UNIVERSITY OF GAZİANTEP
GRADUATE SCHOOL OF
NATURAL & APPLIED SCIENCES

IMAGE AND SPEECH SIGNAL ENHANCEMENT IN TIME-FREQUENCY DOMAIN VIA ADAPTIVE LIFTING STRUCTURES

Ph.D. THESIS
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

BY
HACİ TAŞMAZ
AUGUST 2009

Image and Speech Signal Enhancement in Time-Frequency Domain via Adaptive Lifting Structures

PhD Thesis
in
Electrical and Electronics Engineering
University of Gaziantep

Supervisor
Assoc. Prof. Dr. Ergun ERÇELEBİ

by
Haci TAŞMAZ
August 2009

UNIVERSITY OF GAZİANTEP
GRADUATE SCHOOL OF
NATURAL & APPLIED SCIENCES
ELECTRICAL AND ELECTRONICS ENGINEERING DEPARTMENT

Name of the thesis: Image and Speech Signal Enhancement in Time-Frequency Domain via Adaptive Lifting Structures
Name of the student: Haci TAŞMAZ

Exam date: 19.08.2009

Approval of the Graduate School of Natural and Applied Sciences

Prof. Dr. Ramazan KOÇ
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Savaş UÇKUN
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Assoc. Prof. Dr. Ergun ERÇELEBİ
Supervisor

Examining Committee Members	signature
Prof. Dr. Muhammet KÖKSAL	_____
Assoc. Prof. Dr. H. Gökhan İLK	_____
Assoc. Prof. Dr. Ergun ERÇELEBİ	_____
Asst. Prof. Dr. Vedat Mehmet KARSLI	_____
Asst. Prof. Dr. Sema KOÇ KAYHAN	_____

ABSTRACT

IMAGE AND SPEECH SIGNAL ENHANCEMENT IN TIME-FREQUENCY DOMAIN VIA ADAPTIVE LIFTING STRUCTURES

TAŞMAZ, Haci
PhD in Electrical and Electronics Engineering
Supervisor: Assoc. Prof. Dr. Ergun ERÇELEBİ
August 2009, 114 pages

This thesis addresses the problem of image and speech enhancement for various noise environments using adaptive lifting schemes. A new space adaptive lifting scheme algorithm is proposed for 1-D (speech) and 2-D (image) signals. The space adaptive lifting schemes provide better signal representation and better enhancement results. The proposed speech enhancement method aims to remove the noise in order to improve the quality and the intelligibility of the enhanced speech signal. In order to improve the quality of the enhanced speech signal, an auditory model (Critical Bands) is integrated with the proposed speech enhancement method. The single channel estimators are employed for subbband speech enhancement since they are practical. The proposed image enhancement method is based on space adaptive 2-D lifting scheme. The aim of proposed image enhancement method is to remove the noise while retaining significant features of the image. The gray-level noisy images are decomposed into subbands using the proposed space adaptive 2-D lifting scheme algorithm. Spatial domain estimators and wavelet thresholding-based estimators are used for subband image enhancement. The experimental and objective evaluation results show the performance of proposed speech and image enhancement methods.

Keywords: speech enhancement, space-adaptive lifting, wavelet, critical band analysis, single channel estimators, image enhancement.

ÖZET

GÖRÜNTÜ VE KONUŞMA SİNYALLERİNİN ZAMAN-FREKANS BÖLGESİNDE UYARLAMALI LİFTİNG YAPILARI KULLANILARAK PEKİŞTİRİLMESİ

TAŞMAZ, Haci
Doktora Tezi, Elektrik-Elektronik Mühendisliği
Tez yöneticisi: Doç. Dr. Ergun ERÇELEBİ
Ağustos 2009, 114 sayfa

Bu tezde, çeşitli gürültü ortamları için uyarlamalı kaldıraç yapıları kullanılarak konuşma ve görüntü pekiştirme problemi ele alınmaktadır. Tek boyutlu (konuşma) ve iki boyutlu (görüntü) işaretler için yeni bir uzam uyarlamalı kaldıraç yapısı önerilmektedir. Uzam uyarlamalı kaldıraç yapıları daha iyi bir işaret temsili ve daha iyi pekiştirme sonuçları sağlar. Önerilen konuşma pekiştirme yöntemi, pekiştirilen konuşma işaretinin kalitesini ve anlaşılabilirliğini geliştirmek için gürültüyü ayırmayı amaçlar. Pekiştirilen konuşma işaretinin kalitesini arttırmak için, bir işitsel model (Kritik Bandlar) önerilen konuşma pekiştirme metoduyla bütünleştirilmektedir. Altband konuşma pekiştirmesi için pratik olduklarından tek kanallı konuşma pekiştirme kestirimcileri kullanılmaktadır. Önerilen görüntü pekiştirme yöntemi uzam uyarlamalı iki boyutlu kaldıraç yapısına dayanmaktadır. Önerilen görüntü pekiştirme yönteminin amacı, görüntünün önemli detaylarını korurken gürültüyü ayırmaktır. Gri-düzeyi gürültülü görüntüler, önerilen iki boyutlu uzam uyarlamalı kaldıraç yapısı algoritması kullanılarak altbandlara ayrıştırılmıştır. Altband görüntü pekiştirmesi için uzamsal bölge kestirimcileri ve dalgacık eşik temelli kestirimciler kullanılmaktadır. Deneysel ve objektif değerlendirme sonuçları önerilen konuşma ve görüntü pekiştirme yönteminin başarısını göstermektedir.

Anahtar kelimeler: konuşma pekiştirme, uzam-uyarlamalı kaldıraç, dalgacık, kritik band çözümlemesi, tek kanallı kestirimci, görüntü pekiştirme.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Assoc. Prof. Dr. Ergun ERÇELEBİ for his supervision, guidance and insight throughout this study.

I would like to thank Assoc. Prof. Dr. H. Gökhan İLK for his helpful criticism, guidance and encouragement.

I also like to thank Ercan UYGUN and M. Emin ÖZCAN for their valuable support.

Finally, I wish to thank my family for their patience and continuous support throughout my study.

CONTENTS

LIST OF FIGURES

LIST OF TABLES page

LIST OF SYMBOLS

a	inner product
$w(t)$	Windowing function
f	Frequency
τ	Shifting (time) parameter
*	Complex conjugate operator
$\psi(t)$	Mother wavelet
$\psi_{ab}(t)$	Daughter wavelets
$\hat{\Phi}_{xx}(k,l)$	Short-time power spectral density of $x(n)$
$\hat{\Phi}_{dd}(k,l)$	Short-time power spectral density of $d(n)$
ξ_k	Priori SNR estimation (or suppression factor)
$E\{.\|.\}$	Conditional expectation operator
$P\{.\|.\}$	Conditional probability operator
$P(H_1^k \| Y_k)$	Soft decision modification of optimal estimator under the signal presence uncertainty
$p\{.\|.\}$	Conditional probability density operator
$\Lambda(k)$	Generalized likelihood ratio
q_k	Probability of speech absence in the kth frequency bin
$G_{STSA}(k)$	Original gain function for MMSE-STSA estimator
$G_{LSA}(k)$	Original gain functions for the MMSE-LSA estimator
$G_{MM-LSA}(k)$	Multiplicatively-modified gain functions for the MMSE-LSA estimator
$G_{W\,\mathrm{mod}}(k)$	Multiplicatively-modified gain function for STSA-Wiener estimator
γ_k	A posteriori SNR estimation
η_k	Unconditional a priori SNR estimation
$\lambda_b(k)$	Noise power spectral density estimation
α_b	Smoothing factor for noise PSD estimation
$R_k^2(l)$	Noisy speech signal power spectrum
$P[.]$	Half-wave rectification
$\hat{A}_k(l-1)$	Estimated speech spectrum at previous frame
$bw(j,p)$	Frequency bandwidth [Hz] value corresponding to node (j,n) of the full WPD tree
$x(i,j)$	Original image
$y(i,j)$	Noisy image

$b(i,j)$	2-D noise signal
$\hat{\sigma}_b$	Standard deviation estimate of noise
$\hat{\sigma}_b^{\,2}$	Estimated noise variance
$\hat{\sigma}_x^{\,2}$	Estimated signal variance
$\hat{\sigma}_Y$	Noisy observations at kth scale
β	Scale parameter
L_k	Length of the subband at kth scale
J	Total number of scales $(k = 1,2,3....J)$
$x[n]$	Original signal
$\hat{x}[n]$	Processed (enhanced) signal
$y[n]$	Noisy signal
$b[n]$	Noise signal
$\vec{a}_x$	Linear prediction coefficient vector of clean
$\vec{a}_{\hat{x}}$	Processed speech signal
R_x	(R+1) x (R+1) (Toeplitz) autocorrelation matrix

LIST OF ABBREVIATIONS

ASR	Automatic speech recognition
Ad-WF2	Adaptive 2-D Wiener filter
CBW	Critical bandwidth
CB-WPD	Critical-band wavelet packet decomposition
CDF	Cohen-Daubechies-Feauveau
CWT	Continuous wavelet transform
DD	Decision directed
DFT	Discrete Fourier transform
DRT	Diagnostic rhyme test
DWPT	Discrete wavelet packet transform
DWT	Discrete wavelet transform
FIR	Finite impulse response
FWT	Fast wavelet transform
HH	(high-high) diagonal wavelet coefficient
HL	(high-low) vertical wavelet coefficient
HMM	Hidden Markov model
IIR	Infinite impulse response
IS	Itakura-Saito distance
LAR	Log-area ratio
LDT	Level-dependent thresholding
LH	(low-high) horizontal wavelet coefficient
LL	(low-low) approximation coefficient
LPC	Linear prediction coefficient
LSA	Log-spectral amplitude
MAP	Maximum a posteriori
MAP	Maximum a posteriori
ML	Maximum likelihood
ML-STSA	Maximum likelihood-short time spectral amplitude
MM-LSA	Multiplicatively modified log-spectral amplitude
MMSE	Minimum min square error
MMSE LSA	Minimum mean-square error log-spectral amplitude
MMSE-STSA	Minimum mean-square error short-time spectral amplitude
Mod-STSA	Modified short time spectral amplitude
Mod-WF	Modified Wiener filter
MOS	Mean opinion score
MRA	Multiresolution analysis
MRT	Modified rhyme test
MSE	Mean-square error
MSS	Magnitude spectral subtraction
PD	Probability distribution
PDF	Probability density function
PESQ	Perceptual evaluation of speech quality

PR	Perfect reconstruction filter
PSD	Power spectral density
PSNR	Peak signal to noise ratio
PSQM	Perceptual speech quality measure
QMF	Quadrature mirror filter
RDWT	Redundant discrete wavelet transform
RGB	Red-Green-Blue
SegSNR	Segmental signal to noise ratio
SNR	Signal to noise ratio
STD	Standard deviation
STFT	Short time Fourier transform
STSA	Short time spectral amplitude
SWT	Stationary wavelets transform
UWT	Undecimated wavelet transform
VAD	Voice activity detector
WF	Wiener filter
WGN	White Gaussian noise
WPD	Wavelet packet decomposition
WPT	Wavelet packet transform
WT	Wavelet transform

CHAPTER 1

INTRODUCTION

1.1 Overview of Speech Enhancement

The most natural and efficient tool for communication between people is speech. Speech communication is often affected by noise and environmental conditions. The human auditory system is known to be robust against the most common adverse conditions; however, the speech acquisition devices are not so much robust against the adverse environments.

When a speech acquisition device is used in such a noisy environment, the quality and the intelligibility of the transmitted speech is degraded due to the noise. This degradation may be very troublesome, especially in mobile communications where hands-free devices are used. Use of speech enhancement algorithms is always advisable in such communication devices [1].

The problem of enhancing speech degraded by noise is largely open to research. Many effective techniques have been introduced over the past decade since there are many areas where it is necessary to enhance the quality of speech degraded by background noise. Some of these areas include:

- Car interiors for cellular
- Helicopter and aircraft cockpits
- Hands free telephones
- Automatic speech recognition (ASR) systems
- Hearing aids and cochlear implants
- Restoration of historical recordings

Although, the speech enhancement systems may vary depending on the place it is used, a general speech enhancement system is given in Figure 1.

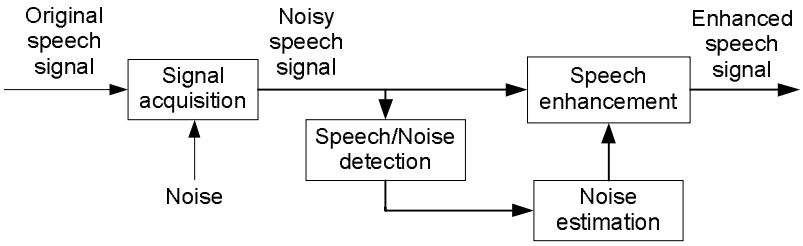

Figure 1.1 Description of a general speech enhancement system

The aim of a speech enhancement system may be to improve the perceived quality while preserving the intelligibility of processed speech. This is achieved by minimizing the effect of noise in order to reduce the listener's fatigue or to obtain as much noise free records as possible. The speech quality is a subjective concept since the final speech quality should be assessed by the human listener. The quality of speech is the degree to which listeners perceive the naturalness or pleasantness while the intelligibility is the degree to which the speech is correctly understood by the listeners.

1.1.1 Speech Characteristics

Speech is non-stationary signal carrying information in its fluctuations varying with time and frequency. Moreover, the consecutive samples of speech signal are highly correlated. Generally, the speech signals are processed frame-by-frame, with frames having 10-30 ms durations during which the speech signal is considered to be quasi-stationary. Speech bandwidth varies approximately from 50-4000 Hz. Speech signals are composed of voiced and unvoiced sounds. The voiced speech has high amplitude and energy at low frequencies and unvoiced speech has lower energy at higher frequencies. The speech is produced by human vocal tract. The vocal track is an acoustic tube which is limited by the vocal cords and the lips. The vocal track is characterized by its natural frequencies (formants) corresponding to resonances in its sound transmission characteristics [2, 3].

A simplified speech production model based on parameters of vocal tract is given in Figure 1.2.

