 3.4 Perceptual / Bark Scaled Wavelet Transform

The Bark-scaled wavelet transform BS-WPD can be considered as another application of the WPD, where the speech signal is decomposed into non-uniform bands. Such decomposition is used to exploit the auditory perceptual system where the resultant bands are approximations to the critical bands.

A terminal node \((l,n) \in E_m\) in BS-WPD is associated with each sub-band (subspace) whose centre frequency and bandwidth are roughly given by [46]:

\[
\begin{align*}
\tilde{f}_{l,n} &= 2^{-l}[GC^{-1}(n) + 0.5].F_s / 2 \\
B_{l,n} &= 2^{-l}.F_s / 2
\end{align*}
\]

(3.23)

where \(GC^{-1}\) is the inverse Gray code permutation and \(F_s\) is the sampling frequency in the signal space. To obtain the critical bands with the WPD, a decomposition tree has been constructed such that the distance between the centre frequencies of one subband to the frequency of the next subband is 1 Bark. The relation between frequency \(f\) in Hz and critical band rate \(c\) in bark is given by [44]:

\[c = 26.81/(1+1960/f) – 0.53\]

(3.24)

Fig 3.8 shows an approximation of the `constructed BS-WPD. The corresponding decomposition tree used in this construction is depicted in Fig 3.9. As can be seen
from this tree, the BS-WPD splits the frequency range 0-8 kHz into 21 sub-bands. The approximated and true critical bands are given in Table 3.1. In Chapter 4, a similar perceptual WPD will be used where only 18 approximated critical bands are considered.

**Table 3.1: BS-WPD approximated and true critical bands**

<table>
<thead>
<tr>
<th>BANDS NO.</th>
<th>BS-WPD CRITICAL BANDS (KHz)</th>
<th>TRUE CRITICAL BANDS (KHz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.125</td>
<td>0.100</td>
</tr>
<tr>
<td>2</td>
<td>0.187</td>
<td>0.150</td>
</tr>
<tr>
<td>3</td>
<td>0.312</td>
<td>0.250</td>
</tr>
<tr>
<td>4</td>
<td>0.437</td>
<td>0.350</td>
</tr>
<tr>
<td>5</td>
<td>0.562</td>
<td>0.450</td>
</tr>
<tr>
<td>6</td>
<td>0.687</td>
<td>0.570</td>
</tr>
<tr>
<td>7</td>
<td>0.812</td>
<td>0.700</td>
</tr>
<tr>
<td>8</td>
<td>0.937</td>
<td>0.840</td>
</tr>
<tr>
<td>9</td>
<td>1.125</td>
<td>1.000</td>
</tr>
<tr>
<td>10</td>
<td>1.375</td>
<td>1.170</td>
</tr>
<tr>
<td>11</td>
<td>1.562</td>
<td>1.370</td>
</tr>
<tr>
<td>12</td>
<td>1.687</td>
<td>1.600</td>
</tr>
<tr>
<td>13</td>
<td>1.875</td>
<td>1.850</td>
</tr>
<tr>
<td>14</td>
<td>2.250</td>
<td>2.150</td>
</tr>
<tr>
<td>15</td>
<td>2.750</td>
<td>2.500</td>
</tr>
<tr>
<td>16</td>
<td>3.250</td>
<td>3.000</td>
</tr>
<tr>
<td>17</td>
<td>3.750</td>
<td>3.400</td>
</tr>
<tr>
<td>18</td>
<td>4.500</td>
<td>4.000</td>
</tr>
<tr>
<td>19</td>
<td>5.500</td>
<td>4.800</td>
</tr>
<tr>
<td>20</td>
<td>6.500</td>
<td>5.800</td>
</tr>
<tr>
<td>21</td>
<td>7.500</td>
<td>7.000</td>
</tr>
</tbody>
</table>
Figure 3.8: Approximation of the Bark scale (solid line) by BS-WPD (*).

Figure 3.9: BS-WPD expansion tree
3.5 Wavelet-Based Denoising Using Thresholding

Wavelet Transform (WT) has been progressively applied to removing the additive white Gaussian noise (WGN) from speech signals. The most commonly used WT-based speech denoising techniques are those based on wavelet thresholding or shrinkage (see Chapter 2). In general, most speech denoising techniques that use wavelet thresholding perform well in enhancing corrupted speech. However, in most cases, the resulting speech suffers from serious residual noise and distortion. WT thresholding, or shrinkage, techniques are based on the ability of the wavelet transforms to compress data, which is represented by the WT coefficients. The noisy signal is transformed into few, relatively large coefficients, with the white noise spread evenly over all coefficients. The shrinkage process is a key part of any wavelet-based denoising algorithm. It is based on the following algorithm, which is known as the Donoho-Johnstone algorithm [47].

Suppose we are given noisy data $y = \{y(n)\}_{n=0}^{N-1}$ with $N=2^j$, where

$$y(n) = f(n) + e(n), \quad n=0,\ldots,N-1$$  \hspace{1cm} (3.25)

$f = \{f(n)\}_{n=0}^{N-1}$ is an unknown discrete real-valued signal which we would like to recover, and $e = \{e(n)\}_{n=0}^{N-1}$ is WGN with zero mean and presumably known power spectral density (PSD) $\sigma^2$. Let $\hat{f} = \{\hat{f}(n)\}_{n=0}^{N-1}$ denote the vector of estimated sample values of $f$. A basic denoising technique can be introduced now using the example of a discrete wavelet transform DWT.

The vector $w = \{w_{l,n,k}\}$ of wavelet expansions coefficients of the noisy data $y$ is defined by:

$$w = W_y$$  \hspace{1cm} (3.26)
Where \( l \) is the resolution level index (the scaling parameter), \( n \) is the oscillation (modulation) index, \( k \) is the time-domain position index and \( W \) is the finite orthonormal wavelet transform matrix. The orthonormality of \( W \) yields the following reconstruction formula:

\[
\mathbf{y} = W^T \mathbf{w}
\]  

(3.27)

Under the noise model underlying (3.25), noise contaminates all wavelet coefficients equally. The noise vector \( \mathbf{e} \) is assumed to represent WGN so that its orthogonal transform \( \mathbf{z} = W \mathbf{e} = \{ z_{l,n,k} \} \) is also WGN. Consequently, applying the wavelet transform to \( \mathbf{y} \) yields:

\[
w_{l,n,k} = \Theta_{l,n,k} + z_{l,n,k}
\]  

(3.28)

where \( \Theta = W \mathbf{f} = \{ \theta_{l,n,k} \} \) is the vector representing to the unknown wavelet transform coefficients of the noiseless data \( \mathbf{f} \). Therefore every empirical wavelet coefficient \( w_{l,n,k} \) contributes noise of variance \( \sigma^2 \), but only very few wavelet coefficients contain significant signal energy. This is a heuristic basis to wavelet –based denoising.

Denoising in the wavelet domain is based on the principle of selective wavelet reconstruction. Given \( \mathbf{w} \) we determine a final set \( \Gamma \) of indexes \( (l,n,k) \) of wavelet coefficients, that have to be modified (multiplied by an appropriate gain), so the estimate \( \hat{\Theta} \) of \( \Theta \) implies:

\[
\hat{\mathbf{f}} = W^T \hat{\Theta} = W^T \{ TH \{ W \mathbf{y} \} \}
\]  

(3.29)

where \( TH(\) \) is generally a nonlinear operator that determines the set \( \Gamma \) and performs noise subtraction in the wavelet domain. Clearly, the quality of the resulting estimation \( \hat{\mathbf{f}} \) depends on the algorithm, that determines \( \Gamma \) and the gain function.
3.5.1 Thresholding types

Wavelet thresholding can be performed as hard, soft, semisoft and nonlinear adaptive thresholding which are defined as follows [48]:

a. **Hard Thresholding** [49]:

\[
\text{TH}_h(x, \lambda) = x \cdot 1_{|x| > \lambda}
\]  \hspace{1cm} (3.30)

where \(x\) represents unknown input signal to be thresholded and \(\lambda\) the predetermined threshold value.

\[
1_u = \begin{cases} 
1, & u > 0 \\
0, & u \leq 0 
\end{cases}
\]  \hspace{1cm} (3.31)

b. **Soft Thresholding** [20]:

\[
\text{TH}_s(x, \lambda) = (|x| - \lambda) \cdot \text{sign}(x) \cdot 1_{|x| > \lambda}
\]  \hspace{1cm} (3.32)

c. **Semisoft Thresholding** [50]:

\[
\text{TH}_{\text{semi}}(x, \lambda_1, \lambda_2) = \begin{cases} 
x & \text{if } |x| > \lambda_1 \\
\text{sgn}(x) \frac{\lambda_2}{\lambda_2 - \lambda_1} (|x| - \lambda_1) & \text{if } \lambda_1 < |x| < \lambda_2 \\
0 & \text{otherwise}
\end{cases}
\]  \hspace{1cm} (3.33)

where \(\lambda_1\) and \(\lambda_2\) represent a lower and upper predetermined threshold values.
d. Nonlinear Aadaptive Thresholding [51]:

\[
T_{th_{\text{nonlinear}}}(x, \lambda) = \begin{cases} 
    x + \lambda - \frac{1}{2k + 1} & \text{if } x < -\lambda \\
    \frac{1}{(2k + 1)^2 \lambda} & \text{if } |x| \leq \lambda \\
    x - \lambda + \frac{1}{2k + 1} & \text{if } x > \lambda 
\end{cases}
\]  

(3.34)

Other thresholding techniques also exist. For example in [52], a super-soft threshold is developed. Another possibility is to have different thresholds for different levels of the transformed coefficients. These methods are called Hybrid schemes.

To determine the value of the threshold, Donho and Johnstone[20], proposed a global threshold defined by:

\[
\lambda \approx \sigma \sqrt{2 \log N}
\]

(3.35)

where \(\sigma\) is the noise level (variance) and \(N\) is the length of the coefficient, expressed in terms of signal samples. The noise level in the case of WGN can be estimated as:

\[
\sigma = \text{MAD} / 0.6745
\]

(3.36)

where \(\text{MAD}\) is median absolute deviation of the noisy samples. The factor 0.6745 in the denominator rescales the numerator so that \(\sigma\) is also a suitable estimator for the standard deviation of WGN. For the wavelet packets WP case, the threshold becomes:

\[
\lambda \approx \sigma \sqrt{2 \log( N \log_2 N )}
\]

(3.37)

At first site, it might look strange that the optimal threshold depends on the number of the data samples. There is a intuitive explanation for this behaviour. Adding more samples enhances redundancy in the signal. In wavelet domain, this means that the number of important coefficients is hardly growing, and all information remains concentrated in a limited number of coefficients [53]. Since the number of important
coefficients is approximately constant, the reconstruction would become noisier. Therefore, it is better to let the threshold increase slowly to catch all noise coefficients while leaving the faster growing signal coefficients intact.

3.7 Summary

In this Chapter, the principles of the WT have been introduced and the various WT decomposition schemes have been presented. Wavelet thresholding using different techniques has also been described.

The two proceeding Chapters will deal with issues of adaptive noise profile estimation (Chapter 4) and the voiced/unvoiced classification process (Chapter 5), that are suggested in our proposed speech enhancement system as outlined in Section 2.5.
Chapter 4

Adaptive Noise Estimation

4.1 Introduction

As indicated in Chapter 2, accurate noise power estimation is crucial part of the proposed speech enhancement system. As a part of this work, an extensive investigation has been carried out to develop an efficient noise power estimation technique that is able to work in various stationary or non-stationary conditions.

In this chapter, we consider three different, recently reported algorithms for estimation of noise level/power from the noisy speech alone without the need for a voice activity detector or signal statistics learning histograms. We investigate their performance when implemented using two different wavelet decompositions: Second-Generation wavelet transform (SGWT) and perceptual wavelet packet decomposition (PWPD).

First detailed descriptions of the three noise estimation techniques are presented. Then a new noise estimation approach based on features of these techniques that have been identified as advantageous is proposed and implemented in wavelet domain. Finally, a systematic evaluation and comparison of the performance of all presented noise estimation techniques is presented and discussed.
4.2 Overview of Noise Tracking Techniques For Speech Signals

Reliable noise estimation remains a challenging task in many speech enhancement and noise compensation systems. Accurate instantaneous noise power estimation is crucial for the success and robustness of any single-channel speech enhancement system. In fact the noise estimator has a major impact on the overall quality of the speech enhancement system.

In Mobile telephony area in particular, there is a requirement for low resource, efficient noise estimation algorithms that typically must perform in real-time. Over the last few years, various noise estimation techniques have been proposed and their performance evaluated. These include techniques that are based on tracking the minima of the noise power [16,36,65], recursive estimation of subband spectra [66], or on quantile computation [67,68]. However most of the reported techniques have some limitations. For example the spectral analysis method in [36] requires a long segment of speech and a good frequency resolution to work effectively and overcome windowing errors caused by the discrete Fourier transform (DFT). The subband recursive spectral estimation proposed in [66], works well in most situations, but being based on relative energy levels, cannot distinguish between rising noise energy and the presence of speech. A new technique based on the idea of the higher order statistics (HOS) has been proposed in to deal with both Gaussian and Gaussian unlike signals [69]. The noise is estimated in each band so that it is possible to separate signal and the noise energy, but the performance may be affected significantly when the noise is Gaussian unlike. Another introduced approach is that using the minima controlled recursive averaging (MCRA) for noise estimation [70]. The noise estimate here is obtained by averaging past spectral power values, using a smoothing parameter that is adjusted by the signal presence probability in subbands. Also, most of these algorithms have been implemented using the frequency domain which although efficient, but may involve relatively high computational complexity.

The three recently reported noise estimation algorithms are investigated in the following Sections. These are: (a) an adaptive technique with a smoothing parameter that depends on the estimated subband posterior signal-to-ratio (SNR)[72]; (b) a one-pass quantile-based technique [37] and (c) a technique that is based on tracking the minimum variance of the subband noisy signal [72]. The implementation of these
algorithms is described using two wavelet decomposition schemes that provide different time-frequency resolutions. The first is based on the application of the second-generation wavelet transform (SGWT) (see Chapter 3), and the second is based on critical-band noise motivated perceptual wavelet packet decomposition (PWPD)[30]. After describing the above three techniques, a new robust wavelet-based noise estimation technique based on the algorithms (a) and (c) has been proposed. The performance of the presented noise estimation techniques is evaluated in two stages using a variety of speech signals distorted by different types of noise. In the second stage, two wavelet-denoising approaches have been used, these are: Soft and Weiner filter gain thresholding. Different objective measure parameters are used through the performance evaluation tests.