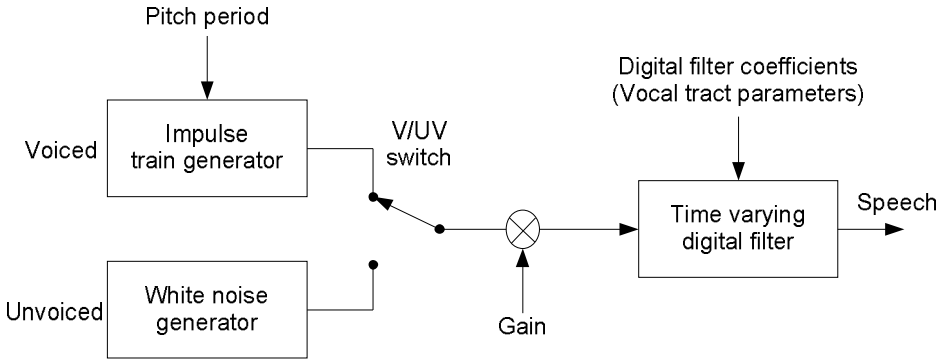

Figure 1.2 Digital model for speech production

Another important concept in speech enhancement is the speech perception. The models taking into account the aspects of human perception are generally based on properties of the human auditory system. A speech enhancement system taking into account the aspects of human perception may lead to improved perceived quality and intelligibility of the enhanced speech.

1.1.2 Noise Characteristics

The nature of the noise is an important criterion in choosing a particular speech enhancement method. Choosing an appropriate noise model is also significant. Noises may have different statistical and spectral properties as given below. The following classification can be made based on the characteristics of the noise.

- Uncorrelated noise: Additive background noise existing in many noisy environments (cars, offices, street, machines, windy conditions, factories and aircraft cockpits). It can be stationary, slowly varying or non-stationary.
- Speech babble noise: Noise interfering due to other speakers.
- Correlated noise: Reverberation and echoes.
- Non-additive noise: Noise caused by transmission line or transmission

channel distortion.

Among these, the non-stationary noise is the most difficult noise type to remove when a priori information of noise is not available.

1.1.3 Literature Summary for Speech Enhancement

Speech enhancement systems designed so far, differ in the number of channel they use (single or dual-channel approaches) and the domain in which they operate (time domain, frequency domain or time-frequency domain).

In dual-channel approach a second microphone provides a reference noise to better characterize changing noise statistics, which is necessary to deal with the non-stationary noise. A well known dual-channel approach for noise cancellation has been proposed by Widrow et al.[4] where a primary sensor is used for the corrupted speech signal, while a second sensor for noise. One of the advantages of the dual-channel approach is that it enables to process the speech corrupted with either stationary or non-stationary noise.

Single-channel approaches are more difficult to implement since there is no reference noise source however, they are more general and widely preferred by researchers. In the single-channel approach the noise must be eliminated during the silence frames of the noisy speech and it is assumed that the noise is stationary during speech activity [5].

Some of the popular single channel speech enhancement methods which have been developed in the last decade are as follows:

The spectral subtraction based methods are the most popular among the many available single-channel speech enhancement methods for its effectiveness and easiness [6, 7]. The method is based on subtracting the estimate of average noise spectrum from the noisy speech spectrum to obtain the magnitude estimate of clean speech. The main drawback of spectral subtraction is that, it causes residual and unnatural musical noise. The residual noise refers to the broadband noise that has the same perceptual characteristics as the original noise. The musical noise is the

synthetic musical tones due to the presence of the random short-duration spectral peaks in the noise spectrum [8].

The single channel speech enhancement algorithms based on minimum mean-square error (MMSE) estimation have received considerable attention in the past two decades [9-11] and widely used by researchers owing to their success in elimination of musical noise. The methods are based on a priori signal-to-noise ratio (SNR) estimation, Gaussian statistics and short–time Fourier transform (STFT). By applying a spectral gain to each frequency bin in a short-time frame of the noisy speech signal the spectral components of clean speech can be estimated. Since the spectral components are assumed to be statistically independent Gaussian variables, the gain is adjusted individually as a function of the relative local SNR at each frequency bin.

A vast amount of work has been emerged on the development of the soft decision noise suppression filters [12, 13]. In this approach, a spectral decomposition of a frame of noisy speech is performed and a specific spectral line is attenuated depending on the amount of measured noisy speech power exceeding an estimate of the background noise power.

Model based speech enhancement methods can be found in [14, 15]. Model based approaches can be classified in two groups.

- Speech enhancement based on speech production model (AR model)
- Speech enhancement based on Hidden Markov Model (HMM). The HMM is based on statistical model of speech and noise which is estimated during a training sequences.

The main disadvantage of AR models is that they are memoryless. If a given AR model is chosen for the current speech frame, certain AR models are more likely to occur in the following frame.

The HMM is used to model the probability distribution (PD). The HMM based systems give better results than the traditional speech enhancement methods, especially at low SNRs for speech corrupted by non-stationary noise. However, HMM based models require a training sequence and they cause high computational

cost. After obtaining the statistical parameters of speech and noise by a training sequence, the speech is estimated either by maximum a posteriori (MAP) estimation, leading to an iterative algorithm or MMSE estimation, where the filter weights are directly estimated from the noisy signal.

Most of the single channel speech enhancement methods designed so far aims to reduce the noise to improve the SNR of the enhanced speech signal. However, they can not improve the intelligibility of the enhanced speech signal. Recently, an increasing number of researchers have designed speech enhancement methods which include characteristics of the human auditory system [16-18] in order to improve the intelligibility of the speech signals. These perceptual models are generally based on critical-band decomposition or noise masking properties. The improvement in the intelligibility does not affect the objective quality of the speech signal, which means that a speech signal with good intelligibility may have poor objective quality or vise versa.

In the past decade, the wavelet transform has become a popular tool in speech enhancement applications for analyzing the non-stationary signals. It was developed to overcome the shortcomings of the STFT which is also capable of analyzing non-stationary signals. However, the limitation of STFT is that it uses a fixed window length for all frequencies. Once the window length is chosen, the time resolution is the same for all frequencies. On the other hand, the wavelet transform uses a variable window length. It provides good time resolution (poor frequency resolution) at high frequencies, while providing good frequency resolution (poor time resolution) at low frequencies [19-23].

A well-known wavelet-based speech enhancement method is wavelet thresholding (or shrinkage) proposed by Donoho at al. [24-26]. However, the STFT-based speech enhancement filters (or estimators) can be extended to the wavelet domain [27].

Furthermore, the wavelet packet transform (WPT) provides easy handling of the spectral content of speech signal, variable frequency resolution in each subband and exploiting the frequency subbands of interest [28].

By adjusting the subbands of WPT according to critical-bands of the human auditory system, a perceptual filterbanks which lead to efficient speech enhancement algorithms can be designed [29-31].

An alternative method for constructing biorthogonal wavelet transform, the lifting scheme [32, 33], has the following advantages over the classical wavelet transform.

- it is a spatial domain method
- easier to implement
- allows faster and in-place calculations
- allows nonlinear, adaptive, irregularly sampled and integer to integer wavelet transforms
- easier to obtain inverse transform

Furthermore, any wavelet transform can be factored into lifting steps [34]. Besides the above given advantages, the lifting scheme has a limitation. Since the filter structure is fixed, it can not adapt to the sudden changes in the input signal. However, a lifting scheme which adapts itself to the signal structure is desirable in many applications. This is achieved by allowing the lifting scheme to adapt its prediction or update filters to the local properties of the signal, which leads to adaptive lifting schemes [35].

The motivation behind introducing adaptivity into the lifting scheme is that, choosing better prediction filters (in the update-first lifting scheme) will give rise to more efficient signal representations. Some of the adaptive lifting algorithms developed by the researchers in the last decade include:

G. Piella and H. Heijmans [35] designed an adaptive update lifting scheme. The

adaptivity is achieved in the update stage and no bookkeeping is required for perfect reconstruction.

Yonghong at al. [36] proposed a spatially adaptive lifting scheme for 1-D signal denoising. Their adaptive algorithm chooses the Haar wavelet near the edges and

CDF (2, 2) filter for smooth parts of the signal, based on comparing the derivative of samples with a threshold coefficient.

R. L. Claypoole at al. [37,38], proposed "scale adaptive" and "space adaptive" update-first lifting algorithms for 1-D signal denoising. In the scale adaptive case, the prediction filter is adapted to the signal structure within each scale by minimizing prediction errors. In the space adaptive case, the prediction filter chosen from a family of prediction filter is adapted to the signal structure for each sample point, based on "edge avoiding prediction" method.

1.2 Overview of Image Enhancement

The image degradation caused by transmission errors, faulty acquisition devices or atmospheric disturbances is an important issue in image processing applications. Interpretation or visual perception of a noisy image is difficult for human observers. Such a noisy image needs to be enhanced. Furthermore, noisy images sent by satellites or medical devices cannot be directly processed. A pre-processing is always required where image enhancement is one of the important steps in the pre-processing stage.

The difficulties encountered in image enhancement are in general two types. The image edges are often blurred when noise is removed or when edges are retained and enhanced, image noise is strengthened too. Therefore, finding an image enhancement method which can both remove the noise and retain the edges or the other significant details of image is still a challenging task for researchers [39].

The aim of image enhancement is to remove the noise content while preserving significant features of the image. Improved visual quality of an image (or better input for other image processing stages) can be achieved by aid of image enhancement.

1.2.1 Image Characteristics

An image may be represented by a two-dimensional function $f(x,y)$, where x and y are spatial coordinates. The amplitude of f at any pair of coordinate (x,y) is

called as the intensity or gray level of the image at that point. The term gray level is used to describe the intensity of monochrome images.

Color images are composed of a combination of individual 2-D images. For instance, in the RGB color system, a color image consists of three (red, green and blue) individual component images. Therefore, many of the techniques developed for gray-level images can be extended to color images by processing the tree component images individually.

An image may be continuous with respect to coordinates (x, y) and amplitude of f. Both the coordinates and the amplitude need to be digitized to convert such an image to digital form. The process is called as sampling and quantizing. Thus, when x, y and amplitude of f are all finite discrete quantities (pixels), the image is called as digital image. According to coordinate convention used in MATLAB a digital image is represented as given in Figure 1.3.

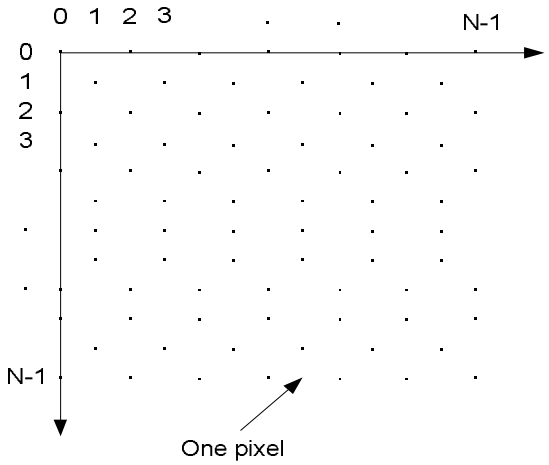

Figure 1.3 Coordinate convention for digital images

The digital images are represented by matrices. For example, an image $f(x, y)$ of M rows and N columns is called as an image of size MxN and can be given in the matrix form as follows.

$$f(x,y)=\begin{bmatrix} f(0,0) & f(0,1) & .... & f(0,N-1) \\ f(1,0) & f(1,1) & .... & f(1,N-1) \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ f(M-1;0) & f(M-1,1) & & f(M-1,N-1) \end{bmatrix}$$

Some of the image (or graphic) formats supported by MATLAB is given in the Table1.1.

Table 1.1 Frequently used the image/graphic formats

Format name	Description	
TIFF	Tagged Image file Format	.tif, .tiff
JPEG	Joint Photographic Expert Group	.jpg, .jpeg
GIF	Graphic Interchange Format	.gif
BMP	Windows Bitmap	.bmp
PNG	Portable Network Graphics	.png

There are many data classes representing pixels in images. The numeric computations in MATLAB are performed using double quantities; hence this is also a frequently used data class used in image processing applications. Class uint8 is also used frequently, especially when reading image data from storage devices. The most frequently used data classes are as follows:

- double — Double-precision floating point numbers in the approximate range -10^{308} *to* 10^{308} (8 bytes per element)
- uint8 — Unsigned 8-bit integers in the range [0 , 255] (1 byte per element)

- uint16 Unsigned 16-bit integers in the range [0 , 62535] (2 bytes per element)
- int8 Signed 8-bit integers in the range [-128 , 127] (1 byte per element)
- int16 Signed 16-bit integers in the range [-32768 , 32767] (2 byte per element)
- int32 Signed 32-bit integers in the range [-214748368 , 2147483647] (4 byte per element)
- single Single-precision floating-point numbers with values in the approximate range -10^{38} *to* 10^{38} (4 bytes per element)
- char Characters (2-byte per element)
- logical Values 0 or 1 (1 byte per element)

The most common image types used in image processing applications are the intensity images, binary images, indexed images and RGB images. However, most gray-level image processing operations are performed using binary or intensity images [40, 41].

- Intensity images: A data matrix whose values are scaled to represent intensities. (i.e., if the elements of an intensity image are of class uint8, it has integer values in the range [0, 255]).
- Binary images: A binary image is a logical array of 0s and 1s.
- Indexed images: An indexed image has two components: a data matrix of integers and a colormap matrix, map. Matrix map is an $mx3$ array of class double containing floating point values in the range [0, 1]. An intensity image maps pixel intensity values to colormap values.

- RGB images: An RGB color image is an $MxNx3$ array of color pixels, where each color pixel consists of red, green and blue components of an RGB image.

1.2.2 2-D Noise Characteristics

The effect of noise is significant in image enhancement applications. Noise in spatial domain is defined by noise mean and noise variance. The general noise types in spatial domain are as follows:

- Gaussian noise: Image is corrupted by Gaussian noise with mean m and variance v.

- Salt & Pepper noise: Image is corrupted by Salt & Pepper noise with density d (a percentage of the image area contains noise values).

- Poisson noise: Poisson noise generated from the data, instead of adding noise to the data.

1.2.3 Literature Summary for Image enhancement

The image enhancement methods can be broadly divided into two groups. The spatial domain methods, based on direct handling of the pixels in an image and frequency domain methods, based on modification of the Fourier transform of an image [42].