4.3 Noise Estimation Techniques

4.3.1 Adaptive smoothing-based noise estimation algorithm

This adaptive noise estimation technique is based on the use of a smoothing parameter that is controlled by the estimated subband posteriori SNR [72]. Here, the noise and speech are assumed to be independent signals and that the noise power changes slowly. The subband noisy signal power (or variance), \( \sigma_{x_i}^2 (p) = E\{y_i^2(n)\} \), is estimated on a frame-by-frame basis using:

\[
\hat{\sigma}_{x_i}^2 (p) = \frac{1}{N} \sum_{n=0}^{N-1} y_i (pN + n)
\]

(4.1)

where \( \hat{\sigma}_{x_i}^2 (p) \) is the estimated noisy signal power calculated at frame \( p \), and \( N \) is the size of the frame. Similarly, the subband noise power is estimated using the smoothing filter, such that:

\[
\hat{\sigma}_{e_i}^2 (p) = \alpha_s (p) \hat{\sigma}_{e_i}^2 (p - 1) + (1 - \alpha_s (p)) \hat{\sigma}_{x_i}^2 (p)
\]

(4.2)
where $\hat{\sigma}_i^2(p)$ is the estimate of subband noise power at frame $p$ and $\alpha(p)$ represents a smoothing filter, which is:

$$\alpha_i(p) = 1 - \min \left\{ 1, \left( \frac{\hat{\sigma}_i^2(p)}{\sigma_i^2(p-1)} \right)^{-Q} \right\} \quad (4.3)$$

where $Q$ is an integer and $\sigma_i^2(p-1)$ is the average of the noise estimates of the previous 5 to 10 frames, given by:

$$\sigma_i^2(p-1) = 1/10 \sum_{k=1}^{10} \hat{\sigma}_i^2(p-k) \quad (4.4)$$

The integer $Q$ controls the way in which $\alpha_i(p)$ changes with $\hat{\sigma}_i^2(p)/\sigma_i^2(p-1)$. Generally, larger values of $Q$ lead to larger values of $\alpha$ and slower noise updates, whereas smaller values of $Q$ give faster noise updates, at the risk of possible overestimation during long voiced intervals. The value of $Q$ is usually chosen in the range 3 to 5. The ratio $\hat{\sigma}_i^2/\sigma_i^2(p-1)$ can be considered to be an approximation to the a posteriori signal-to-noise ratio. The above algorithm can be explained as follows: if the speech is absent in frame $p$, the new calculation of the noisy speech power $\hat{\sigma}_i^2(p)$ should be very close to the average noise estimate $\sigma_i^2(p-1)$, so that $\alpha(p) \approx 0$. That is the estimate of noise power in frame $p$ is almost immediately close to the power of the noisy signal in the absence of speech. On the other hand, if the speech is present, the new signal power calculation is much larger than the previous noise estimate $\hat{\sigma}_i^2(p) \gg \sigma_i^2(p-1)$, and hence from (4.3), $\alpha_i(p) \approx 1$.

### 4.3.2 Quantile based noise spectrum estimation

For any speech signal, it is known that even in the speech sections, not all frequency bands are permanently occupied with speech [67]. In fact, a significant percentage of the energy in each frequency band is due to the noise in the signal. This observation can be used to estimate the noise power spectrum $E(k)$ from the observed noise signal.
denoted by $Y(k,t)$ by computing the $q$-th quantile over time in every frequency band.

More precisely, for every band, the frames of the entire utterance $Y(k,t)$, $t=0,\ldots,T$ are sorted such that

$$Y(k,t_0) \leq Y(k,t_1) \leq \ldots \leq Y(k,t_T).$$

The $q$-quantile noise estimation is defined as [5]:

$$E(k) = \frac{\sum_{i=0}^{qT} Y(k,t)}{qT} \quad (4.5)$$

The above assumption is true for very small values of $q$. In [67], it was shown experimentally that the probability of having more than 20% duration being silence for various segment lengths ranging from 200ms to 2000ms. Accordingly, realistic value for $q$ needs to be very close to 0.2.

In order to implement the quantile technique in wavelet domain, we need to modify the following notations as explained below: First, the decomposed coefficients, $y_{l,n}(k)$, are framed into frames of length $L_{\text{fr}}$. Let $L_{\text{win}} > L_{\text{fr}}$ be the length of a finite window observation of the coefficients $y_{l,n}(k)$ spanning a number of frames. Also, let $\hat{\sigma}_{i,l,n}$ be the noise level of the $p$-th frame in the ($l$, $n$th) band estimated using the previous set of coefficients $\{y_{i,l,n}(k), k=0,\Lambda, L_{\text{win}}-1\}$, after been sorted according to the requirement of the quantile-based approach. The noise level in the $p$th frame of the $l$, $n$th band is estimated by:

$$\hat{\sigma}_{l,n} = \beta \sum_{p=0}^{\text{int}(q,L_{\text{win}})} y_{l,n}(p) / (t_{\text{win}}) q \quad (4.6)$$

where $\beta$ is an appropriate scaling factor. This process is illustrated in Fig 4.1.
4.3.3 Minimum variance tracking-based noise estimation algorithm

In this algorithm, both the noisy signal and the noise are considered to be stationary over a short period of time, such that the variance can be estimated on a frame-by-frame basis. The noisy signal variance, $\sigma^2_{y_i}$, for each band is calculated as [72]:

$$
\sigma^2_{y_i}(p) = \alpha_i \sigma^2_{y_i}(p-1) + (1-\alpha_i)\sigma^2_{y_i,\text{new}}(p)
$$

(4.7)

where

$$
\sigma^2_{y_i,\text{new}}(p) = \frac{1}{N} \sum_{k=0}^{N-1} y_i^2(pN+k)
$$

(4.8)

is the most recent approximation of the noisy signal variance using the new data at frame $p$. The parameter $\alpha_i$ is a smoothing factor whose values are between $0.45 \leq \alpha_i \leq 0.95$. The noise estimate $\sigma^2_{e_i}(p)$ is updated such that

$$
\sigma^2_{e_i}(p) = \alpha_i \sigma^2_{e_i}(p-1) + (1-\alpha_i)\sigma^2_{e_i,\text{new}}(p)
$$

(4.9)

where $\sigma^2_{e_i,\text{new}}$ is the minimum value of $\sigma^2_{y_i}(p)$ in the neighboring frames, i.e. if

$$
\sigma^2_{y_i}(p-1) < \sigma^2_{y_i}(p) \quad \& \quad \sigma^2_{y_i}(p-1) < \sigma^2_{y_i}(p-2) K \quad \& \quad \sigma^2_{y_i}(p-1) < 2\sigma^2_{e_i}(p-1)
$$

(4.10a)

then

$$
\sigma^2_{e_i,\text{new}}(p) = \sigma^2_{e_i}(p-1)
$$

(4.10b)

Otherwise

$$
\sigma^2_{e_i,\text{new}}(p) = \sigma^2_{e_i}(p-1) .
$$
The last term in (4.10a) is introduced to avoid picking up spurious minimum during voiced frames and is based on the fact that noise intensity usually changes slowly.

4.3.4 Proposed noise estimation algorithm

A new noise estimation technique based on the modification of the quantile-based method presented in section 4.3.2 is proposed here. The modification is based on the addition of a smoothing parameter that depends on the estimated subband SNR, similar to that used in the smoothing-based technique presented in section 4.3.1, such that a new quantile-based noise estimate that can be updated adaptively is obtained. The new technique proceeds as follows. The noise power in the $i$th subband of the $p$th frame, $\hat{\sigma}_{v,i}$, is estimated as in the standard quantile-based method (see Section 4.2.2). To enhance the adaptability and trackability of the quantile-based method, a smoothing factor based on the estimated subband posteriori SNR defined by:

\[
\alpha_i(p) = 1 - \min \left\{ 1, \left( \frac{\hat{\sigma}_{v,i}^2(p)}{\sigma_{v,\text{quantile}}^2(p-1)} \right)^{-Q} \right\} \tag{4.11}
\]

where $\sigma_{v,\text{quantile}}^2$ is the noise power in the $i$th subband of the $p$th frame as calculated using (4.6), is introduced.

4.4 Wavelet Decompositions:

The performance of the noise estimation techniques presented in previous Sections has been evaluated using two different wavelet decompositions. The first decomposition scheme uses Daubechies(9-7) wavelet filter based on lifting step SGWT, whose aspects have been explained in Chapter 3, and the second is based on
perceptual wavelet transform. Appendix-I gives details of various common wavelet filters designs are shown in

### 4.4.1 Perceptual wavelet packet decomposition PWPD

For the implementation of the PWPD, a similar WPD to that used for BS-WPD scheme (see Chapter 3) is utilised to represent the 18 critical bands of speech signals sampled at 8kHz. The scheme, which was first proposed by Black and Zeytinoglu [30], is based on an efficient 6-stage tree structure decomposition using 16-tap FIR filters derived from the Daubechies wavelet function. Fig 4.2 depicts the PWPD decomposition tree, which also which provides an exact invertible decomposition. Here, the dB16 wavelet is used which represents the optimal wavelet among the class of all equal length wavelets [30].

It should be noted here that the above PWPD decomposition provides a close approximation to the critical bands as defined in the auditory psychoacoustics model, as demonstrated in Fig 4.3. This perceptual auditory model uses a set of strongly overlapping bandpass filters.

### 4.5 Performance Evaluation and Experimental Results

To evaluate the performance of the described noise estimation algorithms, various (TIMIT) speech signals sampled at 8 kHz and framed into 32 ms long frames with 50% overlap have been used. For the implementation of the quantile technique, the corresponding time lengths for $L_{\text{frm}}^{i,n}$ and $L_{\text{win}}^{i,n}$ are 64 ms and 512 ms, respectively. The scaling factor $\beta$ has been experimentally selected to be 0.4. The performance evaluation has been done in two stages [73, 74]. In the first stage, each noise
Figure 4.2: PWPD expansion tree

Figure 4.3: Exact and approximation of the critical bands using 18-subbands PWPD
estimation algorithm was implemented using the two chosen wavelet decompositions schemes, and then tested over various noisy conditions.

In the second stage of the evaluation, the four-noise estimation algorithms were incorporated into speech enhancement/denoising system and their performance was assessed accordingly. For the speech enhancement, two different denoising approaches were employed, as well be explained in the following two sections.

4.5.1 Performance of the four noise estimation techniques

In the first part of the evaluation process, different speech signals acquired from the TIMIT database were degraded by different types of noise and used to test the four noise estimation techniques. The noisy speech signals were decomposed into 6 bands (details) using the dB (7-9) second generation wavelet filter, and each of the four techniques was then used to estimate the spectrum of the added noise power. Experimental results for the performance of each noise estimation algorithm, implemented using SGWT scheme, are shown in Figures 4.4 and 4.5 for the cases of additive white Gaussian noise (AWGN) and pink noise, respectively. Here the solid line waveforms represent the real noise for band 2 of the decomposed signal, which corresponds to 0.5-1 kHz. In Figures 4.4-4.9 (a) corresponds to the adaptive smoothing-based, (b) the quantile-based, (c) the minimum variance tracking-based, and (d) the proposed technique. Fig 4.6 shows, similar results to the above, regarding the performance evaluation of the noise estimation techniques in band 3 for the case of F16 cockpit noise. Using the perceptual wavelet decomposition, Figures 4.7 and 4.8 show the real and estimated noise waveforms for band 7 of the decomposition, for the cases of AWGN and Pink noise, respectively. Fig 4.9 shows similar results to those of Figures 5.7 and 5.8, regarding the performance of the noise estimation techniques in band 10 for the case of F16 cockpit noise. It can be noticed from Figures 4.7-4.9, that there is a near perfect matching between the real and estimated noise when the proposed noise estimation of Section 4.3.4 is used. In Fig 4.10, the updating with time of the values of the parameter $\alpha(p)$, associated with the adaptive smoothing based algorithm which is defined in (4.3) is shown for the case of band 3 of SGWT decomposition of the noisy speech.
Figure 4.4: Real and estimated noise using SGWT-based noise estimation with White noise at 0dB.

Figure 4.5: Real and estimated noise power using SGWT-based noise estimation with Pink noise at 0dB.
Figure 4.6: Real and estimated noise power using SGWT-based noise estimation with F16 cockpit noise at 0dB.

Figure 4.7: Real and estimated noise power using PWPD-based noise estimation with White noise at 0dB.
Figure 4.8: Real and estimated noise power using PWPD-based noise estimation with Pink noise at 0dB.

Figure 4.9: Real and estimated noise power using PWPD-based noise estimation with F16 cockpit noise at 0dB.
Figure 4.10: (a) A subband noisy signal at 10 dB with, (b) corresponding updating of values of $\alpha(p)$
To provide an objective performance measure, the average relative error (ARE) factor in the estimated noise was also computed by the following formula:

\[
ARE = \frac{1}{N_{\text{frame}}} \sum_{p} \frac{\hat{\sigma}^2_{e_i}(p) - \sigma^2_{e_i}(p)}{\sigma^2_{e_i}}
\]

(4.12)

where \(N_{\text{frame}}\) represents the number of frames in the test signal. Using this factor, Tables 4.1-4.4 illustrate the performance of the four presented noise estimation techniques for two SGWT subbands (bands 2 and 3) over different noise types (white, pink and F16 cockpit).

Here, T1, T2, T3 and T4 refer to the first, second, third and the proposed noise estimation techniques in the sequence of their presentation in Section 4.3. In Tables 4.5 and 4.6, the performance of the noise estimation techniques are illustrated, in a similar fashion to that used used in Tables 4.1-4.4 for two PWPD subbands (bands 7 and 15) for noise types (White and F16 cockpit).