A vast literature has emerged recently on image enhancement using linear or nonlinear techniques. The linear techniques, such as low pass filtering, tend to blur edges and destroy important image details. One of the popular nonlinear techniques is the median filtering. In median filtering, a pixel (whether corrupted by noise or not) is replaced by its local median value within a window. Therefore, median filtering not only removes noise but also cause distortion. There is a trade off between noise removal and signal distortion [43].

Another method for noise removal is the nonlinear Wiener filtering. The classical Wiener filtering results in blurring edges in images [44].

Over the past decade, wavelet transform has received considerable attention between the scientists and researchers since it provides good time-frequency localization. The most popular wavelet based denoising techniques are wavelet thresholding or shrinkage techniques proposed by Donoho and Johnstone [24-26], where the noisy image is wavelet decomposed into subbands and noise is removed by removing coefficients that are smaller than some threshold. It is a simple and effective method however the choice of thresholding functions and threshold values are critical in enhancement schemes. The threshold value in universal thresholding [25] depends on the number of data samples. If the number of data samples is too small, the enhanced image is still noisy; on the other hand, if it is very large then the important details of the signal are removed. Similar wavelet-based image enhancement methods have been given in [45-50].

The classical wavelet transforms is based on shifting and scaling of a fixed function and on Fourier transform. An alternative way of constructing wavelets is the lifting scheme. The lifting scheme is a spatial domain method having some advantages over the classical lifting scheme [32, 33]. The advantages of the classical lifting scheme are given in detail in Section 1.1. Some of the lifting based image enhancement methods can be seen in [43, 51, 52].

The adaptation of lifting filters to the dominant signal structure leads to “adaptive lifting”. The motivation behind introducing adaptivity into the lifting steps is that, choosing better lifting filters (prediction or update filters) may lead to more efficient signal representations. Some prominent adaptive lifting algorithms proposed in the last decade are as follows:

R. L. Claypoole at al. [37-38] proposed “scale adaptive” and “space adaptive” update-first lifting algorithms for image enhancement. In the scale adaptive case, prediction filter is adapted to signal structure within each scale by minimizing prediction errors. In the space adaptive case, the prediction filter chosen from a family of prediction filter is adapted to the signal structure for each sample point, based on the “edge avoiding prediction” method.

Jacek Stepien at al. [53] proposed an adaptive lifting based image enhancement method based on scale adaptive lifting scheme proposed in [38]. The subband coefficients are modified by using soft thresholding.

1.3 Objective of the Thesis

The objective of the thesis is to develop a single channel speech enhancement method and an image enhancement method in the adaptive lifting-based wavelet domain. The developed speech and image enhancement methods need to be handled individually with the following features.

- The developed single-channel speech enhancement algorithm should be robust to various adverse environments. A vast number of methods have been proposed in the literature however, we should restrict ourselves to single channel speech enhancement methods based on STFT [5-15] to limit our scope. Choosing MMSE-based single channel speech enhancement methods may be advantageous, since it is known that they cause no musical noise which is a common problem encountered in the subtractive type algorithms. The speech enhancement methods developed so far generally use Gaussian noise for performance evaluation. Use of other noise types (white Gaussian noise (WGN), car interior noise, F16 cockpit noise, and speech babble noise) may result in more realistic performance evaluations. Moreover, the noise power spectral density (PSD) should be estimated directly from the noisy speech signal by aid of a voice activity detector (VAD) since no priori information of noise exists in realistic cases.

- The developed speech enhancement algorithm has to operate in adaptive lifting- based wavelet domain: The wavelet transform is an efficient tool for speech enhancement applications since it allows time-frequency localization. A speech signal can be decomposed into different frequency subbands having different time resolutions by using wavelet transform. Moreover, wavelet packet transform provides a more balanced decomposition tree structure, which allows using subbands of interest. Enhancement of a

subband with certain frequency band may be more advantageous than enhancing original speech signal covering whole frequency band.

- An adaptive lifting scheme algorithm may lead to further improvement since the lifting schemes uses a fixed lifting filter throughout the transform; however the adaptive lifting scheme uses different lifting filters depending on the characteristics of the speech signal. This feature of the adaptive lifting scheme may lead to better handling of the speech signal, so better enhancement results.

- The developed speech enhancement algorithm has to take into account the aspects of human perception (critical bands): The MMSE-based estimators are capable of improving objective quality of speech however; they have no effect on the intelligibility or perceived quality of speech. Hence, a perceptual model taking into account the characteristics of human auditory system may result in improved intelligibility or perceived quality of speech.

- The algorithms developed for speech enhancement have to be tested by using objective evaluation tests. Using objective quality evaluation tests which are well correlated to subjective results (such as Segmental SNR, Itakura-Saito distance). The objective tests (i.e. SNR test) are not always meaningful since the speech signals having different perceived qualities may have the same SNR results. Therefore, the objective evaluation results should be verified by using subjective listening tests or alternatively using standard objective tests, such as PESQ-MOS, which is developed for predicting subjective MOS results since the subjective tests cost much time and effort and many listeners.

- The developed image enhancement method is to be based on spatial domain filtering methods and wavelet thresholding-based methods, all in the adaptive lifting based wavelet domain: The gray-level digital images (double or uint8) corrupted by independent white Gaussian noise needs to be used in the objective evaluations.

- The algorithms developed for image enhancement algorithms have to be tested by using standard evaluation tests.
- The experimental results (i.e. signal waveforms, spectrograms and enhanced images) should be presented.

1.4 Organization of the Thesis

The organization of the thesis is as follows:

In Chapter 1, the problems encountered in speech and image enhancement and the importance of speech and image enhancement are outlined. General information on the characteristics of noise, speech and image signals and a brief literature summary of the speech and image enhancement is also given. The objective and organization of the thesis is also given in Chapter 1.

Chapter 2 includes a brief literature survey and mathematical background of wavelets, wavelet transforms and lifting schemes which are basic tools of the thesis.

Chapter 3 is basically devoted to speech enhancement. The literature summary and theoretical background of popular single channel speech enhancement methods (estimators) have been given in detail in this chapter. The procedure of the proposed adaptive lifting scheme, perceptual filterbank and overall speech enhancement method are also given in Chapter 3.

In Chapter 4, the theoretical backgrounds of popular image enhancement methods (spatial domain and wavelet thresholding-based methods) are given in detail. The procedure of the proposed adaptive 2-D lifting scheme and proposed image enhancement method are also given in Chapter 4.

Chapter 5 includes the performance evaluation results of the speech and image enhancement methods proposed in the thesis. The objective and subjective evaluation methods are also outlined. Experimental results for both image and speech enhancement algorithms have also been demonstrated in Chapter 5.

Chapter 6 includes conclusions on the results of the proposed image and speech enhancement methods, main contributions to the subject and suggestions for the future work.

CHAPTER 2

WAVELET TRANSFORMS AND LIFTING SCHEMES

2.1 Introduction

This chapter of the thesis includes a brief literature survey and mathematical background of wavelets, wavelet transforms and lifting schemes.

J. Morlet [21, 57] was the first scientist who introduced the concept of wavelets. He faced with the problem of analyzing signals which had very high frequency components with short time durations, and low frequency components with long time durations. STFT was able to analyze either high frequency components using narrow windows (wideband frequency analysis), or low frequency components using wide windows (narrowband frequency analysis), but not both. He therefore, suggested the idea of using a different window function for analyzing different frequency bands. Furthermore, these windows were all generated by dilation or compression of a prototype Gaussian window. These window functions had compact support both in time and in frequency. Due to the "small and oscillatory" nature of these window functions, Morlet named his basis functions as *wavelets of constant shape*.

Y. Meyer noticed that there was a great deal of redundancy in Morlet's choice of basis functions (wavelets) [21, 57]. Meyer developed wavelets with better localization properties and constructed *orthogonal wavelet* basis functions with very good time and frequency localization. Haar wavelets are the first and the simplest *orthonormal wavelets*; however, they are of little practical use due to their poor frequency localization.

Ingrid Daubechies [57] developed the *wavelet frames* for discretization of time and scale variables of the wavelet transform, which allowed more freedom in the choice of basis functions at an expense of some redundancy. Furthermore, she made contributions on developing the discrete wavelet transforms (DWT).

S. Mallat [58] developed the idea of *multiresolution analysis* (MRA) for discrete wavelet transform (DWT). The main idea was decomposing a discrete signal into its dyadic frequency subbands by a series of low pass and high pass filters to compute its DWT. This new idea was known by electrical engineers for a long time as the quadrature mirror filters (QMF) and subband filtering. Mallat's work led to the good frequency localization idea of QMF and subband coding.

Albert Cohen, Jean Feauveau and Daubechies [64] constructed the *compactly supported biorthogonal wavelets* (CDF family of biorthogonal wavelets), which are preferred by many researchers over the orthonormal basis functions.

R. Coifman, Meyer and Victor Wickerhauser developed wavelet packets, a natural extension of MRA [59, 60]. The wavelet packet system was designed to allow a finer and adjustable frequency resolution at high frequencies. It gives a richer structure which allows adaptation to particular signals.

Wim Sweldens [32, 33] developed an alternative method called as the *lifting scheme* or *second generation wavelets* for construction of biorthogonal wavelets. The main idea was building complicated biorthogonal systems using simple an invertible stages *split*, *predict* and *update*. The new method has the following advantages over the classical wavelet transform: It is a spatial domain method, easier to implement, allows faster and in-place calculations, allows nonlinear, adaptive, irregularly sampled and integer to integer wavelet transforms and inverse transform is easier to obtain. Furthermore, any wavelet transform can be factored into lifting stages [34]. The rest of this chapter will be mainly on CWT, DWT, MRA, filterbanks, wavelet bases and wavelet frames, orthogonal and biorthogonal wavelets, 1-D and 2-D DWT, DWPT, undecimated wavelet transform and lifting schemes.

2.2 Continuous Wavelet Transform (CWT)

The continuous wavelet transform [54-57] was developed because of the above stated limitations of the STFT. The STFT which allowed analysis of non-stationary signals by segmenting them into stationary enough short frames and computing the Fourier transform of each frame.

$$S(\tau, f) = \int w^*(t-\tau)x(t)e^{-j2\pi ft}\,dt$$
$$x(t) = \int_{\tau}\int_{f} S(\tau, f)w^*(t-\tau)e^{j2\pi ft}\,d\tau df \qquad (2.1)$$

Where $w(t)$ is the windowing function, f and τ are frequency and shifting (time) parameters respectively and * is the complex conjugate operator and $S(\tau, f)$ is the STFT of $x(t)$ at frequency f and shifting τ. For each frequency f, time localization is obtained through windowing $x(t)$ by $w(t-\tau)$, the windowing function centered at $t = \tau$. The Fourier transform of this segmented signal provides the frequency localization [21].

The main drawback of STFT is that it provides constant resolution for all frequencies since it uses the same window for the analysis of the whole signal. If the signal has high frequency components for a short-time interval, a narrow window (compactly supported in time) would be enough for good time resolution. However, narrow windows mean wider frequency bands, resulting in poor frequency resolution. On the other hand, if the signal contains low frequency components of longer time interval, than a wider window is needed to obtain good frequency resolution (at the expense of time resolution). This is the motivation behind the wavelet transform (WT), which provides varying time and frequency resolutions by using variable window lengths [57].

The uncertainty principle prevents the possibility of having arbitrarily high resolution in both time and frequency, since it lower limits the time-bandwidth product of possible basis functions by $\Delta T.\Delta\Omega \geq (1/4\pi)$ where ΔT and $\Delta\Omega$ are absolute values of function and its Fourier transform respectively. However, there is a trade off between time and frequency resolutions in case of using a variable window size [20].
The continuous wavelet transform of a square-integrable function $x(t) \in L^2(R)$ is given as

$$W(a,b) = \int_{-\infty}^{\infty} \psi^*_{a,b}(t)x(t)dt \quad \text{for} \quad a \in R^+,\ b \in R\ , \qquad (2.2)$$

where

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right). \tag{2.3}$$

The *mother wavelet* $\psi(t)$ is used to derive all the basis functions $\psi_{ab}(t)$ which are sometimes called as *daughter wavelets*. The shifting (or translation) parameter b refers to the location of the wavelet function, as it is shifted through the signal. Hence, it corresponds to the time information in the wavelet transform. The scaling (or dilation) parameter a refers to the scale (1/frequency) and corresponds to frequency information. The large scales expand the signal and provide the detailed low frequency information (transients or peaks) in the signal. On the other hand, small scales contract the signal and provide global high frequency information in the signal. Accordingly, the high frequencies appear as short bursts while low frequencies last throughout the entire signal.

Fourier transform of the mother wavelet and the daughter wavelets can be given as

$$\psi(t) \Leftrightarrow \Psi(\omega), \tag{2.4}$$

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \Leftrightarrow \sqrt{a}\Psi(a\omega)e^{-jb\omega}. \tag{2.5}$$

The scaling operation shows that an increase in resolution in one domain results in loss of resolution in the other domain. This actually reflects the trade-off that exists between time and frequency domain resolutions. The daughter wavelets $\psi_{a,b}(t)$ (wavelet basis functions) are used in decomposing the signal $x(t)$ and the CWT coefficients $W(a,b)$ represent the projections of the signal on these bases.