**Table 4.1: Average relative error ARE in band-2 SGWT for the four noise estimation methods**

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.51</td>
<td>2.71</td>
<td>0.24</td>
<td>0.39</td>
</tr>
<tr>
<td>5</td>
<td>0.26</td>
<td>1.12</td>
<td>0.29</td>
<td>0.22</td>
</tr>
<tr>
<td>0</td>
<td>0.28</td>
<td>1.07</td>
<td>0.59</td>
<td>0.20</td>
</tr>
<tr>
<td>-5</td>
<td>0.048</td>
<td>0.40</td>
<td>0.13</td>
<td>0.053</td>
</tr>
<tr>
<td>-10</td>
<td>0.037</td>
<td>0.67</td>
<td>0.15</td>
<td>0.015</td>
</tr>
</tbody>
</table>

**Table 4.2: Average relative error ARE in band-2 SGWT for the four noise estimation methods**

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.45</td>
<td>5.92</td>
<td>0.45</td>
<td>1.22</td>
</tr>
<tr>
<td>5</td>
<td>0.60</td>
<td>4.68</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>0</td>
<td>0.17</td>
<td>0.70</td>
<td>0.14</td>
<td>0.21</td>
</tr>
<tr>
<td>-5</td>
<td>0.078</td>
<td>0.16</td>
<td>0.13</td>
<td>0.076</td>
</tr>
<tr>
<td>-10</td>
<td>0.050</td>
<td>0.50</td>
<td>0.14</td>
<td>0.082</td>
</tr>
</tbody>
</table>
Table 4.3: Average relative error ARE in band-3 SGWT for the four noise estimation methods

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>ARE – WHITE NOISE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
</tr>
<tr>
<td>10</td>
<td>0.84</td>
</tr>
<tr>
<td>5</td>
<td>0.31</td>
</tr>
<tr>
<td>0</td>
<td>0.11</td>
</tr>
<tr>
<td>-5</td>
<td>0.061</td>
</tr>
<tr>
<td>-10</td>
<td>0.051</td>
</tr>
</tbody>
</table>

Table 4.4: Average relative error ARE in band-3 SGWT for the four noise estimation methods

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>ARE – F16 COCKPIT NOISE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
</tr>
<tr>
<td>10</td>
<td>4.00</td>
</tr>
<tr>
<td>5</td>
<td>2.54</td>
</tr>
<tr>
<td>0</td>
<td>0.49</td>
</tr>
<tr>
<td>-5</td>
<td>0.48</td>
</tr>
<tr>
<td>-10</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table 4.5: Average relative error ARE in band-7 PWPD for the four noise estimation methods.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>ARE – WHITE NOISE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
</tr>
<tr>
<td>10</td>
<td>2.56</td>
</tr>
<tr>
<td>5</td>
<td>0.9</td>
</tr>
<tr>
<td>0</td>
<td>0.43</td>
</tr>
<tr>
<td>-5</td>
<td>0.25</td>
</tr>
<tr>
<td>-10</td>
<td>0.073</td>
</tr>
</tbody>
</table>

Table 4.6: Average relative error ARE in band-15 PWPD for the four noise estimation methods.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>ARE – F16 COCKPIT NOISE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
</tr>
<tr>
<td>10</td>
<td>0.96</td>
</tr>
<tr>
<td>5</td>
<td>0.36</td>
</tr>
<tr>
<td>0</td>
<td>0.16</td>
</tr>
<tr>
<td>-5</td>
<td>0.10</td>
</tr>
<tr>
<td>-10</td>
<td>0.18</td>
</tr>
</tbody>
</table>
4.5.2 Performance evaluation using speech enhancement

In this part of the evaluation process, the four noise estimation methods have been used to enhance speech signals contaminated with noise over different conditions. Two types of speech denoising approaches have been used to test the performance of the noise estimation techniques. First approach uses a soft thresholding as defined in (3.32). Here, the value of the threshold $TH$ when using SGWT is computed by [47]:

$$TH = \sigma \sqrt{2 \log(N)}$$

(4.13)

For the PWPD, the threshold value is given by [26]:

$$TH = \sigma \sqrt{2 \log(N \log_2 N)}$$

(4.14)

In the second denoising type, Weiner filter gain-based soft thresholding has been used. Here, the estimate for the clean signal variance is obtained by:

$$\sigma^2_e (p) = \max \{0, \sigma^2_y (p) - \sigma^2_v (p)\}$$

(4.15)

The zero term is introduced to avoid negative variance estimation. The denoising gain $K_i$ is calculated as [75]:

$$G_i = \frac{\sigma^2_y (p)}{\sigma^2_f (p) + \mu \sigma^2_e (p)}$$

(4.16)

where $\mu$ is an arbitrary constant that allows a trade-off between signal distortion and noise. For the quantative evaluation, the cepstral distance between the noisy and enhanced speech signals, with respect to the clean and the input average segmental signal-to-ratio ($ASSNR$), has been measured. Tables 4.7-4.9 show the cepstral distances for the enhanced signals using SGWT and soft thresholding, as obtained from the application of each noise estimation algorithm over different noisy
conditions. Similar results for the cepstral distances measure-using PWPD and Weiner gain thresholding are given in Tables 4.10-4.12.

The ASSNR using SGWT and PWPD soft thresholding for the different noise estimation techniques are shown in Figures 4.11 and 4.12. The spectrograms of the noisy and enhanced speech signals using PWPD soft thresholding for noise estimation algorithms T1 are shown in Figures 4.13 and 4.14. In Fig 4.15, the spectrogram of the noisy and enhanced speech signals using the proposed algorithm is shown.

*Table 4.7: Comparison of cepstral distance for the noise estimation algorithms using SGWT and Soft thresholding (AWGN).*

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>CEP(noisy)</td>
<td>CEP(enhanced)</td>
<td>CEP(enhanced)</td>
<td>CEP(enhanced)</td>
</tr>
<tr>
<td>10</td>
<td>4.26</td>
<td>3.70</td>
<td>5.54</td>
<td>4.09</td>
</tr>
<tr>
<td>5</td>
<td>5.14</td>
<td>4.04</td>
<td>6.57</td>
<td>4.49</td>
</tr>
<tr>
<td>0</td>
<td>6.33</td>
<td>4.66</td>
<td>8.04</td>
<td>5.08</td>
</tr>
<tr>
<td>-5</td>
<td>7.76</td>
<td>5.82</td>
<td>10.04</td>
<td>6.14</td>
</tr>
<tr>
<td>-10</td>
<td>9.78</td>
<td>5.87</td>
<td>12.65</td>
<td>7.66</td>
</tr>
</tbody>
</table>

*Table 4.8: Comparison of cepstral distance for the noise estimation algorithms using SGWT and Soft thresholding (Pink noise).*

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>CEP(noisy)</td>
<td>CEP(enhanced)</td>
<td>CEP(enhanced)</td>
<td>CEP(enhanced)</td>
</tr>
<tr>
<td>10</td>
<td>1.91</td>
<td>1.63</td>
<td>2.88</td>
<td>1.86</td>
</tr>
<tr>
<td>5</td>
<td>2.09</td>
<td>1.59</td>
<td>3.64</td>
<td>2.15</td>
</tr>
<tr>
<td>0</td>
<td>3.35</td>
<td>2.21</td>
<td>4.67</td>
<td>2.55</td>
</tr>
<tr>
<td>-5</td>
<td>4.40</td>
<td>2.90</td>
<td>6.10</td>
<td>3.16</td>
</tr>
<tr>
<td>-10</td>
<td>5.88</td>
<td>4.41</td>
<td>8.08</td>
<td>4.28</td>
</tr>
</tbody>
</table>
Table 4.9: Comparison of cepstral distance for the noise estimation algorithms using SGWT and Weiner gain thresholding (F16 cockpit noise).

<table>
<thead>
<tr>
<th>SNR</th>
<th>CEP(noisy)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4.21</td>
<td>3.81</td>
<td>3.89</td>
<td>3.89</td>
<td>1.15</td>
</tr>
<tr>
<td>5</td>
<td>5.18</td>
<td>4.16</td>
<td>4.57</td>
<td>4.57</td>
<td>2.59</td>
</tr>
<tr>
<td>0</td>
<td>6.30</td>
<td>4.95</td>
<td>5.28</td>
<td>5.28</td>
<td>3.87</td>
</tr>
<tr>
<td>-5</td>
<td>7.82</td>
<td>5.75</td>
<td>6.06</td>
<td>6.06</td>
<td>5.61</td>
</tr>
<tr>
<td>-10</td>
<td>9.79</td>
<td>6.76</td>
<td>7.57</td>
<td>7.57</td>
<td>7.39</td>
</tr>
</tbody>
</table>

Table 4.10: Comparison of cepstral distance for the noise estimation algorithms using PWPD and Soft thresholding (AWGN).

<table>
<thead>
<tr>
<th>SNR</th>
<th>CEP(noisy)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4.23</td>
<td>2.75</td>
<td>4.26</td>
<td>3.01</td>
<td>1.05</td>
</tr>
<tr>
<td>5</td>
<td>5.16</td>
<td>2.86</td>
<td>6.50</td>
<td>3.58</td>
<td>1.25</td>
</tr>
<tr>
<td>0</td>
<td>6.31</td>
<td>3.85</td>
<td>7.94</td>
<td>4.86</td>
<td>3.69</td>
</tr>
<tr>
<td>-5</td>
<td>7.78</td>
<td>4.58</td>
<td>9.91</td>
<td>5.15</td>
<td>5.43</td>
</tr>
<tr>
<td>-10</td>
<td>9.8</td>
<td>5.2</td>
<td>12.48</td>
<td>6.81</td>
<td>6.94</td>
</tr>
</tbody>
</table>

Table 4.11: Comparison of cepstral distance for the noise estimation algorithms using PWPD and soft thresholding (Pink noise).

<table>
<thead>
<tr>
<th>SNR</th>
<th>CEP(noisy)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.91</td>
<td>1.05</td>
<td>2.84</td>
<td>1.39</td>
<td>0.98</td>
</tr>
<tr>
<td>5</td>
<td>2.09</td>
<td>1.27</td>
<td>3.58</td>
<td>1.74</td>
<td>1.31</td>
</tr>
<tr>
<td>0</td>
<td>3.35</td>
<td>1.72</td>
<td>4.61</td>
<td>2.25</td>
<td>2.13</td>
</tr>
<tr>
<td>-5</td>
<td>4.40</td>
<td>2.65</td>
<td>6.02</td>
<td>3.00</td>
<td>3.19</td>
</tr>
<tr>
<td>-10</td>
<td>5.88</td>
<td>3.82</td>
<td>7.97</td>
<td>3.94</td>
<td>4.81</td>
</tr>
</tbody>
</table>

Table 4.12: Comparison of cepstral distance for the noise estimation algorithms using PWPD and Weiner gain thresholding (F16 cockpit noise).

<table>
<thead>
<tr>
<th>SNR</th>
<th>CEP(noisy)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
<th>CEP(enhanced)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.82</td>
<td>1.06</td>
<td>1.64</td>
<td>1.73</td>
<td>0.99</td>
</tr>
<tr>
<td>5</td>
<td>2.41</td>
<td>1.98</td>
<td>2.35</td>
<td>2.14</td>
<td>1.64</td>
</tr>
<tr>
<td>0</td>
<td>3.18</td>
<td>2.43</td>
<td>3.25</td>
<td>2.62</td>
<td>2.42</td>
</tr>
<tr>
<td>-5</td>
<td>4.19</td>
<td>2.96</td>
<td>4.42</td>
<td>3.20</td>
<td>3.39</td>
</tr>
<tr>
<td>-10</td>
<td>5.53</td>
<td>3.64</td>
<td>6.08</td>
<td>3.94</td>
<td>4.57</td>
</tr>
</tbody>
</table>
Figure 4.11: Speech enhancement performance evaluation (PWPD) with AWGN and soft thresholding.

Figure 4.12: Speech enhancement performance evaluation (PWPD) with AWGN and Weiner soft thresholding.
Figure 4.13: Spectrograms of: (a) Noisy sentence (AWGN 5dB); (b) The denoised sentence by soft thresholding using adaptive algorithm and PWPD.

Figure 4.14: Spectrograms of: (a) Noisy sentence (pink noise 5dB); (b) The denoised sentence by soft thresholding using adaptive algorithm and PWPD.
Figure 4.15: Spectrograms of: (a) Speech sentence distorted by AWGN at 5dB; (b) The denoised sentence by soft thresholding using the proposed algorithm and PWPD.
4.6 Discussion Of Results

It is obvious from the first stage of evaluation, that a three from the considered techniques here demonstrate high capability in tracking various types of noise, but their performance accuracy varies depending on the rate of change of the noise under test. The minimum variance tracking-based method seems to offer the best performance in tracking the average noise variation with a minimum storage limit (two previous frames). On the other hand, the adaptive smoothing-based method noise offers a good level of precision in tracking the instantaneous variation of the noise with high degree of accuracy for cases of speech signals of relatively low SNRs. Presented results also demonstrate that the adaptability and trackibility of the quantile-based noise estimation method has been enhanced when a smoothing factor based on the \textit{posteriori SNR} is introduced as the proposed method illustrates. In particular, the proposed algorithm is shown to be more robust yielding a consistent performance in terms of both accuracy of estimation and instantaneous tracking of the noise over the investigated range of noise levels and types. This, added to its relatively low computational complexity, makes the proposed algorithm suitable for a wide range of speech enhancement and denoising applications.

In general, the frequency resolution of the subbands affect the capability of the presented techniques to track the noise, where there is a significant improvement in the performance when they have been implemented in PWPD compared with the SGWT under the same conditions. This can be explained as less noisy subbands signals can be obtained when decompose the speech signal into a critical bands. In Figures 4.7-4.9 (d), the proposed algorithm gives a perfect noise power tracking for different noisy types.

From the second stage of evaluation, we can say that in general, algorithms T1 and T3 proved to have a good noise tracking performance over widely varying conditions. In particular, T1 performs efficiently well when the noise level is high. On the other hand, the performance of the quantile based noise estimation seems to degrade for low SNR. However, as indicated before, its performance has been improved when
combined with smoothing factor as explained for the proposed algorithm. For different input SNR input values, algorithm T4 shows a better behaviour than T2 and T3 in terms of the improvement in ASSNR as presented by Figures 4.11 and 4.12 using PWPD. The results of this part support our previous claim about the resolution of the decomposition subbands, which is clearly related to the tracking level of the noise estimation technique.

4.7 Summary

The problem of accurate noise estimation for speech signals has been studied in this Chapter. Four noise power estimation techniques have been considered when implemented using two WT decompositions, and their performance has been evaluated. A discussion of performance of each technique based on comparisons with each other was also affected.

In Chapter 6, two speech enhancement systems will be implemented and tested. The first system combine both a PWPD and subband noise estimation, while the second system has been proposed in Chapter 2 and its based on two stages of speech classification and subband noise estimation as mentioned previously in Chapters 4 and 5.
Chapter 5

Voiced/Unvoiced classification algorithm using dyadic wavelet transform

5.1 Overview of Existing Speech Classification Algorithms

Classification of speech into voiced and unvoiced (V/UV) regions is an important and necessary process in many high-quality speech-processing applications, such as speaker identification and verification, speech coding and synthesis, and speech enhancement using wavelet thresholding as indicated in Chapter 2. Essentially, the classification involves the determination of whether the speech is produced by vibration of the vocal cords [54]. The process can effectively be performed using a single feature or parameter which is closely associated with the voicing and non-voicing activities of the speech. However, due to the fact that the range of values of any single speech parameter overlaps between different regions, the accuracy of a single-feature V/UV classification method would be limited [55]. In the context of a single-feature approach, the problem of speech classification has been traditionally associated with the determination of pitch frequency. However, the vibration of the vocal cords does not necessarily result in periodicity of the speech signal and, therefore, a failure in detection of periodicity in some voiced regions would result in V/UV classification errors [57].

Several classification algorithms that employ more than one speech feature, such as the pitch frequency, zero-crossing rate, Cepstral peaks and/or short-time energy distribution, have been reported to offer better accuracy compared to single-feature
approaches [54-56]. In recent years, the wavelet transform (WT) has been shown to provide a powerful and computationally-efficient tool for a variety of signal and speech processing applications. These include a number of single-feature V/UV classification techniques based on a pitch determination approach [58,59], or based on the per wavelet band average energy distribution [60]. Other techniques use the cross-correlation function between different bands to detect the nature of the speech segment [61,62].

In this Chapter, we describe a new algorithm for V/UV speech classification using two features of the speech signal [63,64]: (a) the frequency distribution of the average energy and (b) zero-crossing rate, for each speech segment. Evaluation of the performance of the proposed algorithm, using a large database of both clean speech and speech degraded with additive noise, and compared to a single-feature wavelet-based method [36], and to a multi-feature Cepstrum-based method [55].

5.2 V/UV Classification Algorithm

The proposed V/UV is based on the computation of two features of the speech signal: (a) the average per band energy distribution in wavelet domain and (b) the zero-crossing rate. A block diagram representation of the algorithm is given in Fig 5.1. First the pre-emphasised speech signal, sampled at 8 kHz, is segmented into 20 ms long segments with 50% overlap using a sliding Hamming window. Prior to that, an optional noise suppression stage has been used to de-emphasis the out-of-band noise in the cases where the signal-to-noise ratio (SNR) of the speech is below 5 dB. This is achieved via the application of Butterworth band-pass filter with lower and upper cut-off frequencies of 200 Hz and 3400 Hz, respectively. After the pre-processing stage, the following analysis and feature-extraction processes are implemented:

5.2.1 Wavelet-based frequency distribution of the average energy

In this stage, each speech segment is decomposed into five different bands using a 4-level dyadic DWT, and the average energy of each band in the wavelet domain is computed.
Figure 5.1: Block diagram of the proposed V/UV speech classification algorithm

Essa Jafer  Chapter 5

Computation of Ratio ($R_i$) of High-band Energy to Low-bands Energy

Determination of Zero-crossing Threshold ($TH_2$)

Computation of Median of Zero-crossing Rate

High pass Filtering & Computation of Zero-crossing Rate ($ZCR_p$)

Computation of per Band Energy

Application of a 4-level DWT

Input Speech

Noise Suppression Filter

Segmentation & Pre- emphasising

V/UV classification based on Computed Values of $R_p$ & $ZCR_p$ Compared to $TH_1$ & $TH_2$

Logic Smoothing of V/UV Classification Data

V  UV
In general, an unvoiced speech segment should show energy concentration in the high frequency bands, while a voiced segment should show energy concentration in low frequency bands of the wavelet domain. Fig 5.2, for example, shows the result of the analysis of the energy distribution in the wavelet domain of a clean speech signal incorporating both voiced and unvoiced regions. It is clear from Fig 5.2 (b), where the per segment ratio ($R_p$) of the accumulated average energy of wavelet low-bands to that of the wavelet highest-band in the $p$th segment is shown, that a reliable V/UV separation could be achieved by comparing the average energy of the wavelet low-bands to those of the wavelet high-bands. Based on this and on an approach reported in [35], the ratio between the energy of the low bands (i.e. below 2 kHz) to that of the highest band (i.e. above 2 kHz) is computed and used in our algorithm as the first fundamental parameter in formulating the V/UV decision.

Let $EG_p$ be the total energy in the $p$th speech segment and $EG_i$ be the energy in wavelet band $i$, such that

$$EG_p = \sum_{i=1}^{5} EG_i$$

(5.1)

where $i=1$ represents the highest frequency band in the wavelet domain (i.e. above 2kHz), and $i=2,3,4$ and 5 represent the frequency bands occupying the frequency range 0-2 kHz. The above-mentioned ratio is then computed as:

$$R_p = \frac{\sum_{i=2}^{5} EG_i}{EG_1}$$

(5.2)

This parameter has been examined using a large database of clean and degraded speech signals of various lengths. Accordingly, a value of 0.95 was found to be the most appropriate criterion to discriminate between voiced and unvoiced segments. Hence, the first threshold for making the required V/UV classification using the DWT energy distribution, is deemed to be

$$TH1 = 0.95$$

(5.3)
such that if $R_p < 0.95$, the segment is most likely to be unvoiced; otherwise, the speech segment is likely to be voiced.

Figure 5.2: (a) The waveform of a clean unvoiced-voiced speech signal, and (b) the corresponding wavelet-based frequency distribution of the average energy in each segment.

5.2.2 Zero-crossing rate

For discrete-time signals, a zero-crossing occurrence is detected when two successive samples have different algebraic signs. Hence, a sufficiently accurate zero-crossing count can be affected by comparing the signs of successive samples. However, in speech signals, the existence of noise and DC offset significantly reduces the accuracy of such a simple measurement approach. In this algorithm, the speech signal is first filtered using a Chebyshev filter with a cut-off frequency of 100 Hz to minimise the above-mentioned effects. The zero-crossing rate corresponding to the $p$th segment of the filtered speech is then computed as follow:
\[
ZCR_p = \sum_{n=0}^{N-1} \left| \text{sgn}[y'_p(n)] - \text{sgn}[y'_p(n-1)] \right|
\]  

(5.4)

where \( N \) denotes the length, in samples, of the filtered speech segment, \( y'_p(n) \). Our investigation has shown that, when the zero-crossing rate for a given speech segment exceeds a given threshold, the segment is likely to be unvoiced. However, our investigation has also shown that the distributions of the voiced and unvoiced segments in the speech inevitably overlap. This is demonstrated in Fig 5.3, where histograms for the distribution of the zero-crossing rates resulting from analysing two speech signals taken from (a & b) two different male speakers and (c & d) two different female speakers are shown. In all cases, the length of speech signal used is 10s. Hence, a threshold for discrimination between voiced and unvoiced speech using the zero-crossing rate needs to be carefully determined. Depending on the above-mentioned property and on an approach reported in [55], it has been decided to use a continually updated value of the median of the zero-crossing rates in this algorithm as a second threshold, \( TH2 \):

\[
TH2 = \text{median} (ZCR_p)
\]

(5.5)

Therefore, if the measured zero-crossing rate of the \( p \)th segment, \( ZCR_p \), is equal or higher than \( TH2 \), the segment is most likely to be an unvoiced segment; otherwise, it is likely to be a voiced segment.

### 5.2.3 Two features V/UV classification

Although each of the extracted features was found to provide sufficiently accurate indication for classification of V/UV segments, further investigation showed that higher classification accuracy is achieved by using a decision based on the two features together. Fig 5.4 shows the wavelet-based per band energy distribution and the corresponding zero-crossing rates for unvoiced-voiced segments of the same speech signal over two conditions: (a) the signal is clean, and (b) the signal is degraded by additive white noise at SNR = 10 dB.
In both cases, values of each feature show total agreement with the criteria discussed in Sections 5.2.1 and 5.2.2 above. In addition, there is a clear correlation between the behaviour of the two features, providing strong basis for marking voiced and unvoiced regions within the signal.

Figure 5.3: Histograms of zero-crossing rates for utterances taken from (a),(b) two different female speakers.
Figure 5.3: Histograms of zero-crossing rates for utterances taken from (c),(d) two different female speakers.
Figure 5.4: The per band energy distribution and the zero-crossing rates feature for each segment of the speech signal used in Figure 4.2, when (a) the signal is clean, and (b) the signal is noise-degraded.
5.3 Implementation And Performance Evaluation Of The Proposed Algorithm

The proposed speech classification algorithm has been implemented and experimentally tested using a large database comprising a wide variety of speech records taken from the TIMIT database. The speech signals, which cover different speakers and utterances, were ranging from 2 to 10 s in length, and organized into 100 speech files corresponding to a total of 30000 speech segments. Different wavelet filters were considered for the dyadic DWT, however the algorithm was found to offer the best performance when a Haar filter is used. The signals were accurately labelled and their V/UV regions marked using a combination of visual inspection, listening and a dynamic pitch-tracking algorithm.

Using the algorithm, Fig 5.5 shows the per segment distribution of the two extracted parameters for the case of a speech signal taken from a male speaker, when the signal is (a) contaminated with White noise at 15dB SNR, and (b) contaminated with noise at 5dB. The accuracy of the V/UV classification obtained by the algorithm was evaluated using an objective error measure, VuVER%, which represents the percentage of the total number of segments that have been erroneously classified. The measure covers both voiced-to-unvoiced and unvoiced-to-voiced error rates. The performance of the proposed algorithm has been evaluated under various noisy conditions. Using a mark-space notation, Fig 5.6 shows V/UV frame-labling of a given speech signal when signal is (a) clean, (b) corrupted by AWGN at SNR=15dB, and (c) corrupted by AWGN at SNR=10dB. Here, the mark indicates a voiced segment and a space indicates an unvoiced segment.

Tables (5.1) and (5.2) show the results of the analysis for male and female speakers at different SNRs, generated by adding White noise to the clean speech signals.
Figure 5.5: The distributions of zero-crossing rates and per band average energy for a speech signal contaminated with noise at (a) SNR = 15dB, and (b) SNR = 5dB.
Figure 5.6: Voiced-Unvoiced labelled segments of speech signal at three different conditions: (a) Clean signal, (b) Noisy at SNR=15dB and (c) Noisy at SNR=10 dB.
Table 5.1: Performance of the proposed algorithm at different SNRs for three different male speakers M1, M2 and M3.

<table>
<thead>
<tr>
<th>SNR</th>
<th>VUVER %</th>
<th>SNR</th>
<th>VUVER %</th>
<th>SNR</th>
<th>VUVER %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>M1: 0.56 M2: 0.83 M3: 0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20dB</td>
<td>M1: 0.88 M2: 1.75 M3: 0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10dB</td>
<td>M1: 2.65 M2: 3.5 M3: 1.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5dB</td>
<td>M1: 3.35 M2: 4.26 M3: 2.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0dB</td>
<td>M1: 4.42 M2: 5.50 M3: 3.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Performance of the proposed algorithm at different SNRs for three different female speakers F1, F2 and F3.

<table>
<thead>
<tr>
<th>SNR</th>
<th>VUVER %</th>
<th>SNR</th>
<th>VUVER %</th>
<th>SNR</th>
<th>VUVER %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>F1: 1.15 F2: 2.17 F3: 1.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20dB</td>
<td>F1: 3.20 F2: 3.40 F3: 3.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10dB</td>
<td>F1: 6.57 F2: 4.64 F3: 6.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5dB</td>
<td>F1: 8.86 F2: 7.96 F3: 8.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0dB</td>
<td>F1: 10.15 F2: 10.28 F3: 10.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.4 Discussion of The Results

The results indicate that our algorithm performs well for both clean and noise-degraded speech. This is particularly clear from Fig 5.6 and Tables 5.1&5.2. The performance of our classification algorithm was also compared with a V/UV classification method based on a wavelet-based per band average energy distribution using the Daubechies (D8) filter [36], and found to offer significantly superior results. Taking the relative computational complexity into account, the algorithm was also found to provide comparable results to the two-pass, three-feature, Cepstrum-based classification technique reported in [43].
The results also show that for a highly noise-contaminated speech, better classification is achieved when dealing with male speakers. We believe that this phenomenon is due to the fact that female utterances usually contain shorter transient voiced segments as compared to male utterances. When contaminated with noise, the characteristics of such transient segments can be easily confused with those of unvoiced speech.

5.5 Summary

In this Chapter a new multi-feature wavelet-based algorithm for classification of speech signals into voiced/unvoiced segments has been introduced and its implementation and performance evaluation have been discussed. Presented results demonstrated that the proposed algorithm performs well for a variety of clean and noisy speech signals taken from different speakers. The algorithm has, hence, been incorporated into our proposed speech enhancement system, as the presented by the outline given in Section 2.5.