The inverse CWT (or reconstruction formula) can be defined as

$$x(t) = \frac{1}{C_\psi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{W(a,b)\psi_{a,b}(t)}{a^2}dadb, \tag{2.6}$$

where

$$C_{\psi} = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty . \tag{2.7}$$

In order for $\psi(t)$ be *admissible* for $C_{\psi} < \infty$, it requires that $\Psi(0) = 0$ and $W(\omega)$ goes to zero $(W(\infty) = 0)$ fast enough for $C_{\psi} < \infty$. This means that $\psi(t)$ is a function with zero-mean and finite energy in the time domain. Moreover, for (2.6) to be satisfied, a wavelet is constructed so that it has *vanishing moments* [61] of order *m* if

$$\int_{-\infty}^{\infty} t^p \psi(t) dt = 0, \qquad p = 0,1,2.....m-1. \tag{2.8}$$

2.3 Discrete Wavelet Transform (DWT)

The reconstruction equation given by (2.6) involves a redundant set of basis functions. A more convenient representation can be obtained by discretizing the shifting and scaling parameters a and b where only the required wavelet coefficients for the reconstruction of $x(t)$ are kept. The new representation is known as the Discrete Wavelet Transform (DWT) [19, 54-56, 58] with the wavelet basis function

$$\psi_{j,k}(t) = a_0^{j/2} \psi(a_0^j t - k b_0), \tag{2.9}$$

where

$$j,k \in \mathrm{Z}, a_0 > 1, \; b_0 \neq 0, \tag{2.10}$$

which correspond to $a = a_0^{-j}$ and $b = k b_0 a_0^{-j}$. Note here that the shifting step depends on the scaling; since long wavelets shift by large steps while short ones by small steps. Thus, discrete wavelet transform becomes

$$W(j,k) = a_0^{j/2} \int_{-\infty}^{\infty} \psi(a_0^j t - k b_0) x(t) dt \tag{2.11}$$

where a_0 and b_0 represents arbitrary reference scale and time values respectively and j,k are the new scaling and shifting parameters respectively. Particularly, choosing $a_0 = 2$ and $b_0 = 1$ for dyadic grids, leads to the fast wavelet transform (FWT). Thus, wavelet basis function $\psi_{j,k}(t)$ becomes

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \tag{2.12}$$

where, $a = 2^{-j}$ and $b = k2^{-j}$. The forward and inverse DWT can be defined as

$$W(j,k) = \int_{-\infty}^{\infty} \psi_{j,k}(t)x(t)dt = < x(t), \psi_{j,k}(t) >, \tag{2.13}$$

$$x(t) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} W(j,k)\psi_{j,k}(t), \tag{2.14}$$

where $j,k \in Z$ and $< . , . >$ in (2.13) represents the inner product between two functions. The basis function $\psi_{j,k}(t)$ in equation (2.9) now provides an orthonormal basis that is no longer redundant. However, equation (2.14) still requires an infinite number of terms to describe the infinitely coarse, i.e. $j \to -\infty$ as well as the infinitely fine, $j \to \infty$.

2.4 Concept of Multiresolution

The concept of multiresolution analysis [54-56] is used to construct orthonormal bases of wavelets. This multiresolution view can be interpreted as a successive approximation procedure.

The multiresolution formulation is obviously designed to represent signals where a single is decomposed into finer and finer details. However, it is also valuable in representing signals where a time-frequency or time-scale description is desired even if no concept of resolution is needed.

In order to talk about the collection of functions or signals that can be represented by a sum of scaling or wavelet functions some ideas and terminology is needed from functional analysis. Some of these terminologies will be summarized here.

A *function space* is a linear vector space (finite or infinite dimensional) where the vectors are the functions and the scalars are the real numbers (or sometimes complex numbers). The *inner product* is a scalar "*a*" obtained *from* two vectors $f(t)$ and $g(t)$ by an integral as given below.

$$a =< f(t), g(t) > = \int f^*(t)g(t)dt \tag{2.15}$$

The inner product is defines a *norm* or *length* of vectors which is defined by

$$\|f\| = \sqrt{|< f, f >|} \tag{2.16}$$

Two signals (vectors) with nonzero norms are called *orthogonal* if their inner product is zero. A space which is particularly important in signal processing is $L^2(R)$ which is the space of the $f(t)$ with a well defined integral of the square of the modulus of the function. The *L* here defines Lebesque integral; the *2* denotes the integral of the square of the modulus of the function and *R* states that the independent variable of integration *t* is a number over the whole real axis.

In order to develop the wavelet expansion described in (2.10) an idea of expansion set or a basis set is needed. Let *S* is given as the vector space, if any $f(t) \in S$ can be expressed as $f(t) = \sum_k a_k \phi_k(t)$, then the set of functions $\phi_k(t)$ are called as an expansion set for the space *S*. If the representation is unique the set is a basis. On the other hand, one can start with the expansion or bases set and define the space *S* as the set of all functions that can be expressed by $f(t) = \sum_k a_k \phi_k(t)$. This is called as the *span* of the basis set.

2.4.1 The Scaling Function

In order to use the idea of multiresolution we will define the scaling function and then define the wavelet in terms of it. Let us describe a set of scaling functions in terms of integer shifts of the basic scaling function by

$$\phi_k(t) = \phi(t-k) \quad \text{for} \quad k \in Z, \quad \phi \in L^2. \tag{2.17}$$

The subspace of $L^2(R)$ spanned by these functions can be defined as

$$V_0 = \overline{\underset{k}{span}\{\phi_k(t)\}} \tag{2.18}$$

for all the integers k from minus to plus infinity. The over bar in (2.18) denotes the closure, which means that

$$f(t) = \sum_k a_k \phi_k(t) \quad \text{for any} \quad f(t) \in V_0. \tag{2.19}$$

The size of the subspace spanned can be changed by changing the time scale of the scaling functions. Accordingly, a two dimensional set of functions is generated from the basic scaling function by shifting and translation by

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k). \tag{2.20}$$

2.4.2 Multiresolution Analysis (MRA)

The basic requirements of multiresolution analysis can be given by requiring a nesting of spanned spaces as

$$\ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots \quad \subset L^2 \tag{2.21}$$

or,

$$V_j \subset V_{j+1} \quad , \quad m \in Z \quad \text{for} \quad V_{-\infty} = \{0\}, \quad V_{\infty} = L^2. \tag{2.22}$$

The space which contains high resolution will also contain those of lower resolutions. Because of the definition of V_j, the spaces have to satisfy a natural scaling condition;

$$f(t) \in V_j \quad \Leftrightarrow \quad f(2t) \in V_{j+1}. \tag{2.23}$$

This insures that the elements in a space are simply the scaled versions of the elements in the next space. From (2.21) to (2.23), if $\phi(t)$ is in V_0, it is also in V_1, the space spanned by $\phi(2t)$. This means that $\phi(t)$ can be expressed in terms of a weighted sum of shifted $\phi(2t)$ as given below

$$\phi(t) = \sum_n h(n)\sqrt{2}\phi(2t-n), \quad n \in Z \tag{2.24}$$

where the coefficients $h(n)$ are a sequence of real or perhaps complex numbers called scaling function coefficients (or the scaling filter or vector) and the $\sqrt{2}$ is the norm of the scaling function of scale 2. The (2.24) is referred to by different names, some of which are the refinement equation, multiresolution equation, or scaling equation, to describe different interpretations or different points of view.

For example, the Haar scaling function is the simple, unit-width, unit-height pulse function $\phi(t)$ shown in (2.25) and it is clear that $\phi(t)$ can be constructed by using $\phi(2t)$.

$$\phi(t) = \phi(2t) - \phi(2t-1) \tag{2.25}$$

The (2.25) is satisfied for the coefficients $h(0) = 1/\sqrt{2}$ and $h(1) = 1/\sqrt{2}$.

2.4.3 Multiresolution Characteristics of DWT

The important characteristics of a signal can better be described by defining a slightly different set of functions $\psi_{j,k}(t)$ that span the differences between the

spaces spanned by the various scales of the scaling function. These functions are called the wavelets as discussed before.

There are many advantages to requiring that the scaling functions and wavelets be orthogonal since the orthogonal basis functions allow simple calculations of expansion coefficients. The orthogonal complement of V_j and V_{j+1} is defined as W_j. This means that all members of V_j are orthogonal to all members of W_j. We get

$$< \phi_{j,k}(t), \psi_{j,l}(t) > = \int \phi_{j,k}(t) \psi_{j,l}(t) dt = 0 \tag{2.26}$$

for all $j,k,l \in Z$. The following expression represents the relationship between various subspaces. Starting from W_j when $j = 0$ we have

$$.V_0 \subset V_1 \subset V_2 \subset \ldots \quad \subset L^2 . \tag{2.27}$$

The wavelet spanned by the subspace W_0 can be defined as

$$V_1 = V_0 \oplus W_0 . \tag{2.28}$$

This gives in general,

$$L^2 = V_0 \oplus W_0 \oplus W_1 \oplus \ldots \tag{2.29}$$

The scale of the scaling function can be chosen arbitrarily. In practice, it is chosen to represent the coarsest detail of interest in a signal.

Since the wavelets reside in the space spanned by the next narrower scaling function, $W_0 \in V_1$, they can be represented by a weighted sum of shifted scaling function $\phi(2t)$ as defined in (2.30) as:

$$\psi(t) = \sum_n h_1(n) \sqrt{2} \phi(2t - n), \quad n \in Z \tag{2.30}$$

for some set of coefficients $h_1(n)$. From the requirement that the wavelets span the "difference" or orthogonal complement spaces and the orthogonality of integer translates of the wavelet (or scaling) function. The wavelet coefficients are required by orthogonality to be related to the scaling function coefficients by $h_1(n) = (-1)^n h(1-n)$.

For $h(n)$ with finite even length-N $h_1(n) = (-1)^n h(N-1-n)$. The function generated by (2.24) gives the prototype or mother wavelet $\psi(t)$ for a class of expansion coefficients of the form $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$.

For example the Haar wavelet function is given in Figure (2.1) where, $\psi(t) = \phi(2t) - \phi(2t-1)$ and the wavelet coefficients are $h(0) = 1/\sqrt{2}$ and $h(1) = -1/\sqrt{2}$.

We have now constructed a set of functions $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ that span all of $L^2(R)$ according to (2.29). For lower and upper resolution limits for scaling indexes are $j = 0$ and $j = L$, any function $g(t) \in L^2(R)$ can be written as

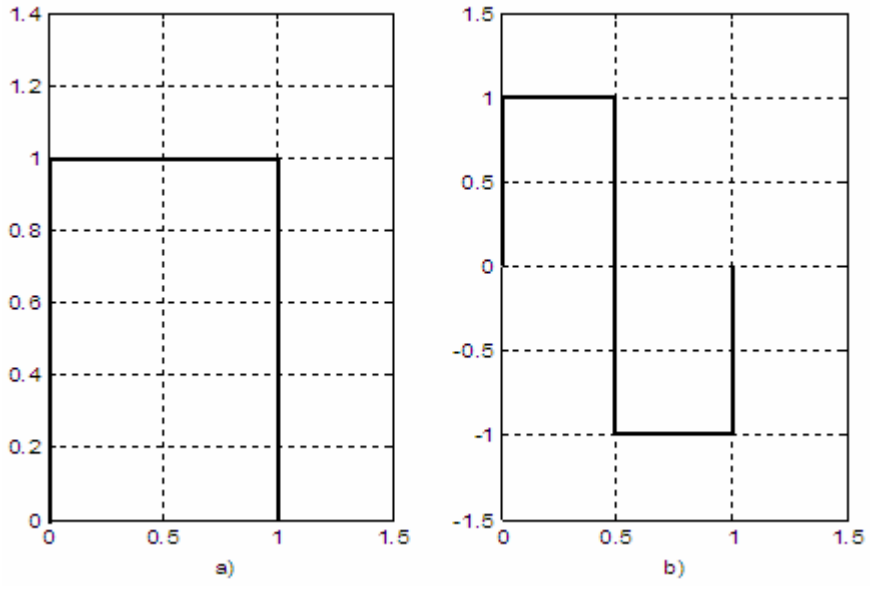

Figure 2.1 Haar wavelet: a) Haar scaling function b) Haar wavelet function

$$g(t) = \sum_k c(0,k)\phi_{0,k}(t) + \sum_{j=0}^{L-1}\sum_n d(j,k)\psi_{j,k}(t) \tag{2.31}$$

where,

$$c(j,k) = c(0,k) =< g(t)\phi_{j,k}(t) >,$$
$$d(j,k) =< g(t)\psi_{j,k}(t) >. \tag{2.32}$$

The coefficients $c(j,k)$ and $d(j,k)$ are referred to as *approximation coefficients* and *detail coefficients* respectively.

2.5 DWT and Filterbank Representations

The scaling and wavelet functions are not needed to be dealt with directly in many applications. Only the coefficients $h(n)$ and $h_1(n)$ in (2.24) and (2.30) and $c(j,k)$, $d(j,k)$ in (2.32) need to be considered and they can be viewed as digital filters and digital signals respectively [20,44,58]. Although, it is possible to develop most of the results of wavelet theory using only filterbanks, we think that both the wavelet expansion and filterbank point of view are necessary for better understanding of this new concept.

2.5.1 Analysis: From Fine to Coarse Scale

Let us drive the relationship between the expansion coefficients at lower scale level in terms of those at higher scale levels in order to be able to work directly with the wavelet transform coefficients. From the equation (2.24) after scaling and translating (shifting) time variable, we have

$$\phi(2^j t - k) = \sum_n h(m-2k)\sqrt{2}\phi(2(2^j t - k) - n) = \sum_n h(n)\sqrt{2}\phi(2^{j+1}t - 2k - n) \tag{2.33}$$

Substituting $m = 2k + n$, we have

$$\phi(2^j t - k) = \sum_n h(m-2k)\sqrt{2}\phi(2^{j+1}t - m). \tag{2.34}$$

If V_j is defined as

$$V_j = \overline{\underset{k}{span}}\left\{2^{j/2}\phi(2^j t-k)\right\}, \tag{2.35}$$

then,

$$f(t)\in V_{j+1} \Rightarrow f(t)=\sum_k c_{j+1}(k)2^{(j+1)/2}\phi(2^{j+1}t-k) \tag{2.36}$$

can be given at a scale of $j+1$ with scaling functions only and no wavelets. At one scale lower resolution, wavelets are necessary for the details not available at scale j.

$$f(t)=\sum_k c_j(k)2^{j/2}\phi(2^j t-k)+\sum_k d_j(k)2^{j/2}\psi(2^j t-k) \tag{2.37}$$

Where, the term $2^{j/2}$ represents the unity norm of the basis functions at various scales. If $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ are orthonormal or tight frame, the level j scaling coefficients can be found by using inner product.

$$c_j(k) =< f(t),\phi_{j,k}(t) >= \int f(t)2^{j/2}\phi(2^{j/2}t-k)dt \tag{2.38}$$

By using (2.34) and interchanging sum and integral we have

$$c_j(k)=\sum_k h(m-2k)\int f(t)2^{(j+1)/2}\phi(2^{(j+1)/2}t-m)dt\,. \tag{2.39}$$

Since the integral part is equal to $c_{j+1}(m)$, we obtain

$$c_j(k)=\sum_m h(m-2k)c_{j+1}(m)\,. \tag{2.40}$$

The wavelet coefficients can be obtained in the similar manner

$$d_j(k)=\sum_m h_1(m-2k)c_{j+1}(m)\,. \tag{2.41}$$

2.5.2 Filtering and Down-Sampling

In the digital signal processing, the filtering of a digital signal is achieved by convolving the sequence with filter coefficients, weights or impulse response. For input sequence is $x(n)$ and filter coefficients is $h(n)$, the output sequence $y(n)$ can be represented as

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k). \tag{2.42}$$

If the number of filter coefficients N is finite, the filter is called as *Finite Impulse Response* (FIR) filter, if the number is infinite; it is called as *Infinite Impulse Response* (IIR) filter. The design problem is the selection of $h(n)$ to obtain some desired effects, removal of noise or some separate signals.

What we deduce from (2.40) and (2.41) is filtering and down-sampling. It is seen from these equations that the scaling and wavelet coefficients at scale j are filtered by two FIR filters $h(-n)$ and $h_1(-n)$ then down-sampling or decimating to give the expansion coefficients at the next coarser scale $j-1$.

These structures implement Mallat's algorithm [58] and have been developed in the engineering literature on filterbanks, Quadrature Mirror Filters (QMF), conjugate filters and perfect reconstruction (PR) filters. The perfect reconstruction criteria is given in Appendix A2. A two-band two stage analysis filterbank tree is given in Figure 2.2.

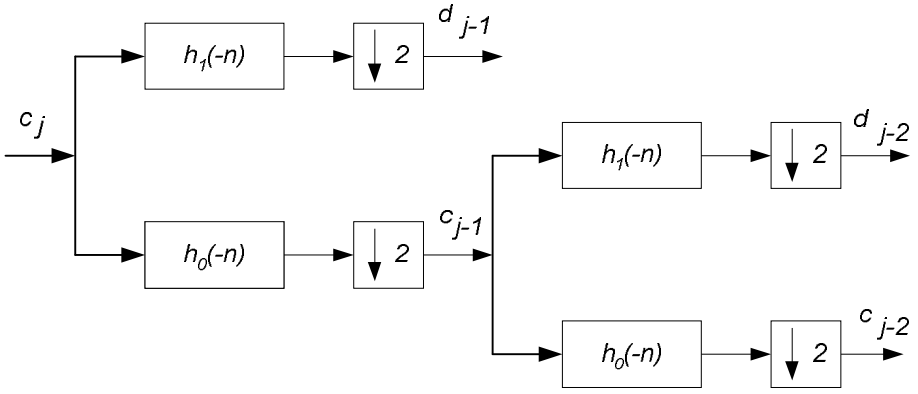

Figure 2.2 Two-band, two-level wavelet analysis (decomposition) tree

Note that the FIR filter implemented by $h(-n)$ or ($h_o(-n)$) is low-pass filter while the one implemented by $h_1(-n)$ is high-pass filter. The number of data points is doubled by having two filters then it is halved after down-sampling (or decimation). Hence, the average number of data points before and after the system is the same. This means that there is no information lost and it is possible to completely recover the original signal. This is the basic idea behind the perfect reconstruction filterbanks. The filtering and decimation process can be continued on the scaling coefficients to obtain the tree structures with two, three or more scale levels. The decomposition tree structure obtained is called as constant-Q filterbank or octave-band filterbanks in the filterbank terminology.

2.5.3 Synthesis: From Coarse to Fine Scale

As stated before, fine scale coefficients of the original signal can be reconstructed from the combination of the scale coefficients and the detail coefficients at a coarser scale. To derive this, let us consider a signal in the $j+1$ scaling function space $f(t) \in V_{j+1}$,

$$f(t) = \sum_k c_{j+1}(k) 2^{(j+1)/2} \phi(2^{j+1}t - k). \tag{2.43}$$

In terms of the next scale which also requires the wavelets

$$f(t) = \sum_k c_j(k) 2^{j/2} \phi(2^j t - k) + \sum_k d_j(k) 2^{j/2} \psi(2^{j/2} t - k). \tag{2.44}$$

Substituting (2.24) and (2.30) into (2.44) we have

$$\begin{aligned} f(t) = &\sum_k c_j(k) \sum_n h(n) 2^{(j+1)/2} \phi(2^{(j+1)/2} t - 2k - n) + \ldots \\ &+ \sum_k d_j(k) \sum_n h_1(n) 2^{(j+1)/2} \phi(2^{(j+1)/2} t - 2k - n) \end{aligned} \tag{2.45}$$

Since all these functions are orthonormal, multiplying (2.43) and (2.45) by $\phi(2^{j+1}t - k)$ and integrating we obtain the original signal coefficients as,

$$c_{j+1}(k) = \sum_m h(k-2m)c_j(m) + \sum_m h_1(k-2m)d_j(m). \quad (2.46)$$

2.5.4 Up-sampling and Filtering

The synthesis or reconstruction stage in the filterbank includes up-sampling followed by filtering. This is what (2.46) exactly does. The equation is evaluated by up-sampling the j level scale coefficient sequence $c_{j-1}(k)$ then filtered by $h(n)$. The j level detail coefficient sequence $d_{j-1}(k)$ is up-sampled and filtered by $h_1(n)$ in the same way. The results are added to give the j level scaling coefficients. The operation is repeated for j level scaling and detail coefficients to obtain the $j+1$ level scaling coefficients (original signal). A two stages two band synthesis filterbank tree is given in Figure 2.3.

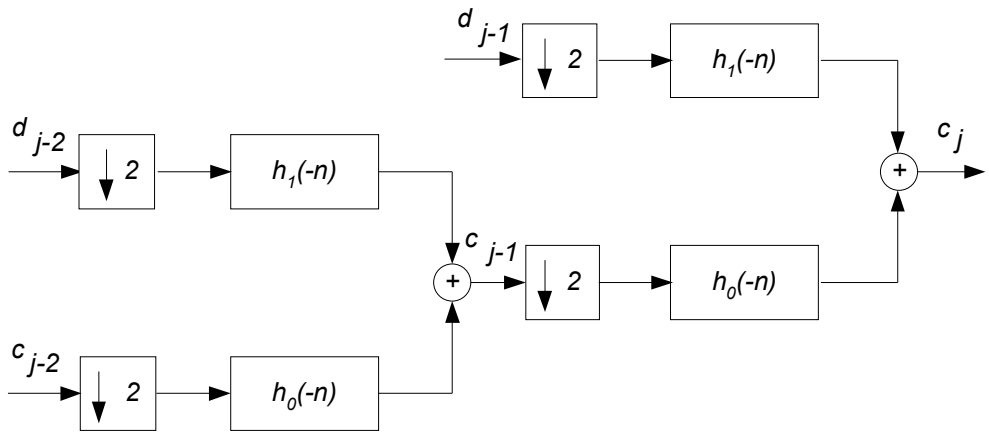

Figure 2.3 Two-level two-band wavelet synthesis (reconstruction) tree

2.6 Wavelet Packet Transform

The wavelet packet system was proposed by Ronald Coifman [59] to allow a finer and adjustable frequency resolution at high frequencies. It gives a richer structure which allows adaptation to particular signals [54, 60].

The standard DWT involves a dyadic (2-Band) tree structure where only the low-channel side is split down to a certain level. However, in the wavelet packet decomposition, each detail coefficient vector is also decomposed into two parts using the same approach as in approximation coefficients. This offers a richer analysis and

a more balanced decomposition tree structure. The full wavelet packet decomposition tree is given in Figure 2.4.

Note here that a signal may be represented by a selected numbers of subbands without using every subband for a given resolution level. An algorithm can be constructed to choose the subbands for an optimization criterion (such as energy, entropy, variance etc.). Best-basis and best-level algorithms are the most popular algorithms for signal representation [55, 59-60].

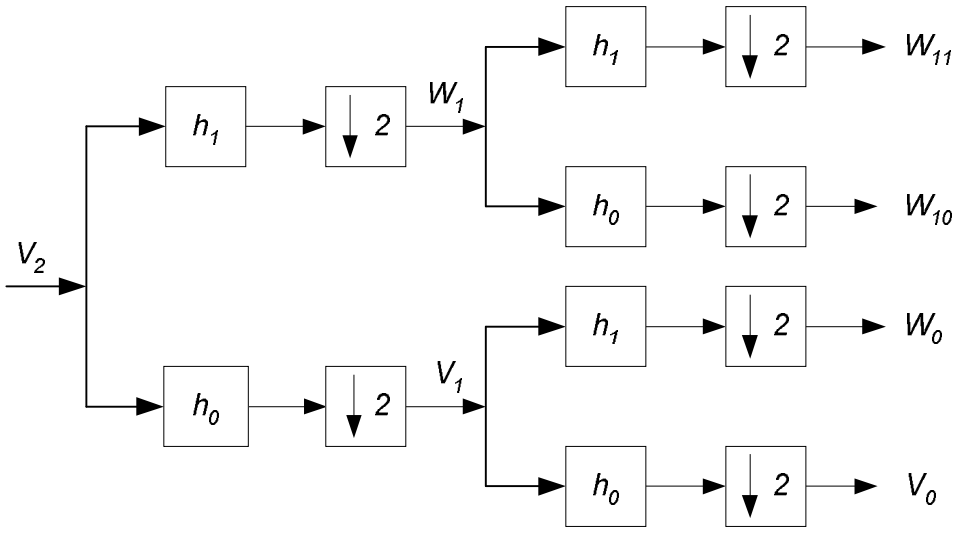

Figure 2.4 Wavelet packet decomposition tree at level 2

2.7 Orthogonal and Biorthogonal Bases and Frame

2.7.1 Orthogonal and Biorthogonal Bases

A set of finite or infinite functions $f_k(t)$ *spans* a vector space F if any element of that space can be expressed as a linear combination of members of that set. We define $\underset{k}{span}\{f_k\} = F$ as the vector space with elements of the space of the form:

$$g(t) = \sum_k a_k f_k(t) \tag{2.47}$$

with $k \in Z$ and $t, a \in R$. An *inner product* is usually defined for this space and is denoted $< f(t), g(t) >$. A *norm* is defined as $\|f\| = \sqrt{< f, f >}$. We say that the set

$f_k(t)$ is a *basis set* or *basis* for a given space F if the set of $\{a_k\}$ in (2.47) are unique for any particular $g(t) \in F$. The set is called an *orthogonal basis* if $< f_k(t), f_l(t) >= 0$ for all $k \neq l$. In three dimensional Euclidean spaces, orthogonal basis vectors are coordinate vectors that are at right (90^o) angles to each other. The set is sad to be an orthonormal basis if $< f_k(t), f_l(t) >= \delta(k-l)$ if in addition to being orthogonal, the basis vectors are normalized to unity i.e. $\|f_k(t)\| = 1$ for all *k*.
It is clear from these definitions that, if we have an orthonormal basis, we can express any element in the vector space, $g(t) \in F$ as in (2.47),

$$g(t) = \sum_k < g(t), f_k(t) > f_k(t) . \tag{2.48}$$

Since by taking inner the product of $f_k(t)$ with both sides of (2.48) we have

$$a_k =< g(t), f_k(t) > \tag{2.49}$$

Although the orthonormal bases are advantageous in many applications, there are cases where the basis system dictated by the problem is not and cannot be made orthogonal. In such cases one should use *dual basis set* $\tilde{f}_k(t)$ whose elements are not orthogonal $\tilde{f}_k(t)$ to each other, but to the corresponding elements of the expansion set is

$$< f_l(t), \tilde{f}_k(t) >= \delta(l-k) . \tag{2.50}$$

Since this type of orthogonality requires two sets of vectors, the expansion set and the dual set, the system is called as *biorthogonal.* From (2.47) and (2.50) we have

$$g(t) = \sum_k < g(t), \tilde{f}_k(t) > f_k(t) . \tag{2.51}$$

The calculation of the expansion coefficients using inner product in (2.49) is called as the *analysis* part while the calculation of the signal from the coefficients and expansion vectors in (2.47) is called as *synthesis* part.

The analysis and synthesis operations are matrix-vector multiplications in finite dimensions. If the expansion vectors in (2.47) are a basis, the synthesis matrix has these basis vectors as columns and the matrix is a square and non-singular. If the matrix is orthogonal, its rows and columns are orthogonal; its inverse is its transpose and the identity operator is the matrix multiplied by its transpose. If it is not orthogonal, then the identity is the matrix multiplied by its inverse and the dual basis consists of the rows of the inverse. If the matrix is singular, then its columns are not independent and, therefore, do not form a basis [54-56].

2.7.2 Wavelet Frames

Although, the conditions for a set function being an orthonormal basis are sufficient for the representation in (2.48) and the requirement of the set being a basis is sufficient for (2.51), they are not necessary. To be a basis, coefficients are required to be unique. In other words, the set is required to be independent and no element can be written as the linear combination of the others [54-56, 57].

If the set of functions or vectors is dependent and allow the expansion described in (2.51), then the set is called as a *frame*. A frame is a spanning set which requires finite limits of an inequality bound of inner products.

An expansion set $\varphi_k(t)$ must satisfy the following equation in order to be a frame in a signal space

$$A\|g\|^2 \le \sum_k |< \varphi_k, g >|^2 \le B\|g\|^2 \tag{2.52}$$

for some $0 < A$ and $B < \infty$, for all signals $g(t)$ in the space. By dividing (2.52) by $\|g\|^2$ shows that, A and B are bounds on the normalized energy of the inner product. They frame the normalized coefficient energy. If $A=B$, then the expansion set is called to be a *tight frame* which gives

$$A\|g\|^2 = \sum_k |< \varphi_k, g >|^2 . \tag{2.53}$$

The (2.54) is the generalized Parseval's theorem for tight frames. For $A=B=1$, the tight frame becomes an orthonormal basis. It can be shown for a tight frame that

$$g(t) = A^{-1} \sum_{k} < \phi_k(t), g(t) > \phi_k(t) . \tag{2.54}$$

This is the same as the expansion using an orthonormal basis except for the A^{-1} term which represents a quantity for redundancy for the expansion set.

2.8 Undecimated Wavelet Transform

It is known from the basic laws of wavelets in the wavelet theory that, the DWT is not *shift-invariant* [54, 55].