Having laid the grounds and developed all the required sub-algorithms, Chapter 6 proceeds to give details of the implementation and performance evaluation of the proposed speech enhancement system we outlined in Chapter 2.
Chapter 6

New Adaptive Speech Enhancement System

6.1 Introduction

Based on the outline given in Chapter 2, the implementation of a new novel speech enhancement system is considered in this Chapter. The system is implemented using two different approaches; the first approach is based on integrating a perceptual auditory model into a novel time-frequency adaptive wavelet based thresholding. The auditory model is based on the fact that the human auditory system tolerates additive noise as long as it is below some masking threshold. The use of a human auditory model has recently been proposed in subtractive-type enhancement techniques [6,76], to reduce the musical noise effect caused by most of the proposed speech enhancement techniques. However, although these techniques resulted in good quality speech with reduced level of musical noise, the excessive expansion of high frequency sub-bands, resulting from the use of the STFT or uniform subband wavelet packet decomposition UB-WPD, gives rise to the following drawbacks [75]:

(1) Increasing computational complexity.
(2) Degradation of the perceptual quality of the unvoiced sounds.
(3) Insufficient spectral resolution for dealing with voiced sounds.
Our speech enhancement system in its first approach utilises a Bark-scaled wavelet packet decomposition -see Chapter 3- integrated into a modified Weiner filtering process. The Weiner filtering uses a novel threshold estimation technique and a magnitude decision-directed approach [78,79]. First, a Bark-scaled wavelet packet decomposition BS-WPD is used to decompose the speech signal into critical bands. Threshold estimation is then performed for each wavelet band according to an adaptive noise level-tracking algorithm. Finally, the speech is estimated by incorporating the computed threshold into a Wiener filtering process, using the magnitude decision-directed approach.

The second implementation approach of our proposed system is based on two main processes: V/UV speech classification and subband noise estimation. First, the input noisy speech signal is decomposed into 6-bands using 5-levels discrete wavelet decomposition. In order to track the variation of noisy speech, wavelet coefficient threshold (WCT) is adapted according to the SNR of each frame. In a speech-dominated frame, the smaller WCTs are used to minimize the speech distortion. For those noise-dominated frames, we use larger WCTs to remove the wavelet coefficients WCs of noise. In parallel with this process, the speech segments will be classified into voiced/unvoiced using the technique described in Chapter 5, and the WCTs will be adapted according to the classification decision to preserve ineligibility of the speech signal.

The performance of the developed system is evaluated using different noise types and conditions. For the first implementation approach, two wavelet decompositions; the DWT and the SGWT, are employed to analyse the input signal and study the effect of each one on the overall performance of the presented system. The noise estimation is carried out using two different selected methods from those considered in Chapter 4. The assessment study of the second approach given here will identify the improvement in system’s performance that would be obtained when introducing each stage. Also we will discuss in comparative study the advantages and drawbacks of each approach and how further improvement can be achieved. At the end, a new Matlab based graphical interface system, which has been developed and used for implementing and evaluating all the techniques introduced in this research.
6.2 First Implementation Approach Speech Enhancement System

In the first approach, the proposed speech enhancement system begins by decomposing $y(k)$ into a number of critical bands using the BS-WPD method explained in Section 5.3, such that:

$$\{y^{l,n}(k)\} = BS-WPD[y(k)], \quad k = 0, \Lambda, M - 1$$  \hspace{1cm} (6.1)

where $\{y^{l,n}(k)\}$ represents the set of BS-WPD coefficients in the ($l$th, $n$th) band of the decomposition. The clean speech is estimated based on the outcome of the two following operations: a time-frequency dependent threshold estimation, and an adaptive soft thresholding technique using a modified Weiner filtering approach, as will be described in the proceeding Sections. Subsequently, the estimate of the clean speech is transformed back into the time-domain using an inverse BS-WPD, as illustrated in Fig 6.1.

The novelty of the proposed system is based on the exploitation of a new adaptive soft thresholding that is incorporated with a suppression filter [78,79].

![Figure 6.1: Schematic representation of the proposed speech enhancement system using BS-WPD.](image-url)

92
6.2.1 Time-frequency dependent threshold estimation

Signals denoising by classical wavelet thresholding can be performed as either ‘hard’ or ‘soft’ thresholding, as discussed in Section 3.5.2. However, for speech signals, hard/soft thresholding often results in time-frequency discontinuities, which leads to unnatural effects and further degradation of the perceptual quality of the enhanced speech. This effect is more pronounced when high noise levels corrupt the speech. Considering this drawback, a new adaptive time-frequency dependent threshold estimation method is proposed here. The method involves first the estimation of the noise level, $\sigma$, for every wavelet band and time frame using the newly proposed noise estimation algorithm described in Section 4.3.4.

6.2.2 Soft-thresholding using modified Weiner filtering & magnitude decision directed approach

The wavelet threshold technique used here is based on the modification of the noise suppression rule of the Ephraim and Malah algorithm (see Section 2.2.2). Based on the level of noise for the $(lth, nth)$ band at the $p$-th frame, we first calculate a posteriori signal to noise ratio as:

$$\left( y^{l,n}_p(k) \right)^{post} = \left| y^{l,n}_p(k) \right| \sigma_p^{l,n}$$  \hspace{1cm} (6.2)

It can be noticed that the above equation is identical to (2.9). The corresponding priori signal-to-noise ratio is estimated using a magnitude decision-directed approach (2.14) such that:
\[
\left(y^{l,n}(k)\right)_{\text{priori}} = \alpha \frac{\hat{y}^{l,n}(k - L_{\text{frm}})}{\sigma^{l,n}_{p-1}} + \left(1 - \alpha\right) \left[0.0 \left(y^{l,n}(k)\right)_{\text{post}} - 1\right]
\]

(6.3)

where \(\sigma^{l,n}_{p-1}\) is the estimated noise level of the previous frame, and \(0 \leq \alpha \leq 1\) is a factor that is used to control the degree of the suppression.

It has been reported that the noise suppression rule proposed by Ephraim and Malah [1,33], makes it possible to obtain a significant noise reduction while avoiding the musical noise phenomenon. Some modifications have been proposed to overcome the problem of the basic suppression rules that are incapable of eliminating completely the musical noise in the short-time frame as in [77]. Using the two estimated SNRs, a suppression function is computed as:

\[
G^{l,n}(k) = \frac{\left(y^{l,n}(k)\right)_{\text{priori}}}{1 + \left(y^{l,n}(k)\right)_{\text{priori}}} \left[\frac{1}{\left(y^{l,n}(k)\right)_{\text{post}}} + \left(y^{l,n}(k)\right)_{\text{priori}}\right]
\]

(6.4)

Finally, to obtain the estimate for the clean speech, the values of the wavelet coefficients of the noisy speech, \(y^{l,n}(k)\), are adjusted using the suppression function such that:

\[
\hat{y}^{l,n}(k) = G^{l,n}(k).y^{l,n}(k)
\]

(6.5)

6.2.3 Implementation and performance evaluation of the first approach

The evaluation of the proposed speech enhancement system has been performed using two objective measures: the segmental signal to noise ratio (SegSNR) of the enhanced speech compared to the that of the noisy speech, and the Cepstral CEP distance between the original clean speech signal and both the noisy and enhanced speech signals. The evaluation involved the use of a number of speech records, for male and female speakers, taken from TIMIT database. The speech signals were sampled at 16 kHz, and corrupted by different types of noise taken from the Noisex 92 database. In
all the evaluation cases, the BS-WPD is implemented with the discrete Meyer wavelet filters, which provides good separation of bands due to their regularity property. By means of illustration, Fig 6.2 and Fig 6.3 show the waveforms resulting from applying the proposed technique to speech signals corrupted by white and pink noise, respectively, both at SNR = 5 dB. For these tests, a value of 0.75 was used for controlling factor $\alpha$. Figure 6.4, on the other hand, shows the improvement in SegSNR of the enhanced signal obtained from the application of the proposed algorithm to speech signal distorted by three different noise types. For cases (1), (2) and (3), results are obtained from full-implementation of the proposed speech enhancement algorithm using White, F16 cockpit and Pink noise respectively as a function of the input SNR.

A relatively high improvement in the output SNR can be noted here since the suppression factor is 0.9. In case (4), the magnitude decision-directed soft thresholding approach of the proposed system is replaced by classical soft thresholding process. Using the Cepstral distance measure, tables 6.1-6.4 provide further indication about the performance of the proposed algorithm. In tables 6.1 and 6.2, two different speech signals taken from two different speakers (male and female), corrupted by AWGN at different SNRs, were tested. Tables 6.3 and 6.4 consider a male speech signal corrupted by two different type of additive noise: pink noise and F16 cockpit noise.

In Fig 6.5 the cepstral distance of the enhanced speech signal versus the input SNR, for $\alpha$ taken three different values of 0.50, 0.75 and 0.95 is shown. It is clear from the results given that the proposed speech enhancement algorithm gives a good robust performance over a wide range of noise types and levels.

6.2.4 Discussion of the results

For comparison, Tables 6.1-6.4 also indicates the performance of the speech enhancement using a classical soft thresholding algorithm where the value of $\alpha$ fixed to be 0.75. Our investigation has shown that for cases of high SNR (above 10dB), setting $\alpha$ to a value between 0.5 and 0.75 yields the best results. For lower SNR (below 5 dB), however, $\alpha \geq 0.9$ has been found to provide the best performance in terms of noise suppression as well as preservation of the naturalness and quality of
Figure 6.2: Waveforms of the (a) noisy speech (AWGN) with SNR=5dB, and (b) enhanced speech.

Figure 6.3: Waveforms of the (a) noisy speech (pink noise) with SNR=5dB, and (b) enhanced speech.
Fig 6.4: Performance of the proposed speech enhancement algorithm for three different noise types in terms of the improved SSNR.
the original speech signal. Fig 6.4 shows a little improvement in the output SNR of the enhanced speech signal (0-4dB) for cases (1), (2) and (3) where input SNRs is above 0dB. For input SNRs below 0dB, a relatively high improvement in the output SNR is obtained since the suppression factor is 0.9. The results presented by Figure 6.4 also show that a moderate performance for a wide input SNR range can be obtained with $\alpha=0.75$. This indicates that the system performance can be improved further by introducing an adaptive process for selection of the suppression factor $\alpha$ according to the estimated SNR in each frame.

The proposed noise estimation technique (Section 4.3.4) used in the system was selected since it gives a better performance in the perceptual wavelet domain and provides relatively low computational complexity compared with other techniques.

### Table 6.1: Comparison of Cepstral distances resulting form the application of the Classical Soft Thresholding and proposed algorithm on a Male speech corrupted by White noise.

<table>
<thead>
<tr>
<th>Input SNR (dB)</th>
<th>CEPSTRAL DISTANCE (MALE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noisy</td>
</tr>
<tr>
<td>10</td>
<td>0.85</td>
</tr>
<tr>
<td>5</td>
<td>1.56</td>
</tr>
<tr>
<td>0</td>
<td>2.67</td>
</tr>
<tr>
<td>-5</td>
<td>4.33</td>
</tr>
<tr>
<td>-10</td>
<td>6.57</td>
</tr>
</tbody>
</table>
Table 6.2: Comparison of Cepstral distances resulting from the application of the Classical Soft Thresholding and proposed algorithm on a Female speech corrupted by White noise.

<table>
<thead>
<tr>
<th>Input SNR (dB)</th>
<th>CEPSTRAL DISTANCE (FEMALE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noisy</td>
</tr>
<tr>
<td>10</td>
<td>0.78</td>
</tr>
<tr>
<td>5</td>
<td>1.38</td>
</tr>
<tr>
<td>0</td>
<td>2.43</td>
</tr>
<tr>
<td>-5</td>
<td>4.00</td>
</tr>
<tr>
<td>-10</td>
<td>6.12</td>
</tr>
</tbody>
</table>

Table 6.3: Comparison of Cepstral distances resulting from the application of the Classical Soft Thresholding and proposed algorithm on speech corrupted by pink noise.

<table>
<thead>
<tr>
<th>Input SNR (dB)</th>
<th>CEPSTRAL DISTANCE (PINK)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noisy</td>
</tr>
<tr>
<td>10</td>
<td>0.42</td>
</tr>
<tr>
<td>5</td>
<td>0.67</td>
</tr>
<tr>
<td>0</td>
<td>1.15</td>
</tr>
<tr>
<td>-5</td>
<td>3.34</td>
</tr>
<tr>
<td>-10</td>
<td>4.10</td>
</tr>
</tbody>
</table>
Table 6.4: Comparison of Cepstral distances resulting from the application of the Classical Soft Thresholding and proposed algorithm on speech corrupted by F16 cockpit noise.

<table>
<thead>
<tr>
<th>Input SNR (dB)</th>
<th>CEPSTRAL DISTANCE (F16 COCKPIT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noisy</td>
</tr>
<tr>
<td>10</td>
<td>031</td>
</tr>
<tr>
<td>5</td>
<td>0.57</td>
</tr>
<tr>
<td>0</td>
<td>1.23</td>
</tr>
<tr>
<td>-5</td>
<td>2.42</td>
</tr>
<tr>
<td>-10</td>
<td>3.10</td>
</tr>
</tbody>
</table>

6.3 Second Implementation Approach Speech Enhancement System

In this section, the system implementation approach that is based on adaptive wavelet thresholding using both V/UV speech classification and accurate subband noise estimation will be implemented and evaluated. The approach is outlined in the diagram shown in Fig 6.6. First the structure of the used wavelet decomposition is described. The adaptive threshold setting based on V/UV speech classification and noise estimation will be explained in details since it represents the core of the proposed implementation. The performance of the system is then evaluated in different noisy conditions and the results obtained are analysed and discussed.

6.3.1 Structure of the wavelet decomposition

Wavelet coefficients are obtained by decomposing each frame in a wavelet orthogonal basis as described in Chapter 3. This algorithm is cascades discrete convolution of low-pass and high pass filters, and then down samples the filtered
Figure 6.5: Performance of the proposed speech enhancement algorithm in a White noise: The enhanced cepstral distance in terms of $\alpha$ taken three different values.
The output from high-pass filter represents the detail coefficient and that of the low-pass filter represents the scale coefficient. A simple diagram of the filter bank is shown in Fig 6.7. The bandwidth of the input signal is 8kHz since the sampling frequency is 16kHz.

### 6.3.2 Voiced/Unvoiced speech classification

The algorithm developed in Chapter 5, will be employed in this implementation of the system to classify the speech segment into voiced/unvoiced. However, as described then, such algorithm is based on the statistical energy distribution of the AWGN in wavelet domain where the unvoiced parts are concentrated on the higher bands while the voiced parts can be detected in the lower bands. For other types of noise, different wavelet band energy distributions have been noticed by our investigation for the voice and unvoiced parts. Tables 6.5-6.10, show the distribution of the average energy for Pink noise, Car noise, and F16 cockpit noise added to speech signals at different SNR using 6-levels DWT. It can be concluded that such types of noise tend to be of low frequency random nature where the energy is more concentrated in the low frequency bands of the signal. Hence, the V/UV speech classification algorithm as proposed in Chapter 5 may not perform well for all noise types and may affect the performance of the speech enhancement system. According to the above, a second wavelet based V/UV speech classification algorithm based on pitch detection and independent of the subband energy distribution has been used to classify speech segments in different non-Gaussian noisy conditions [80,81]. To implement this algorithm, first the 5-dilation details obtained from the DWT decomposition have been examined. Results showed that when there is a voiced segment, the DWT exhibits local maxima that coincide with the instants of sharpest signal variation [80]. For voiced speech these points are close to the instants of glottal contraction or complete closure.

The wavelet based pitch detection algorithm used here is described in Table 6.11. The algorithm is essentially computed for a pair of consecutive details. A speech segment is classified as unvoiced when the local maxima of the DWT pairs, which are above 0.9 of the maximum of the upper leg of the first pair, occur at different instants.
Fig 6.6: Block diagram of the proposed speech enhancement implementation

Fig 6.7: Decomposition tree of the filter bank. H and L respectively represent the low-pass and high-pass filters that cascaded with down 2 operations.
The dilation details, \(i_{\text{min}} \leq i \leq i_{\text{max}}\), where \(i_{\text{min}} = 3\) and \(i_{\text{max}} = 5\), have been shown to filter out the unwanted higher frequency information, while providing good representation of the signal properties in the fundamental frequency range.

**Table 6.5: Per band average energy distribution for Pink noise using 5-levels DWT as obtained for voiced segments.**

<table>
<thead>
<tr>
<th>SNR</th>
<th>detail (1)</th>
<th>detail (2)</th>
<th>detail (3)</th>
<th>detail (4)</th>
<th>detail (5)</th>
<th>scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.70</td>
<td>1.11</td>
<td>6.26</td>
<td>32.40</td>
<td>18.43</td>
<td>41.00</td>
</tr>
<tr>
<td>5</td>
<td>0.66</td>
<td>0.66</td>
<td>6.70</td>
<td>37.00</td>
<td>20.15</td>
<td>34.74</td>
</tr>
<tr>
<td>0</td>
<td>1.27</td>
<td>3.25</td>
<td>6.56</td>
<td>20.15</td>
<td>14.00</td>
<td>54.80</td>
</tr>
</tbody>
</table>

**Table 6.6: Per band average energy distribution for Pink noise using 5-levels DWT as obtained for unvoiced segments.**

<table>
<thead>
<tr>
<th>SNR</th>
<th>detail (1)</th>
<th>detail (2)</th>
<th>detail (3)</th>
<th>detail (4)</th>
<th>detail (5)</th>
<th>scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>74.50</td>
<td>10.00</td>
<td>5.20</td>
<td>2.80</td>
<td>2.46</td>
<td>4.80</td>
</tr>
<tr>
<td>5 dB</td>
<td>65.67</td>
<td>8.75</td>
<td>5.30</td>
<td>3.80</td>
<td>3.70</td>
<td>12.60</td>
</tr>
<tr>
<td>0 dB</td>
<td>32.00</td>
<td>7.40</td>
<td>7.10</td>
<td>6.60</td>
<td>7.60</td>
<td>39.00</td>
</tr>
</tbody>
</table>

**Table 6.7: Per band average energy distribution for Car(Volvo) noise using 5-levels DWT as obtained for voiced segments.**

<table>
<thead>
<tr>
<th>SNR</th>
<th>detail (1)</th>
<th>detail (2)</th>
<th>detail (3)</th>
<th>detail (4)</th>
<th>detail (5)</th>
<th>scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 dB</td>
<td>0.01</td>
<td>0.11</td>
<td>6.36</td>
<td>54.80</td>
<td>15.33</td>
<td>23.37</td>
</tr>
<tr>
<td>5 dB</td>
<td>0.012</td>
<td>0.10</td>
<td>6.00</td>
<td>48.71</td>
<td>12.39</td>
<td>32.85</td>
</tr>
<tr>
<td>0</td>
<td>0.02</td>
<td>0.15</td>
<td>3.81</td>
<td>28.90</td>
<td>5.72</td>
<td>61.36</td>
</tr>
</tbody>
</table>

**Table 6.8: Per band average energy distribution for Car(Volvo) noise using 5-levels DWT as obtained for unvoiced segments.**

<table>
<thead>
<tr>
<th>SNR</th>
<th>detail (1)</th>
<th>detail (2)</th>
<th>detail (3)</th>
<th>detail (4)</th>
<th>detail (5)</th>
<th>scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 dB</td>
<td>72.00</td>
<td>8.70</td>
<td>4.25</td>
<td>2.27</td>
<td>1.54</td>
<td>10.48</td>
</tr>
<tr>
<td>5 dB</td>
<td>59.27</td>
<td>5.37</td>
<td>2.10</td>
<td>1.27</td>
<td>0.70</td>
<td>31.25</td>
</tr>
<tr>
<td>0 dB</td>
<td>29.70</td>
<td>2.03</td>
<td>0.50</td>
<td>0.40</td>
<td>0.30</td>
<td>67.00</td>
</tr>
</tbody>
</table>
Table 6.9: Per band average energy distribution for F16 cockpit noise using 5-levels DWT as obtained for voiced segments.

<table>
<thead>
<tr>
<th>SNR</th>
<th>detail (1)</th>
<th>detail (2)</th>
<th>detail (3)</th>
<th>detail (4)</th>
<th>detail (5)</th>
<th>scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>10dB</td>
<td>0.014</td>
<td>0.21</td>
<td>6.62</td>
<td>56.80</td>
<td>16.45</td>
<td>20.00</td>
</tr>
<tr>
<td>5</td>
<td>0.036</td>
<td>0.87</td>
<td>6.51</td>
<td>51.60</td>
<td>13.71</td>
<td>27.25</td>
</tr>
<tr>
<td>0</td>
<td>0.11</td>
<td>3.38</td>
<td>5.78</td>
<td>31.46</td>
<td>7.82</td>
<td>51.43</td>
</tr>
</tbody>
</table>

Table 6.10: Per band average energy distribution for F16 cockpit noise using 5-levels DWT as obtained for unvoiced segments.

<table>
<thead>
<tr>
<th>SNR</th>
<th>detail (1)</th>
<th>detail (2)</th>
<th>detail (3)</th>
<th>detail (4)</th>
<th>detail (5)</th>
<th>scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>10dB</td>
<td>74.48</td>
<td>10.00</td>
<td>5.30</td>
<td>3.06</td>
<td>2.13</td>
<td>5.11</td>
</tr>
<tr>
<td>5</td>
<td>63.80</td>
<td>6.61</td>
<td>3.42</td>
<td>3.11</td>
<td>2.03</td>
<td>21.02</td>
</tr>
<tr>
<td>0</td>
<td>32.00</td>
<td>6.00</td>
<td>4.43</td>
<td>6.00</td>
<td>4.72</td>
<td>46.83</td>
</tr>
</tbody>
</table>

It should be noted here that the two versions of the speech classification algorithms (the original described in Chapter 5 and the described above) implemented using SGWT with no further modifications although for the second algorithm, the choice of wavelet function is critical for the performance.

Table 6.11: Outline of the V/UV speech classification algorithm based wavelet pitch detection

**step1:** segment the speech signal into frames of 20 ms

**step2:** compute the DWT coefficients at scale $j$ and $j+1$.

**step3:** locate local maxima of the DWT at scale $j$ and $j+1$, which exceed 0.9 of the global maxima of scale $j$.

**step4:** compare locations of local maxima at scale $j$ with those at scale $j+1$. If they have the same locations, then indicate this segment to be voiced, and then go to the next step. If not, increase the scale counter $j$, by 1. When $j$ is 4 and no local maxima have been detected, then indicate the segment to be unvoiced.

**step5:** Move to next segment. If the segment counter reaches end of the file, then go to end, otherwise, to step2.
6.3.3 Estimation of noise profile and level

According to the evaluation tests given in Chapter 4, the adaptive smoothing and minimum variance noise power estimation techniques proved to have superior performance, among the presented techniques, over the range of noise conditions of interest. Hence in this implementation of the system we use these techniques. At this stage, the noise level $\sigma_e$ will employed to set an adaptive threshold. Both techniques involve the following operation: for each segment, the per band posteriori segmental signal to noise ratio will be computed as:

$$ SNR_{post}(p,i) = 10 \log \frac{\sigma_{y_i}^2}{\sigma_{e_i}^2} $$  \hspace{1cm} (6.6)

where

$$ \sigma_{y_i}^2 = \sum_k y_i^2 (pk + N) $$  \hspace{1cm} (6.7)

In (6.7), $p$ represents the frame index and $N$ is the length of the frame. The quantity calculated above is then used to introduce the adaptive threshold estimation as will be explained in the following section.

6.3.4 The proposed threshold setting

This stage, which represents the core of this approach of the proposed speech enhancement system, involves the integration of a new threshold based of the speech classification and noise power estimation. The threshold is determined by the following equation:

$$ G_i(p) = F_{SNR} \cdot F_{V/UV} \cdot \sigma \cdot \sqrt{2 \log_e N} $$  \hspace{1cm} (6.8)

where $F_{SNR}$ is the first threshold gain factor which corresponds to the estimated posteriori segmental signal to noise in band $i$, and $F_{V/UV}$ is the second threshold gain factor and related to the segment type, i.e. whether it is voiced or unvoiced. A detailed description of the two threshold gain factors are given in the proceeding sections.

106
6.3.4.1 Mean square error of the enhanced speech signal

If \( f'_\lambda(n) \) is the output of the thresholding algorithm with some threshold value \( \lambda \), the mean square error (MSE) is defined as the average power of the difference between the enhanced and the and the clean speech:

\[
MSE = E\{[f(n) - f'_\lambda(n)]^2\} \tag{6.9}
\]

where \( f(n) \) denote the clean speech signal. The goal of speech enhancement is to minimize this MSE. In terms of wavelet coefficients, the MSE can be expressed as:

\[
f'_\lambda(n) = \sum_k TH_s(C_{L,k}) \cdot \phi_{L,k}(n) + \sum_k \sum_{j=1}^L TH_s(D_{j,k}) \cdot \psi_{j,k}(n) \tag{6.10}
\]

where \( TH_s(D_{j,k}) \) and \( TH_s(C_{L,k}) \) are the details and scale wavelet denoised coefficients using soft-thresholding. By substituting (6.10) into (6.9), we get:

\[
MSE = E\left\{\left[\sum_k (c_{L,k} - C'_{L,k}) \phi_{L,k}(n) + \sum_k \sum_{j=1}^L (d_{j,k} - D'_{j,k}) \psi_{j,k}(n)\right]^2\right\} \tag{6.11}
\]

where \( C'_{L,k} = TH_s(C_{L,k}) \), and \( D'_{j,k} = TH_s(D_{j,k}) \), equation (6.11) can be further simplified by using two wavelet error notations \( E_{1,L,h}(n) \) and \( E_{2,L,h}(n) \) such that: \( E_{1,L,h}(n) = \sum_k (c_{L,k} - C'_{L,k}) \phi_{L,k}(n) \) represents the wavelet error estimation and \( E_{2,L,h}(n) = \sum_k \sum_{j=1}^L (d_{j,k} - D'_{j,k}) \psi_{j,k}(n) \) represents the scaling error estimation. By substituting these two terms in (6.11), we obtain the following:
The above equation indicates that the amount of error between the clean and enhanced speech signals depends on the estimated errors associated with various wavelet coefficients. Hence in order to minimize the total error, we have to adapt the threshold setting for each band according to the level of noise and the nature of the segment.

### 6.3.4.2 Setting the threshold gain $F_{SNR}$

A frame with large $SegSNR$ implies that the frame is speech-dominated. Furthermore, the speech is likely to mask the noise, so we can use smaller wavelet coefficients thresholds (WCTs). On the contrary, a frame with small $SegSNR$ implies that the frame is either a speech-pause region (Silence) or in a very noisy environment. Using larger WCTs means reducing the WCs of a signal. In order to reduce the MSE given by (6.12), the WCTs tend to be small for speech-dominated segments, and large for noise-dominated segments.

For speech-dominated segment, the speech signal may mask the background noise. Hence, reserving the WCs is sufficient to decrease the speech distortion. In this situation, the WCT of the segment must be adapted to a smaller value. On the contrary, the WCs are almost determined by the noise for a noise-dominated segment. Removing background noise can be accomplished by reducing value of the WCs. In this case WCT tends to be larger than that of speech dominated segment. Since a speech-dominated segment contains larger $SegSNR$ than that of noise-dominated segment, we employ a sigmoid function [53], to interpolate the WCT between subbands of noise-dominated and speech dominated segments. Usually the universal WCT tends to give an overestimation. Hence, a compensation factor $\tau$ is used in threshold gain factor $F_{SNR}(p, i)$, so that it can be evaluated as:

$$F_{SNR}(p, i) = \tau \left[ 1 - 1/(1 + e^{-c_{SNR}(p, i) + \delta}) \right]$$

(6.13)
Here both $\varepsilon$ and $\delta$ are the factors of the sigmoid function. The factor $\varepsilon$ controls the slope of transition, and $\delta$ is used for the centre-offset of the transition curve. Elevating $\varepsilon$ can decrease the transition range according to SegSNR. On the contrary, decreasing it would increase the transition range. Fig 6.8 shows three plots for a sigmoid function where $\varepsilon$ takes three different values 0.2, 0.4 and 1.0.

### 6.3.4.3 Setting the threshold gain $F_{V/UV}$

Generally, the corrupting noise is not a white signal. Most of the traditional wavelet-based speech enhancement techniques use WCT for each band. They can work well for White Gaussian noise conditions, but their performance deteriorates in colored-noise environments. In this work we adapt the WCT in each wavelet subband according to the segment type depending on whether it is voiced or unvoiced. The output of the speech classifier stage is employed to adjust the WCT in each band to decrease the degradation in the subjective criterion and alleviate the problem of over-filtering of high-frequency components of the UV segments. In applying the thresholding method to speech, it is important not to harm the unvoiced sound in the speech signal. Since the unvoiced sound contains many noise-like high intelligibility components, unvoiced regions should be first separated from the noisy speech, and then a different thresholding method should be applied.