A typical wavelet based signal processing framework consists of the following steps. Wavelet transform, point processing of wavelet coefficients and inverse wavelet transform. As the wavelet transform is not shift-invariant, if the signal is shifted and processed as through the above explained steps and shift the output, the results are different for different shifts.

A method to create a linear-shift invariant DWT can be achieved by constructing a frame from the orthogonal DWT added by shifted orthogonal DWT. Doing this, the result is still a frame but because of the redundancy, it is called as *redundant discrete wavelet transform* (RDWT).

Another way to construct the RDWT is the wavelet filterbank approach. The wavelet filterbanks can be modified by removing the decimators (down samplers) between each stage to give the coefficients of the tight frame expansion (the RDWT) of the signal. The new structure is called as *undecimated filterbank*. The undecimated DWT is shift-invariant; less effected by noise, quantization and error and has $N \log(N)$ computational complexity.

The general procedure for undecimated wavelet transform (UWT) (sometimes called as *stationary* wavelets transform (SWT)) structure is given below in Figure 2.5.

Given a signal c_j of length N, the first step of the SWT produces approximation coefficients c_{j-1} and detail coefficients d_{j-1}. These vectors are obtained by convolving c_j with the low-pass filter $Lo\_(j)$ and with the high-pass

filter $Hi\_(j)$. Note that c_{j-1} and d_{j-1} are of length N instead of N/2 as in the DWT case.

The next step splits the approximation coefficients c_{j-1} in two parts using the same scheme, but with modified filters obtained by up-sampling the filters used for the previous step and replacing c_j by c_{j-1}. Then, the SWT produces c_{j-2} and d_{j-2}. The decomposition can further be continued in the same way.

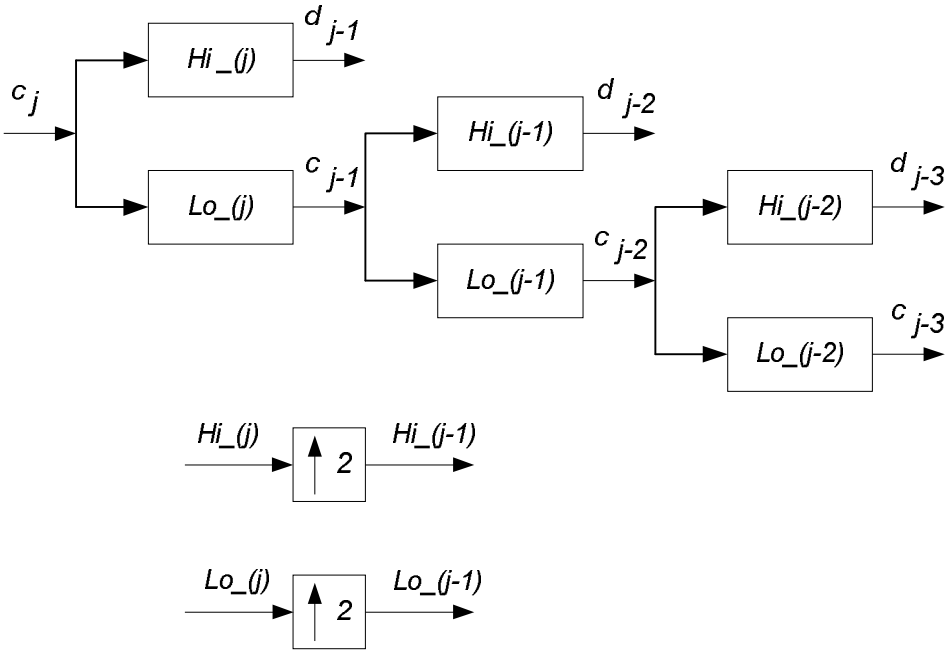

Figure 2.5 Undecimated wavelets transform

2.9 2-D Wavelet and Wavelet Packet Transform

When the input signal is two dimensional (2-D) the signal must be represented by 2-D wavelet and 2-D scaling functions. For a given set of wavelet scaling function (ψ, ϕ) one 2-D scaling function and tree different 2-D wavelet can be constructed. The 2-D scaling function is

$$\phi_{i,j}(x,y) = \phi(x-i)\phi(y-j), \tag{2.55}$$

and the 2-D wavelet functions are

$$\psi_{i,j}^{[1]}(x,y) = \phi(x-i)\psi(y-j),$$
$$\psi_{i,j}^{[2]}(x,y) = \psi(x-i)\phi(y-j),$$
$$\psi_{i,j}^{[1]}(x,y) = \psi(x-i)\psi(y-j). \quad (2.56)$$

The $\psi_{i,j}^{[1]}(x,y)$, $\psi_{i,j}^{[2]}(x,y)$ and $\psi_{i,j}^{[3]}(x,y)$ are horizontal, vertical and diagonal wavelet functions respectively. Hence, the wavelet functions satisfy

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\psi_{i,j}^{[M]}(x,y)dxdy = 0 \text{ for M=1, 2, 3.} \quad (2.57)$$

The spectral subbands that are obtained from one level wavelet decomposition are labeled as LL, LH, HL, and HH. Where, LL denotes (low pass-low pass) approximation coefficients and LH (low-high), HL (high-high), HH (high-high) denote horizontal, vertical and diagonal wavelet coefficients respectively. After 2-D wavelet decomposition, an image signal is decomposed into four sub-images with sizes quarter of the original image due to down sampling. Figure 2.6 demonstrates the 2-D wavelet subbands at level one [55].

Indeed, the 2-D wavelet transform of an image is nothing but the 1-D wavelet transform applied to an image in both x and y direction. Let we are given a 2-D input signal (image) of size (NxN). If 1-D wavelet transform is applied to the image in x-direction (row operation) first, two sub-images say (L and H) of size NxN/2 respectively will be obtained.

If the 1-D wavelet transform is applied to the sub-images L and H along y-direction (column operation) four sub-images (LL1, LH1, HL1 and HH1) with size N/2x/N/2 respectively will be obtained. The reduction in the size of the images is due to down-sampling. The block diagram of the two dimensional (2-D) wavelet decomposition is given in Figure 2.6. The inverse 2-D wavelet transform can be performed by applying 1-D inverse wavelet transform in the reverse order (first in y-direction and then in x-direction) by using the reconstruction filters $\{g_0(k), g_1(k)\}$ instead of decomposition filters $\{h_0(k), h_1(k)\}$.

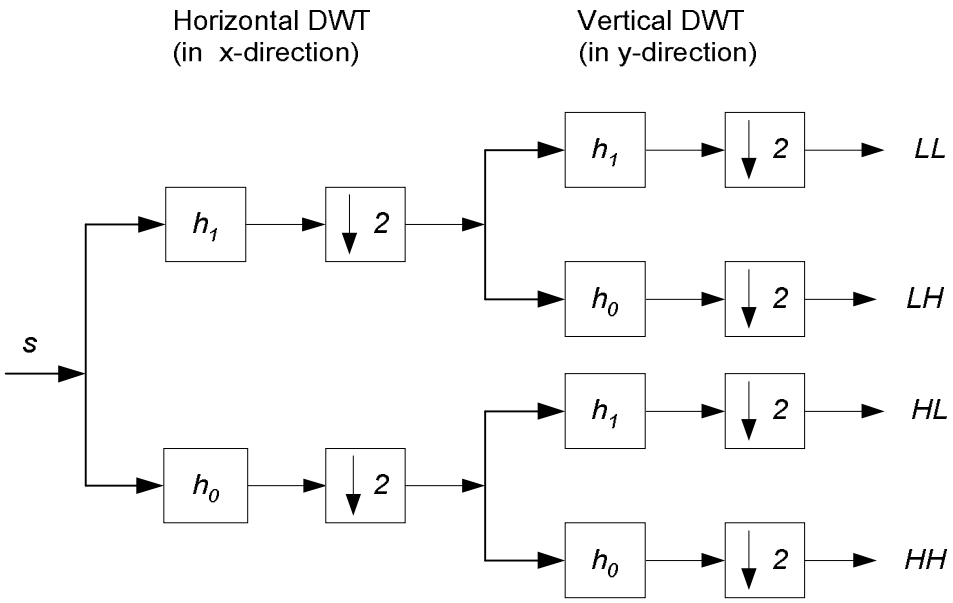

Figure 2.6 Discrete 2-D wavelet packet decomposition tree at level one

The 2-D wavelet packet transform mimics the 2-D wavelet transform. It represents the 1-D wavelet transform in x-direction first and then in y-direction as in the 2-D wavelet case. The difference is that, not only the approximation coefficients (LL1) but also the detail coefficients (LH1, HL1 and HH1) are decomposed at each level of decomposition. The 2-D wavelet packet algorithm is no more difficult than the 2-D wavelets; however it has more computational complexity.

2.10 Biorthogonal Wavelet Systems

A wavelet expansion system that is orthogonal across both translation and scale gives a clean, robust and symmetric representation with Parseval's theorem. However, it also produces some limitations. Requiring orthogonality results in complicated design equations, prevents linear phase analysis and synthesis filterbanks and prevents asymmetric analysis and synthesis systems. This section is devoted to develop the biorthogonal wavelet system using non-orthogonal basis and dual basis to allow greater flexibility in achieving other goals at the expense of energy partitioning property which Parseval's theorem states [54-56, 64-66].

2.10.1 Two Channel Biorthogonal Filter Banks

As explained previously, the analysis and synthesis filters are time reversal of each other for orthogonal wavelets, i.e. $\tilde{h}(n) = h(-n),\ \tilde{g}(n) = g(-n)$. In case of biorthogonal wavelets we relax these limitations however; these four filters still have to satisfy some set of conditions for perfect reconstruction of the input. Figure 2.7 shows a two channel biorthogonal filterbanks [54].

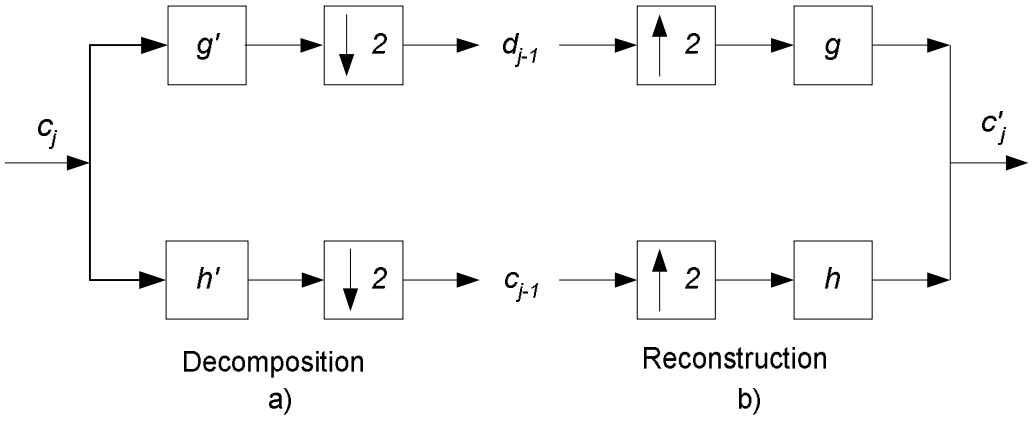

Figure 2.7 Two channel biorthogonal filterbank; a) Decomposition b) Reconstruction

$$\tilde{g}(n) = (-1)^n h(1-n), \qquad g(n) = (-1)^n \tilde{h}(1-n) \tag{2.58}$$

The four filters are cross-related by time reversal and by flipping signs of every other element. When $\tilde{h} = h$, we get the familiar relationship between the scaling and wavelet coefficients for orthogonal wavelets, $g(n) = (-1)^n h(1-n)$.

For biorthogonal case, we have

$$\sum_n \tilde{h}(n) h(n+2k) = \delta(k) \tag{2.59}$$

where $\tilde{h}(n)$ is orthogonal to $h(n)$, but in orthogonal case we have $\sum_n h(n)h(n+2k) = \delta(k)$, i.e., $h(n)$ is orthogonal to even translation of itself.

2.10.2 Advantages of Biorthogonal Wavelets

The well known advantages of the biorthogonal wavelets can be itemized as follows:

- The orthogonal wavelet filter and scaling filters must be the same length and must be even. This restriction is greatly relaxed for biorthogonal wavelets.
- Symmetric wavelet and scaling functions are possible in the structure of biorthogonal wavelets.
- In the biorthogonal systems, if we switch the role of the primary and dual, the system is still sound. Hence, we can choose the best arrangement for our applications. For example in image enhancement, we should use the smoother pairs to reconstruct the enhanced image for better visual appearance.
- In statistical signal processing, white Gaussian noise remains white after orthogonal transform. If the transforms are non-orthogonal, the noise becomes correlated or colored. Thus, when biorthogonal wavelets are used in estimation and detection one may need to adjust the algorithm to better address the colored noise.

Since biorthogonal wavelet systems are very flexible, there are many approaches to design different biorthogonal systems. The key point is to design h and $\widetilde{h}$ that satisfy $\sum_n \widetilde{h}(n)h(n+2k) = \delta(k)$ and $\sum_n h(n) = \sum_n \widetilde{h}(n) = \sqrt{2}$, and have other desirable characteristics. Some of the popular biorthogonal wavelet families are Cohen-Daubechies-Feauveau (CDF) and Tian-Wells families of biorthogonal wavelets.

2.11 Lifting Construction of Biorthogonal Wavelets

The lifting wavelet transform or simply *lifting scheme* is an alternative method for construction of biorthogonal wavelets [(32–33, 54, 56, 62–66]. The lifting scheme offers several advantages over the classical wavelet transform. It is a spatial domain method, it is easier to implement, it allows faster and in-place calculations, it allows nonlinear, adaptive, irregularly sampled and integer to integer wavelet transforms, it

is easier to obtain inverse transform. Furthermore, any wavelet transform can be factored into lifting steps [34].

The classical 1-D lifting scheme consists of the following three steps;

Split: Divide the original data set into even and odd indexed subsets. The original data set $x[n]$ is split into even indexed points $x_e[n] = x[2n]$ and odd indexed points $x_o[n] = x[2n+1]$.