By considering the Tables 6.5-6.10, it is obvious that each noise type has a particular different per band energy distribution for both voiced-unvoiced segments. According to this, the thresholding criteria are required to be adapted adequately with the segment type. Hence, we propose two different approaches for thresholding voiced and unvoiced segment for case of White and other low frequency noise types (Pink, Car, and F16cockpit).

A) **White noise thresholding.**

For unvoiced segments, the concentration of the average energy is distributed over the high frequency bands, i.e. bands 1 and 2, on the other hand, the average energy is
Fig 6.8: Plot of sigmoid function for $\alpha$ takes three different values.
distributed over the low bands, i.e. bands 3, 4, 5 and 6 in case of voiced segments. The White noise is a high pass noise signal and hence tends to interfere with the unvoiced segments. The proposed threshold algorithm as follows:

1- IF THE SEGMENT IS UNVOICED
   Negatively bias $F_{V/UV}$ for bands 1 and 2, i.e. a value between 0.25-0.75 is subtracted.
   Soft-thresholding all bands.
2- IF THE SEGMENT IS VOICED
   Positively bias $F_{V/UV}$ for bands 3, 4, 5, and 6, i.e. a value between 1.25-4.00 is added.
   Hard-thresholding bands 1 and 2.
   Soft-thresholding bands 3, 4, 5, and 6.

B) Other Noise types (Pink, Car, F16 cockpit).

For such type of noise, the noise energy is distributed over the low frequency bands (4, 5, 6) with more concentration on bands (4 and 6). The threshold steps as follow:

1- IF THE SEGMENT IS UNVOICED
   Negatively bias $F_{V/UV}$ for bands 1 and 2, i.e. a value between 0.25-0.5 is subtracted.
   Positively bias $F_{V/UV}$ for bands 4 and 6, i.e. a value between 2.00-4.00 is added.
   Soft-thresholding bands 1, 2, 3, and 5.
   Hard-thresholding bands 4 and 6.
2- IF THE SEGMENT IS VOICED
   Positively bias $F_{V/UV}$ for bands 4 and 6, i.e. a value between 1.20-1.25 is added.
   Soft-thresholding all bands.
The biasing values given above could be controlled empirically. The combination of two threshold approaches Hard - Soft for each band with different biasing could be considered as a new way to adapt the WCT according to the segment type.

As can be noted for the preceding sections, the estimated threshold value is based on three factors, noise level, subband SNR and segment type whether its voiced or unvoiced. Accordingly, we expect to have an efficient thresholding approach, which reduces the noise, and at the same time preserve the speech contents undistorted. One assumption can be made here, is that we have a prior knowledge about the nature of the noise whether it is Gaussian or not. The two-thresholding algorithms could be combined in stage-F (see Fig 6.6), such that the system will be more adaptable to deal with different noisy conditions. In the following section, the implementation and the performance in its both implementation approaches will be examined.

### 6.3.5 Performance evaluation of the second approach

The proposed system has been implemented over different noisy conditions and using both DWT and SGWT decomposition schemes. For the evaluation process, a wide range of speech signals selected from TIMIT database and different types of noise taken from Noisex92 database has been used. It should be mentioned that for all the evaluation tests, the speech signals are segmented into 50% overlapped frames of 32 ms length. The value of $\varepsilon$ was selected experimentally to be 0.2. Tables 6.12 and 6.13 show the evaluation results of the implemented system using DWT and SGWT respectively. The input SNR (0-20 dB) was chosen as an appropriate range combining both possible low and high noisy speech signals that are usually encountered in many systems like mobile telephony [82].

Four objective measures have been used to evaluate the system performance. These are: the output enhanced $\text{SNR}$, the average signal-to-noise ratio $\text{ASSNR}$, the Cepstral distance $\text{CEP}$ and the log area ratio $\text{lar}$. In both tests (DWT and SGWT), four different noisy types have been considered: AWGN, Pink, Car, and F16 cockpit noise.
Figures 6.9 and 6.10 show the waveforms of the clean and enhanced speech signals superimposed on the noisy one for two types of noise and using DWT. The spectrograms for both noisy and enhanced signals are also given. Figures 6.11 and 6.12 shows similar results to the above but for a speech signal decomposed using the SGWT.

The effect of the noise estimation technique on the performance of the system has been examined. It has been noted in Chapter 4, that both the adaptive smoothing and the minimum variance techniques have a superior overall performance. Figures 6.13 and 6.14 show the results of system implementation, in terms of improvement in the SNR of the enhanced speech, using the two noise estimation techniques and over two different noisy conditions. The tests have been repeated using SGWT as in Figures 6.15 and 6.16. A comparative study between the proposed implementation and three other systems, using DWT and SGWT are shown in Figures 6.17 and 6.20 for two different noise types. The three systems are: (a) classical soft thresholding, (b) soft thresholding and speech classification, and (d) soft thresholding and subband SNR estimation. The effect of wavelet decomposition bands number has been investigated, using DWT and two different noise types, as shown in Figures 6.21 and 6.22.

The effect of the wavelet filter on the performance of speech enhancement system has been investigated using DWT, and evaluated in terms of the improved output SNR and ASSNR measures, as shown in Figures 6.23 and 6.24. Four different wavelet filter types have been considered here: the Haar, the symlet(10), the coiflet(5), and the discrete Meyer. In Figures 6.25 and 6.26, four SGWT filters are tested: the Haar, the Daubechies 7-9, the Daubechies 6, and the cubic spline. The design steps for these filters are given in Appendix-I.

The last part of the evaluation process dealt with the subjective quality measure using the minimum opinion score MOS, as explained in Chapter 2. The MOS for the noisy, enhanced using DWT, enhanced using SGWT for different noise types are given in Tables 6.14-6.17.
Table 6.12: Performance of the proposed system using DWT

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>SNR (dB)</th>
<th>Improved GSNR</th>
<th>Improved SSNR</th>
<th>Cep-D (noisy)</th>
<th>Cep-D (enh)</th>
<th>Lar-D (noisy)</th>
<th>Lar-D (enh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>79.22</td>
<td>3.02</td>
<td>1.91</td>
<td>0.96</td>
<td>9.52</td>
<td>6.39</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>21.00</td>
<td>5.22</td>
<td>5.29</td>
<td>0.98</td>
<td>11.55</td>
<td>6.99</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>11.29</td>
<td>6.72</td>
<td>8.22</td>
<td>1.67</td>
<td>12.22</td>
<td>7.49</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>5.93</td>
<td>7.33</td>
<td>11.45</td>
<td>1.66</td>
<td>12.76</td>
<td>7.99</td>
</tr>
<tr>
<td>Pink</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20.29</td>
<td>1.11</td>
<td>1.19</td>
<td>1.06</td>
<td>3.56</td>
<td>2.36</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10.27</td>
<td>3.71</td>
<td>2.67</td>
<td>1.43</td>
<td>6.78</td>
<td>5.96</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6.30</td>
<td>5.04</td>
<td>3.78</td>
<td>1.47</td>
<td>8.43</td>
<td>6.74</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3.32</td>
<td>5.58</td>
<td>5.18</td>
<td>1.46</td>
<td>10.00</td>
<td>7.49</td>
</tr>
<tr>
<td>Car</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>41.20</td>
<td>2.20</td>
<td>0.24</td>
<td>0.20</td>
<td>2.01</td>
<td>1.79</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>27.33</td>
<td>6.76</td>
<td>0.63</td>
<td>0.37</td>
<td>4.61</td>
<td>4.11</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>20.24</td>
<td>9.44</td>
<td>0.96</td>
<td>0.52</td>
<td>6.39</td>
<td>5.13</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>13.37</td>
<td>11.41</td>
<td>1.48</td>
<td>0.61</td>
<td>8.37</td>
<td>5.72</td>
</tr>
<tr>
<td>F16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20.30</td>
<td>1.11</td>
<td>1.19</td>
<td>1.06</td>
<td>3.56</td>
<td>2.36</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10.27</td>
<td>3.71</td>
<td>2.67</td>
<td>1.42</td>
<td>6.78</td>
<td>5.96</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6.30</td>
<td>5.04</td>
<td>3.78</td>
<td>1.47</td>
<td>8.43</td>
<td>6.74</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3.32</td>
<td>5.58</td>
<td>5.18</td>
<td>1.46</td>
<td>10.00</td>
<td>7.49</td>
</tr>
</tbody>
</table>
Table 6.13: Performance of the proposed system using SGWT

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>SNR (dB)</th>
<th>Improved GSNR</th>
<th>Improved SSNR</th>
<th>Cep-D (noisy)</th>
<th>Cep-D (enh)</th>
<th>Lar-D (noisy)</th>
<th>Lar-D (enh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>20</td>
<td>66.00</td>
<td>2.62</td>
<td>2.14</td>
<td>0.91</td>
<td>9.53</td>
<td>6.76</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>18.03</td>
<td>5.00</td>
<td>5.43</td>
<td>1.26</td>
<td>11.54</td>
<td>7.64</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>10.10</td>
<td>6.23</td>
<td>7.50</td>
<td>1.69</td>
<td>12.21</td>
<td>8.01</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>5.24</td>
<td>7.05</td>
<td>11.71</td>
<td>3.86</td>
<td>12.72</td>
<td>8.50</td>
</tr>
<tr>
<td>Pink</td>
<td>20</td>
<td>17.05</td>
<td>0.84</td>
<td>1.35</td>
<td>1.10</td>
<td>6.15</td>
<td>5.55</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>7.57</td>
<td>2.62</td>
<td>3.63</td>
<td>2.23</td>
<td>8.70</td>
<td>7.35</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4.14</td>
<td>3.51</td>
<td>5.55</td>
<td>3.01</td>
<td>9.55</td>
<td>8.13</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1.78</td>
<td>4.00</td>
<td>8.11</td>
<td>4.52</td>
<td>10.12</td>
<td>9.03</td>
</tr>
<tr>
<td>Car</td>
<td>20</td>
<td>33.50</td>
<td>1.31</td>
<td>0.24</td>
<td>0.12</td>
<td>2.01</td>
<td>1.92</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>23.69</td>
<td>5.20</td>
<td>0.63</td>
<td>0.25</td>
<td>4.61</td>
<td>3.35</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>11.55</td>
<td>6.80</td>
<td>0.96</td>
<td>0.42</td>
<td>6.39</td>
<td>5.38</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>4.20</td>
<td>7.50</td>
<td>1.48</td>
<td>0.77</td>
<td>8.36</td>
<td>7.45</td>
</tr>
<tr>
<td>F16</td>
<td>20</td>
<td>10.41</td>
<td>0.81</td>
<td>1.19</td>
<td>0.52</td>
<td>3.56</td>
<td>2.64</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8.50</td>
<td>3.30</td>
<td>2.67</td>
<td>0.83</td>
<td>6.78</td>
<td>5.53</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5.39</td>
<td>4.54</td>
<td>3.78</td>
<td>1.20</td>
<td>8.43</td>
<td>7.35</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>2.85</td>
<td>5.05</td>
<td>5.18</td>
<td>2.02</td>
<td>10.00</td>
<td>8.18</td>
</tr>
</tbody>
</table>

Table 6.14: Subjective speech quality evaluation with White noise

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>MOS (NOISY)</th>
<th>MOS (ENHANCED) DWT</th>
<th>MOS (ENHANCED) SGWT</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2.804</td>
<td>4.082</td>
<td>3.722</td>
</tr>
<tr>
<td>15</td>
<td>2.645</td>
<td>3.818</td>
<td>3.542</td>
</tr>
<tr>
<td>10</td>
<td>2.292</td>
<td>3.324</td>
<td>3.176</td>
</tr>
<tr>
<td>5</td>
<td>2.075</td>
<td>3.222</td>
<td>2.962</td>
</tr>
<tr>
<td>0</td>
<td>1.884</td>
<td>3.060</td>
<td>2.521</td>
</tr>
</tbody>
</table>
### Table 6.15: Subjective speech quality evaluation with Pink noise

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>MOS (NOISY)</th>
<th>MOS (ENHANCED) DWT</th>
<th>MOS (ENHANCED) SGWT</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2.633</td>
<td>3.482</td>
<td>3.257</td>
</tr>
<tr>
<td>15</td>
<td>2.348</td>
<td>3.263</td>
<td>2.912</td>
</tr>
<tr>
<td>10</td>
<td>2.093</td>
<td>3.050</td>
<td>2.750</td>
</tr>
<tr>
<td>5</td>
<td>2.100</td>
<td>3.143</td>
<td>2.498</td>
</tr>
<tr>
<td>0</td>
<td>1.846</td>
<td>2.612</td>
<td>2.271</td>
</tr>
</tbody>
</table>

### Table 6.16: Subjective speech quality evaluation with Car noise

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>MOS (NOISY)</th>
<th>MOS (ENHANCED) DWT</th>
<th>MOS (ENHANCED) SGWT</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2.866</td>
<td>4.021</td>
<td>3.822</td>
</tr>
<tr>
<td>15</td>
<td>2.605</td>
<td>3.762</td>
<td>3.421</td>
</tr>
<tr>
<td>10</td>
<td>2.244</td>
<td>3.449</td>
<td>3.187</td>
</tr>
<tr>
<td>5</td>
<td>1.913</td>
<td>3.380</td>
<td>2.897</td>
</tr>
<tr>
<td>0</td>
<td>1.646</td>
<td>2.922</td>
<td>2.355</td>
</tr>
</tbody>
</table>

### Table 6.17: Subjective speech quality evaluation with F16 Cockpit noise

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>MOS (NOISY)</th>
<th>MOS (ENHANCED) DWT</th>
<th>MOS (ENHANCED) SGWT</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2.666</td>
<td>3.624</td>
<td>3.308</td>
</tr>
<tr>
<td>15</td>
<td>2.388</td>
<td>3.307</td>
<td>3.002</td>
</tr>
<tr>
<td>10</td>
<td>2.129</td>
<td>3.061</td>
<td>2.877</td>
</tr>
<tr>
<td>5</td>
<td>1.883</td>
<td>3.000</td>
<td>2.679</td>
</tr>
<tr>
<td>0</td>
<td>0.476</td>
<td>2.731</td>
<td>2.172</td>
</tr>
</tbody>
</table>
Fig 6.13: Performance of the proposed system using two noise estimation techniques and DWT with White noise.

Fig 6.14: Performance of the proposed system using two noise estimation techniques and DWT with Pink noise.
Fig 6.15: Performance of the proposed system using two noise estimation techniques and SGWT with White noise.

Fig 6.16: Performance of the proposed system using two noise estimation techniques and SGWT with Pink noise.
Fig 6.17: Comparative performance evaluation for the proposed with other three systems using DWT and White noise.
Fig 6.18: Comparative performance evaluation for the proposed with other three systems using DWT and Pink noise.
Fig 6.19: Comparative performance evaluation for the proposed with other three systems using SGWT and White noise.
Fig 6.20: Comparative performance evaluation for the proposed with other three systems using SGWT and Pink noise.
Fig 6.21: System performance for different decomposition bands, using DWT with White noise.

Fig 6.22: System performance for different decomposition bands, using DWT with F16 cockpit noise.
Fig 6.23: System performance evaluation in term of GSNG using different DWT filters.

Fig 6.24: System performance evaluation in term of Improved ASSNR using different DWT filters.
Essa Jafer

Chapter.6

Fig 6.25: System performance evaluation in term of GSNG using different SGWT filters.

Fig 6.26: System performance evaluation in term of Improved ASSNR using different SGWT filters.
6.3.6 Discussion of the results

For the speech enhancement system implemented in Section 6.3, the following comments have been noticed:

a. The system performance is robust to different types of noise as the presented results showed.

b. In general, distortion of the enhanced speech signals is much reduced for high levels of an additive noise, as the two objective measures $CEP$ and $Lar$ indicate. The two objective measures $GSNR$ and $ASSNR$ tend to have different behaviour for the same enhanced processed signal, such that $ASSNR$ is significantly improved when the noise level increased.

c. The use of SGWT gives a comparable performance to the DWT and speed up the implementation. Hence it is more suitable for real time systems.

d. From Figures 6.13-6.16, the adaptive smoothing noise estimation technique gives a better performance in White noisy conditions, while for other types like Pink noise; the minimum variance technique shows a better behaviour. This fact can be explained as follows: in the case of White noise, the subband SNR is much lower for the higher frequency bands than the lower one, so the adaptive smoothing which is based on the estimated subband SNR, can track the noise in the different bands of the wavelet decomposition adequately. For low frequency noise, the performance of the adaptive smoothing may be degraded due to the mixed distribution of the noise energy between the high and low bands, while the minimum variance technique can track the noise more efficiently regardless the decomposition band frequency.

e. The comparative performance evaluation presented by Figures 6.17-6.20 show the improvement in the ASSNR objective measure obtained by introducing the proposed system implementation. The soft-thresholding based only on speech classification gives a lower performance than the same system but based on the estimation of subband SegSNR. This fact means that the part of the system, which is related to the adaptive SNR estimation, has a greater effect on the overall performance of the implementation over different noisy conditions.

f. As given by Figures 6.21 and 6.22, the increase in the number of decomposition bands may either enhance slightly the performance of the system as the case of
White noise or significantly as for the Pink noise. It is clear that increasing the number of wavelet decomposition bands leads to enhance the frequency subband resolution, which may improve on the subband SegSNR estimation part of the system. As a result of this and based on comment (e), the overall system performance will be improved by increasing the number of decomposition bands but at the expense of the computational complexity.

g. Its obvious from Figures 6.23 and 6.24 that discrete Meyer wavelet filter gives the best performance among the other types of wavelet filters since it provides good separation of subbands due to its regularity property. In the case of SGWT, the daubichies7-9 filter has a good performance comparing to the other considered filters.

h. The tables 6.14-6.17 indicate that speech signals have been perceptually enhanced when applying the proposed speech enhancement system through different noisy conditions. The improved MOS indicates that the noisy speech signal can be enhanced while preserving most of the signal parts undistorted.

i. As a comparison between the approaches proposed in Section 6.2 and that one in Section 6.3, the speech enhancement system based on speech classification and adaptive noise estimation has a better performance over different noisy conditions. The second implementation is more suitable for real time system since the signal first is segmented and then each segment will be decomposed and processed. For the BS-WPD system, each subband signal will be segmented and processed, such that more computations are involved. It should be mentioned that BS-WPD speech enhancement system causes more distortion in the output enhanced speech signal especially for a high noise levels.

j. The drawback of second speech enhancement implementation is the need to have a prior knowledge about the noise type whether its White or not, while the BS-WPD speech enhancement does not require this information.

k. Different objective speech quality measures have been used to evaluate the performance of the system. The SegSNR, CEP and Lar have a high correlating factor to the subjective test and give a good indication to the quality of the enhanced speech signal.


6.4 Matlab based Graphical user Interface for speech enhancement systems

In this section, an efficient computer based system for implementing and evaluating the various algorithms and techniques throughout this work is presented. The system includes different graphical user interfaces (GUIs), which are linked to software code for the implementation of the speech classification, the noise estimation and the speech enhancement techniques. The system has been designed to operate under MS Windows based PC environment. This system can be used to study the performance of any of the given techniques without a prior knowledge of Matlab programming. For each window, a range of input and output parameters can be specified, read, or changed in a user-friendly manner. A simple flow diagram of the GUI links is given in Fig 6.27. The system comprises the following:

**Interface.1:** This is the main Interface where a user can move to any of the other three interfaces (1.1, 1.2, and 1.3) as shown in Fig 6.28.

**Interface.1.1:** This Interface is linked with the noise estimation techniques code as explained in Chapter 4. The multi-display window and other user features are shown in Fig 6.29. A user can specify the name of the tested speech signal, the noise type, the input SNR, the frame length and the overlapping ratio. The domain of implementation, PWPD or SGWT and the decomposition-tested band can be selected in this interface. The real and estimated noise power can be displayed for the selected band by using each technique, as well as the performance evaluation represented by the factor $\text{ARE}$.

**Interface.1.2:** This Interface is designed to deal with the proposed speech enhancement implementation presented in this section 6.3. The input and output user features are shown in Fig 6.30. Two main code are linked with this interface, the speech enhancement system code tested with AWGN and system code tested with other noise types. The user can select the domain of the implementation, the technique of noise power estimation and the noise type. The clean/noisy/enhanced speech signals are displayed in separate figures and can be listened to them at the same time. The user can move to the interface 2.1.1, as shown in Fig 6.31, which deals with the
speech classification algorithms. In this interface, the wavelet based energy distribution developed in Chapter 5 and wavelet based pitch tracking speech classification algorithms code are linked. The total number of the voiced/unvoiced segments can be displayed and indicated on the plot of the noisy signal.

**Interface 1.3** This interface deals with the BS-WPD system described in section 6.2. The input/output user features are shown in Fig 6.32, where the noisy and enhanced speech signals can be displayed separately. The user can select the thresholding approach to be either classical soft-thresholding or the soft-thresholding based suppression filter to examine the performance of the system.

*Figure 6.27: flow diagram of the computer based graphical user interface GUI*
Fig 6.28: Main Interface window.
Fig 6.29: Interface window of the evaluation system for the noise estimation techniques.
Fig 6.30: Interface window of the evaluation system for the speech enhancement based speech classification and noise estimation.
Fig 6.31: Interface window for speech classification test system.
Fig 6.32: Interface window of the evaluation system for speech enhancement based BS-WPD.
6.4 Summary

In this Chapter, the speech enhancement system proposed in Chapter.2 has been implemented and evaluated. The first implementation is based on BS-WPD and a new noise suppression approach. The second implementation is based on speech classification into voiced/unvoiced and subband noise estimation associated with a novel threshold setting gain function.

The performance of these two implementations has been tested over different noisy conditions and using different objective and subjective measures. The obtained results have been presented and discussed. In the next Chapter, the main conclusions drawn from this work and suggestions for future work will be given.
Chapter 7

Conclusions and Suggestions for Further Work

7.1 Conclusions

This thesis has dealt with the development of a practical speech enhancement system that is robust to different noisy conditions. The problem has been addressed in terms of the need to have appropriate algorithms capable of denoising and improving the quality of speech signals transmitted within real life environments. Also, the real time implementation, computational complexity and adaptivity have been taken into consideration when developing the system. The following points summarise the main findings of the work:

➢ As described in Chapter 6, two different versions of the new speech enhancement system have been developed. The first one is based on the use of BS-WPD and a new time-frequency adaptive thresholding. This approach integrates both perceptual critical bands decomposition and subband noise estimation, and utilizes the decision directed approach to estimate a priori SNR. From the results given in section 6.2.3, we have shown that the proposed approach is effective in suppressing additive noise while still preserving the quality of the original speech signal. Using the presented comparative study the proposed system outperforms these approaches based on classical wavelet soft-thresholding. The simulation tests show that such kind of adaptive noise suppression technique performs well when the noise level is relatively high.
The amount of noise reduction is controlled by the suppression factor, which also provides a trade off between noise suppression and keeping most of the speech parts undistorted. Our results also demonstrate that the performance of the used noise estimation technique is crucial for performance of the overall speech enhancement system. The computational complexity involved in this system is essential for the real time implementation. Introducing the Soft-Thresholding that is based on a modified Weiner filtering has alleviated the musical noise effects caused by the residual noise.

- Descriptions of the second version of the proposed speech enhancement system were presented in Section 6.3. From the discussion given in Chapter 2, it has been concluded that most of the existing speech denoising algorithms do not take the type of speech segment to be processed into consideration. Based on that, we proposed and developed a novel speech enhancement algorithm based on the V/UV classification of the speech and on the subband noise power. The final threshold gain value is adjusted by three factors: the first factor represents a voiced/unvoiced related bias; the second depends on the estimated SegSNR and; the third factor is the soft-threshold process which is based on the estimated noise level. From the presented simulation results, it has been shown that this version of the system is capable of removing background noise efficiently and suppressing the residual noise. The results also demonstrated that varying the thresholding approach according to the nature of noise makes the system more robust to different noisy conditions. A remarkable improvement in the output SNR has been obtained from the proposed system when applied and tested for different noise types. By integrating the two gain factors that are based on the estimated subband SegSNR and the V/UV speech classification, our proposed system was proven to give a better performance compared to classic denoising algorithms. Both the objective and subjective measures provided quantitive indications for the improvement in the quality and intelligibility of the enhanced speech signals, and proved that the proposed system preserves the speech contents undistorted and minimizing the musical noise artefacts.
Our investigation has also shown that use of SGWT has a comparable performance and speed up the implementation, which makes it more adequate for real-time implementation. Different DWT and SGWT filters have been tested and the results showed that the choice of the wavelet filter affect the system’s overall performance. The same thing can be said here about the noise estimation technique, which is essential for the proposed system to work effectively. Employing the adaptive smoothing technique in the proposed system is recommended more for AWGN case. According to this, the system could be further improved by introducing another adaptive criteria for both thresholding approach and the noise estimation technique.

The presented experimental results reveal that employing estimates of the subband SegSNR to determine the gain factor of WCT results in efficient removal of the background noise. From the comparative simulation tests, we concluded that SegSNR based gain factor is more effective on the performance of the system compared to that related to the V/UV speech classification. Moreover, increasing the decomposition bands leads to improvement of the system performance but at the expense of increased computational complexity.

The MOS measure was useful for testing the perceptual quality and intelligibility of the enhanced speech signals and there was a correlation between this measure and other objective quality measures.

Despite its need for a prior knowledge about the noise, the second version of the proposed speech enhancement system improves the quality of the speech when tested under different noisy conditions. However the second approach is more appropriate than the first for real-time applications, where the all the processes are applied once for each segment.
7.2 Suggestions For Further Work

As indicated for the above conclusions, the proposed system in its two versions offers a level of performance that is superior to many existing speech enhancement techniques, and adequate to many real-time applications. However, our investigation has also identified a number of issues that, if taken into consideration, would provide further improvement in the performance of the proposed system. These issues, which described in the following sections, would form the basis for further research in this project.

7.2.1 Lowering the complexity of the BS-WPD based system

The evaluation of the auditory model and the noise suppression algorithm of the proposed BS-WPD system were focused primarily on performance. The issue of computational complexity has been mentioned only briefly. However, in a practical setting, this issue can be very important if the processing needs to be performed on a portable device, or on a fixed device that processes several hundred channels independently.

The SGWT can be employed to implement the auditory decomposition and speed up the processing of the speech segments since the wavelet transform using the lifting scheme can be implemented without allocating auxiliary memory.

7.2.2 Improving the adaptivity of the BS-WPD based system

It has been shown in Chapter 6 that the effect of select of an appropriate value for the suppression factor has a considerable effect on the overall performance of the system. Hence it is very important to develop an adaptive stage, which is capable of handling this issue. An estimate of the subband SegSNR could be employed, for example, to pick out appropriate values for the suppression filter from small data storage. Such a process will definitely improve the performance of the system and make it more reliable in coping with different noisy environments.
7.2.3 The use of soft-decision VAD

The subjective simulation tests of the proposed speech enhancement system presented in Section 6.3 have shown that a small amount of an audible musical noise can sometimes be observed in the speech pauses of the enhanced signal. One possible of alleviating problem would be to use a soft-decision VAD that provides more information than a simple speech/silence decision, such as that proposed in [83]. The VAD can be added to the V/UV speech classification stage so that the segment could be classified into voiced/unvoiced/silence. The thresholding approach could be easily adapted when a silence segment detected so that noise could be removed entirely. What kind of information a VAD should generate and how the algorithm would use this information are issues that need to be addressed.

7.2.4 Application of the Perceptual model using SGWT

The wavelet decomposition of the version of the proposed system developed in Section 6.3 could be replaced by PWPD representing an auditory model. We believe that the system performance could be further improved by using such a model in conjunction with other system stages. On the other hand, the SGWT can be employed in order to make the system more adaptable and suitable for real time applications. The thresholding approach can be modified as well by incorporating other suppression techniques based on decision directed such as these presented in [1,34,65]. The suppression factor, $\alpha$, could be adjusted according to the nature of speech segment and the estimated subband SegSNR.
References