Predict: Obtain wavelet coefficients (detail coefficients) $d[n]$, as the error in prediction of $x_o[n]$ from $x_e[n]$, using prediction operator P;

$$d[n] = x_o[n] - P\left(x_e[n]\right), \tag{2.60}$$

Update: Generate scaling coefficients (approximation coefficients) $c[n]$, by combining even indexed points $x_e[n]$ and detail coefficients $d[n]$, by applying update operator U to the detail coefficients $d[n]$;

$$c[n] = x_e[n] + U\left(d[n]\right). \tag{2.61}$$

The general single level forward lifting scheme is given in Figure 2.8.

The lifting steps can easily be inverted even if the lifting filters P and U are nonlinear or adaptive. Again the inverse lifting scheme is given in three steps;

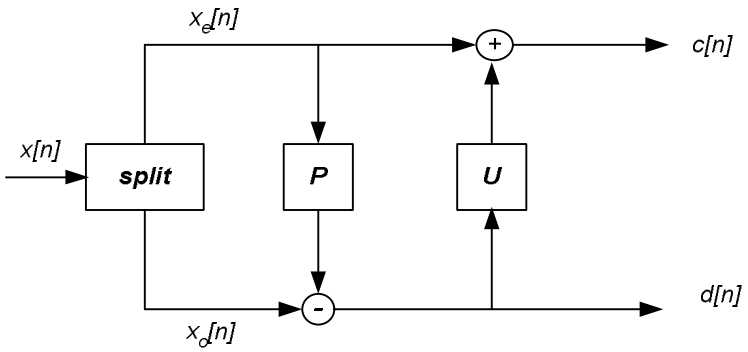

Figure 2.8 Forward lifting scheme

$$x_e[n] = c[n] - U\big(d[n]\big). \tag{2.62}$$

Undo predict: From (2.60) we get,

$$x_o[n] = d[n] + P\big(x_e[n]\big). \tag{2.63}$$

Merge: Combining (2.62) and (2.63) the estimation of the original signal can be obtained.

The inverse lifting scheme is given in Figure 2.9.

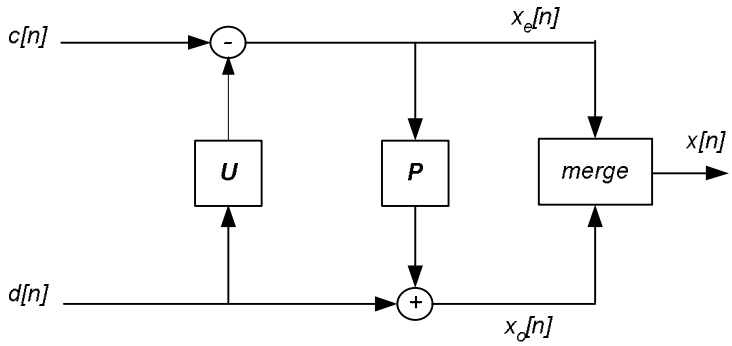

Figure 2.9 The inverse lifting scheme

CHAPTER 3

SPEECH ENHANCEMENT

3.1 Introduction

This chapter of the thesis contains theoretical background, characteristics and literature survey on the recently proposed single channel speech enhancement methods and related concepts, such as noise estimation, a priori SNR estimation and estimation of a priori probability of speech absence. The proposed adaptive lifting scheme and proposed perceptual filterbank are also determined within the first part of this chapter.

A vast amount of literature has been emerged recently on the speech enhancement methods such as multi-channel approaches [4, 5, 65], single channel approaches [6-23], and methods based on wavelet thresholding [24-26]. However, we have limited scope of the thesis (which includes speech enhancement and image enhancement in the lifting-wavelet domain) with the single channel methods. Since, otherwise it would become excessively wide. Moreover, the single channel speech enhancement methods are the most practical and widely used methods requiring no extra sensor for noise source. The clean speech signal is estimated directly from the noisy speech signal in the single channel speech enhancement methods

For practical applications, the most popular single channel approaches proposed in the last two decades have been tested.

The methods depend on frame based analysis of the speech signals using STFT because of the nonstationary nature of the speech signals. The STFT uses a specific window function to multiply the signal and then calculates the DFT coefficients based on the assumption that the speech is stationary in the window. The noise estimation is performed based on speech-pause detection using VAD. The phase of the noisy speech signal is not modified, since the human auditory system is sensitive to magnitude or energy rather than the phase information. The spectral subtraction

method successively reduces the noise level but its main drawback is that, it causes excessive residual and musical noise.

Speech is assumed to always exist in the input speech signal in most of the noise suppression methods. However, speech does not always exist in the input speech and this causes poor enhancement results. The frames which do not contain speech are detected by using voice activity detectors (VAD). The process is called as speech-pause detection or *hard decision* filtering. The main disadvantage of hard decision filters is that, they cause musical noise and degrades the naturalness of the speech.

To overcome the drawbacks of hard decision filters, Mc Cully and Malpass [12] proposed *soft decision* noise suppression filters. The method is based on modification of the gain function, according to probability of speech presence. The proposed estimator is maximum likelihood-short time spectral amplitude (ML-STSA) estimator, modified according to probability of speech presence.

Ephraim and Malah proposed the minimum mean-square error short-time spectral amplitude (MMSE-STSA) and minimum mean-square error log-spectral amplitude MMSE LSA) estimator [9,10]. The estimators are derived based on minimizing the mean square error (MSE) of the short time spectra (or log-short time spectra) of amplitude estimation.

Both methods are based on the Gaussian statistical model and a priori SNR estimation which is reported to be the key parameter in the MMSE-STSA (or MMSE-LSA) estimators [12, 66]. It is widely reported [67] that the MMSE estimators cause no musical background noise.

The gain functions of the above mentioned estimators have been also derived based on soft-decision aspects given in [12] when the probability of speech absence is taken into account [11]. The a priori SNR is estimated from the noisy speech signal via *"Decision Directed"* method. The noise PSD is estimated based on speech pause detection via a recursive equation based on probability of speech absence. The probability of speech absence is estimated frame-based for each frequency bin.

In general, the single channel speech enhancement methods improve the objective quality of the enhanced speech signal. However, they have no effect on the intelligibility of the enhanced speech signal. This means that, a speech signal with good SNR may have poor intelligibility or perceived quality or vise versa. Recently, many speech enhancement methods taking into account characteristics of the human auditory system have been developed [16–18], in order to improve the intelligibility of the speech signals. These perceptual models are generally based on the critical-band decompositions or noise masking properties [16, 30, 31]. By adjusting the subbands of WPD tree according to critical-bands of the human auditory system, perceptual filterbanks which lead to efficient speech enhancement algorithms can be designed [29-31].

The CDF (1, N) group of CDF (Cohen, Daubechies, Feauveau) family lifting filters, based on the average interpolation [68, 69] have been employed in the proposed adaptive lifting scheme as also used in the proposed work in [37,38]. However, we introduced a new adaptive prediction method in the proposed adaptive lifting scheme. The motivation behind introducing adaptivity into the lifting scheme is that, choosing better prediction (or update) filters may give rise to more efficient signal representations.

The subbands of the wavelet packet decomposition (WPD) tree structure obtained via the proposed adaptive lifting scheme have been adjusted according to critical bands. The new decomposition tree structure is called as critical-bands wavelet packet decomposition (CB-WPD) tree. Since the new decomposition tree structure corresponds to a perceptually motivated nonuniform filterbank, it is called as “perceptual filterbank” in the thesis.

3.2 Single Channel Speech Enhancement Methods

The term single channel means that there is only one microphone for both noise and speech signals. In other words, there is no separate microphone for noise source. Therefore, the estimation of noise from a noisy speech signal is critically important in the single channel speech enhancement methods. The noise is generally estimated during noise-only frames. The single channel methods are difficult to implement

however, they are the most widely used methods in the speech enhancement. In this section, the most widely used single channel methods such as spectral subtraction type speech enhancement methods [6-13], soft decision noise suppression filters [12] and MMSE-based estimators (STSA, LSA and MM-LSA) [9-11] are given in detail.

3.2.1 Subtractive-Type Speech Enhancement methods

The spectral subtraction is the most common subtractive type algorithm. It is based on estimation of magnitude (or power) spectrum of the original speech signal, by subtracting an estimate of the average noise spectrum from the noisy speech signal spectrum [6, 7, 67, 70].

The methods are based on short-term spectral analysis of the speech and noise because of the nonstationary nature of the speech signals. The frame based analysis of the speech signal using the discrete Fourier transform (DFT) is called as short time Fourier transform (STFT). The STFT uses a specific window function $w[n]$ to multiply the signal and then calculates the DFT coefficients based on the assumption that the speech is stationary within the window. The noise is estimated during speech pauses using VAD. Generally, Hamming (or Hanning) window with % 50 overlap is used in applications in order to avoid boundary (or block) effect. The discrete STFT is given in Appendix A1.

3.2.1.1 Spectral Subtraction

With $\hat{X}(k,l)$ and $\hat{B}(k,l)$ are estimated short-time spectra (kth frequency bin in lth discrete time frame) of original speech signal $x[n]$ and noise signal $b[n]$ respectively, in discrete time domain. The noise estimate $\hat{B}(k,l)$ is obtained during speech pauses or non-speech frames by using a VAD and $\hat{X}(k,l)$ is estimated as follows. Let,

$$\hat{X}(k,l) = \left[|Y(k,l)|^{\gamma} - \left|\hat{B}(k,l)\right|^{\gamma} \right]^{\gamma} e^{j\ \theta_y(k,l)} \tag{3.1}$$

$$\hat{X}(k,l) = \begin{cases} \hat{X}(k,l), & \text{if } \hat{X}(k,l) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

Gain function for generalized spectral subtraction can be given as

$$G(k,l)=\left[\frac{\left|\hat{X}(k,l)\right|^{\gamma}}{\left|Y(k,l)\right|^{\gamma}}\right]^{\frac{1}{\gamma}}=\left[1-\frac{\left|\hat{B}(k,l)\right|^{\gamma}}{\left|Y(k,l)\right|^{\gamma}}\right]^{\frac{1}{\gamma}}=\left[1-\frac{1}{\frac{\left|Y(k,l)\right|^{\gamma}}{\left|\hat{B}(k,l)\right|^{\gamma}}}\right]^{\frac{1}{\gamma}} \tag{3.3}$$

Where, $\gamma=1$, for *Magnitude Spectral Subtraction (MSS),* and $\gamma=2$ *for Power Spectral Subtraction (PSS)* [6]. The amplitude spectrum (for magnitude spectral subtraction case) of estimated clean speech $\hat{X}(k,l)$ can be obtained as

$$\left|\hat{X}(k,l)\right|=G(k,l)\left|Y(k,l)\right| \tag{3.4}$$

An estimate of the magnitude spectrum is combined with the phase of the noisy signal to restore the enhanced speech signal. The phase of the noisy signal is not modified, as it is known that, human auditory system is more sensitive to magnitude (or energy) and tends to ignore the phase information. After the inverse DFT and overlap add the enhanced speech signal is obtained in the time domain.

The spectral subtraction method successively reduces the noise level but the main drawback of the method is that it causes excessive residual and musical noise. These musical tones have annoying nature for listeners. Block diagram of spectral subtraction method is given in Figure 3.1.

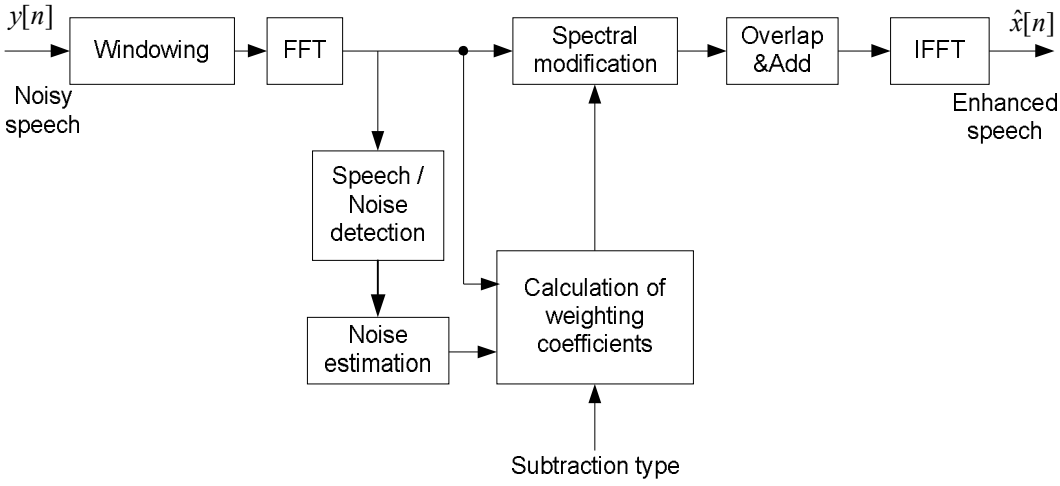

Figure 3.1 Block diagram of spectral subtraction method

3.2.1.2 STFT-Based Wiener Filter

The Wiener *filter (WF)* is generally classified as a subtraction type filter. A widely used gain function for STFT based Wiener filter is given in the form:

$$G(k,l) = \frac{\hat{\Phi}_{xx}(k,l)}{\hat{\Phi}_{xx}(k,l) + \hat{\Phi}_{dd}(k,l)} \tag{3.5}$$

Where, $\hat{\Phi}_{xx}(k,l)$ and $\hat{\Phi}_{dd}(k,l)$ are the short-time power spectral density (PSD) estimates of $x[n]$ and, $b[n]$ respectively. The noise power is estimated during noise only frames as used in power spectral subtraction. The main drawback of Wiener filtering is that, it produces annoying tonal artifacts similar to spectral subtraction.

The gain function for Wiener filter can also be given in the following form:

$$G(k,l) = \frac{|Y(k,l)|^2 - |\hat{B}(k,l)|^2}{|Y(k,l)|^2} = 1 - \frac{1}{\dfrac{|Y(k,l)|^2}{|\hat{D}(k;l)|^2}} \tag{3.6}$$

The term $\dfrac{|Y(k,l)|^2}{|\hat{D}(k,l)|^2}$ in (3.6) is called as *a posteriori SNR* estimation $SNR_{post}(k,l)$

Then rewriting (3.6) we have gain function for Wiener filter based on a posteriori *SNR* estimation as

$$G(k,l)_W = 1 - \frac{1}{SNR_{post}(k,l)} \tag{3.7}$$

Similarly the gain functions of magnitude and power spectral subtraction in (3.3) can be given as follows, based on a posteriori SNR.

$$\begin{aligned} G(k,l)_{MSS} &= 1 - \sqrt{\frac{1}{SNR_{post}(k,l)}} \\ G(k,l)_{PSS} &= \sqrt{1 - \frac{1}{SNR_{post}(k,l)}} \end{aligned} \tag{3.8}$$

The *instantaneous SNR*, a local estimate of the SNR, is given as

$$SNR_{inst} = SNR_{post} - 1 \tag{3.9}$$

A priori SNR estimate (SNR_{prio}) is formulated as,

$$SNR_{prio}(k,l) = \frac{E\left[\left|X(k,l)\right|^2\right]}{\left|\hat{D}(k,l)\right|^2} \tag{3.10}$$

The a priori SNR is estimated from the noisy speech signal by using the "*Decision Directed"* method [9].

3.2.2 MMSE-STSA Estimation-Based Methods

The minimum MMSE STSA estimator and STSA-Wiener estimator [9], minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator have been outlined in this section.

The methods are based on the Gaussian statistical model and a priori SNR estimation. The a priori SNR is reported to be a key parameter (rather than noise

variance) in the reduction of noise removal and speech distortion [12, 66]. Furthermore, it is widely reported [67] that, the MMSE based estimators cause no musical background noise. The gain functions of the above mentioned estimators have been also derived when the probability of speech absence is taken into account [9-11, 71]. The estimation of noise power spectral density (PSD) from the noisy speech signal via *"Decision Directed"* method and the estimation of probability of speech absence have been also outlined in this section.

To derive the MMSE STSA estimator, the a priori probability distribution of the speech and noise Fourier expansion coefficients should be known. Since in practice they are unknown, one can measure each probability distribution or alternatively, assume a reasonable statistical model such as Gaussian statistical model [9-10].

While the distortion measure of mean-square error of the spectra (the original STSA estimator) used in [9] is mathematically tractable and leads also to good results, it is not the most subjectively the meaningful one. It is well known that, a distortion measure which is based on the mean-square error of the log-spectra is more suitable for speech processing [10]. Such a distortion measure is therefore widely used for speech analysis and recognition. Therefore, it is of great interest to examine the STSA estimator which minimizes the mean-square error of the log- spectra (MMSE-LSA estimator) in enhancing noisy speech.

One useful approach to resolve this problem is to derive an MMSE-STSA estimator which takes into account the uncertainty of speech presence in the noisy observations. Such an estimator can be derived on the basis of the above Gaussian statistical model and by assuming that the speech does not always exist in the signal.

3.2.2.1 Soft-Decision Based Gain Modification Taking Into Account Probability of Speech Absence

Noise suppression properties of the above enhancement algorithms have been shown to improve when soft-decision based modification of the gain function depending on the probability of speech absence [9-12].

Let, the noisy input signal $y[n]$ is given by, $y[n] = x[n] + b[n]$ in the discrete time domain, where $x[n]$ is the clean speech signal and $b[n]$ is the noise signal and $x[n]$ and $b[n]$ are assumed to be uncorrelated. The STFT is applied $x[n]$ to compute the overlapping windowed frames. In the frequency domain, $Y_k = X_k + B_k$ where $X_k = A_k \exp(j\alpha_k)$ and $Y_k = R_k \exp(j\beta_k)$. Since the MMSE estimators are based on Gaussian statistics, the DFT coefficients of both speech and noise frames are assumed to be independent Gaussian random variables.

The gain functions for the MMSE based estimators (MMSE-STSA, MMSE-LSA, and STSA-Wiener and MM-LSA) have been derived mainly based on two assumptions: The first one assumes that the speech always exists at *k*th bin (i.e. $q_k = 0)$, and derives the frequency dependent gain functions based on this assumption [9-10]. The second one assumes that, speech does not always exist at *k*th bin, because of the quasi harmonic nature of the speech signal. Furthermore, it takes into account the probability of speech absence and modifies the gain functions based on this probability. The value of probability of speech absence can be chosen as a fixed value (i.e. $q_k = 0.2$ in [9] or $q_k = 0.5$ in [12]) for all frequency bins or more realistically it can be estimated for each frequency bin as $q_k(l)$. The modified gain functions for the MMSE based estimators can be seen in [9-11, 71]

Assume that, C_k is a function of short time amplitude A_k of the clean speech in the k*th* frequency bin taking into account the probability of speech absence, we have

$$\hat{C}_k = E\{C_k | Y_k, H_1^k\} P(H_1^k | Y_k) + E\{C_k | Y_k, H_0^k\} P(H_0^k | Y_k) \qquad (3.11)$$

Where,

H_0^k: Speech absent at k*th* bin,

H_1^k: Speech present at k*th* bin.

Where, $E\{.|.\}$ and $P\{.|.\}$ denote conditional expectation and conditional probability operators, respectively. Since the second term in (3.11) is zero, we have

$$\hat{C}_k = E\left\{C_k \middle| Y_k, H_1{}^k\right\} P\left(H_1{}^k \middle| Y_k\right) \tag{3.12}$$

Where, $P(H_1^k \middle| Y_k)$ is the soft decision modification of optimal estimator under signal presence uncertainty.

Applying Bayes' rule [9, 12] gives

$$\begin{aligned} P(H_1^k \middle| Y_k) &= \frac{p\left(Y_k \middle| H_1^k\right) P\left(H_1^k\right)}{p\left(Y_k \middle| H_0^k\right) + p\left(Y_k \middle| H_1^k\right) P\left(H_1^k\right)} \\ &= \frac{\Lambda(k)}{1+\Lambda(k)} \\ &\stackrel{\Delta}{=} G_M(k) \end{aligned} \tag{3.13}$$

Where $p\{.\middle|.\}$ denotes conditional probability density operator.

$$\Lambda(k) \stackrel{\Delta}{=} \mu_k \frac{p\left(Y_k \middle| H_1{}^k\right)}{p\left(Y_k \middle| H_0^k\right)} \quad \text{and} \quad \mu_k \stackrel{\Delta}{=} \frac{P\left(H_1{}^k\right)}{P\left(H_0{}^k\right)} = \frac{1-q_k}{q_k} \tag{3.14}$$

$\Lambda(k)$ in (3.13) denotes the generalized likelihood ratio taking into account the uncertainty of speech presence and q_k is the probability of speech absence in the *k*th frequency bin.

By considering the soft decision gain modification (or taking into account the probability of speech absence) the modified gain functions for the STSA, LSA and Wiener estimators can be obtained as given in the Sections 3.2.2.2-3.2.2.4.

3.2.2.2 Modified STSA Estimator

Substituting $C_k = A_k$ in (3.12) the modified amplitude estimate $(\hat{A}_{STSA\,\mathrm{mod}})_k$, for the MMSE-STSA estimator [9] can be obtained follows:

$$
\begin{aligned}
(\hat{A}_{STSA\,\mathrm{mod}})_k &= \left[E\left\{A_k \middle| Y_k, H_1^k\right\} G_M(k)\right] \\
&= G_M(k) G_{STSA}(k) R_k
\end{aligned}
\tag{3.15}
$$

Since the gain modification $G_M(k)$ is multiplicative, the modified gain function for the MMSE-STSA estimator $G_{STSA\,\mathrm{mod}}(k)$ is given as follows.

$$
\begin{aligned}
\frac{(\hat{A}_{STSA\,\mathrm{mod}})_k}{R_k} &= G_M(k) G_{STSA}(k) \\
&\stackrel{\Delta}{=} G_{STSA\,\mathrm{mod}}(k)
\end{aligned}
\tag{3.16}
$$

Where, $G_{STSA}(k)$ is the original gain function (without taking into account speech presence uncertainty or without modification) for MMSE-STSA estimator given in detail in [9] as follows

$$
G_{STSA}(k) = \frac{\sqrt{\pi}}{2} \sqrt{\frac{1}{\gamma_k} \frac{\xi_k}{1+\xi_k}} F\left\{\gamma_k \frac{\xi_k}{1+\xi_k}\right\} \tag{3.17}
$$

$$
\text{with,} \quad F(x) = \exp(-x/2)\left[(1+x) I_0\left(\frac{x}{2}\right) + x I_1\left(\frac{x}{2}\right)\right]
$$

From (3.16) and (3.17), the modified gain function for the MMSE-STSA estimator $G_{STSA\,\mathrm{mod}}(k)$ can be obtained as follows:

$$
\begin{aligned}
G_{STSA\,\mathrm{mod}}(k) &= G_M(k) G_{STSA}(k) \\
&= \frac{\Lambda(k)}{1+\Lambda(k)} \frac{\sqrt{\pi}}{2} \sqrt{\frac{1}{\gamma_k} \frac{\xi_k}{1+\xi_k}} F\left\{\gamma_k \frac{\xi_k}{1+\xi_k}\right\}
\end{aligned}
\tag{3.18}
$$

$$
\text{with} \quad F(x) = \exp(-x/2)\left[(1+x) I_0\left(\frac{x}{2}\right) + x I_1\left(\frac{x}{2}\right)\right]
$$

3.2.2.3 MM-LSA Estimator

Substituting $C_k = \log A_k$ in (3.12) the modified amplitude estimate $(\hat{A}_{LSA})_k$ for the MMSE-LSA estimator [10] is obtained as given in (3.19).

$$
\begin{aligned}
(\hat{A}_{LSA})_k &= \exp\left[E\left\{\log A_k \middle| Y_k, H_1^k\right\} G_M(k)\right] \\
&\triangleq \left[G_{LSA}(k) R_k\right]^{G_M(k)}
\end{aligned}
\tag{3.19}
$$

Since the soft decision gain modification of R_k in (14) is not multiplicative and it is explained in [10] that it did not result in meaningful improvement over using $G_{LSA}(k)$ alone, the following multiplicatively modified LSA estimator (MM-LSA) was chosen to use [11,71].

$$
\begin{aligned}
(\hat{A}_{MM-LSA})_k &= G_M(k) G_{LSA}(k) R_k \\
\frac{(\hat{A}_{MM-LSA})_k}{R_k} &= G_M(k) G_{LSA}(k) \\
&\triangleq G_{MM-LSA}(k)
\end{aligned}
\tag{3.20}
$$

Where $G_{LSA}(k)$ and $G_{MM-LSA}(k)$ are the original and the multiplicatively-modified gain functions for the MMSE-LSA estimator.

Since *gain function for MMSE-LSA estimator* is given in [11] as

$$
G_{LSA}(k) = \frac{\xi_k}{1+\xi_k} \exp\left(\frac{1}{2} \int_{\upsilon_k}^{\infty} \frac{e^{-t}}{t} dt \right)
\tag{3.21}
$$

From (3.20) and (3.21), the multiplicatively-modified gain function for MMSE-LSA estimators $G_{MM-LSA}(k)$ is obtained as follows

$$
\begin{aligned}
G_{MM-LSA}(k) &= G_M(k) G_{LSA}(k) \\
&= \frac{\Lambda(k)}{1+\Lambda(k)} \frac{\xi_k}{1+\xi_k} \exp\left(\frac{1}{2} \int_{\upsilon_k}^{\infty} \frac{e^{-t}}{t} dt \right)
\end{aligned}
\tag{3.22}
$$

3.2.2.4 Modified Wiener Estimator

The amplitude estimate for modified STSA-Wiener estimator $G_{W\,\mathrm{mod}}(k)$ can be given as follows:

$$
\begin{aligned}
(\hat{A}_{W\,\mathrm{mod}})_k &= G_M(k)G_W(k)R_k \\
\frac{(\hat{A}_{W\,\mathrm{mod}})_k}{R_k} &= G_M(k)G_W(k) \\
&\triangleq G_{W\,\mathrm{mod}}(k)
\end{aligned}
\tag{3.23}
$$

Since G_W (the original gain function for STSA-Wiener estimator) is given in [9, 67] as

$$
G_W(k) = \frac{\xi_k}{1+\xi_k} \tag{3.24}
$$

From (3.23) and (3.24) the multiplicatively-modified gain function for STSA-Wiener estimator $G_{W\,\mathrm{mod}}(k)$ can be derived as follows

$$
\begin{aligned}
G_{W\,\mathrm{mod}}(k) &= G_M(k)G_W(k) \\
&= \frac{\Lambda(k)}{1+\Lambda(k)}\,\frac{\xi_k}{1+\xi_k}
\end{aligned}
\tag{3.25}
$$

Where

$$
\begin{aligned}
&\xi_k \triangleq \frac{\eta_k}{1-q_k}, \quad \eta_k \triangleq \frac{E\{A_k^{\,2}\}}{\lambda_b(k)}, \quad \lambda_b(k) \triangleq E\{|B_k|^2\}, \quad \upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \\
&\gamma_k \triangleq \frac{R_k^{\,2}}{\lambda_b(k)}, \quad \Lambda(k) = \mu_k\,\frac{\exp(\upsilon_k)}{1+\xi_k}, \quad \mu_k = \frac{q_k}{1-q_k}
\end{aligned}
\tag{3.26}
$$

The γ_k denotes a posteriori SNR, η_k and ξ_k denote unconditional and conditional a priori SNR respectively, $\lambda_b(k)$ denotes noise power spectral density (PSD) (or noise power spectrum variance) and q_k denotes a priori probability of speech absence.

Note here that, when $q_k = 0$ (speech always exists in kth bin), $\Lambda/(1+\Lambda) = 1$ in (3.13). In this case, the modified gain functions for LSA, STSA and Wiener estimators given in (3.18), (3.22) and (3.25) reduce to the standard gains functions

given in (3.17), (3.21) and (3.24). Furthermore, when $q_k = 0$ (speech always exists in *k*th bin), $\xi_k = \eta_k$ in (3.26).

When both the clean speech signal and noise signals are known, it is possible to obtain noise variance $\lambda_b(k)$ and a priori SNR η_k as given in (3.26). However, in practical applications neither noise signal nor clean speech signal is available. We have only the noisy observation signal. Consequently, both the noise variance $\lambda_b(k)$ and the a priori SNR η_k should be estimated from the noisy observation signal as we do in this paper. In the rest of this section the estimation of $\lambda_b(k)$, η_k and q_k adaptively from the noisy observation signal is given in detail.

3.2.2.5 Noise Power Spectral Density Estimation

Noise power spectral density (PSD) estimation is critically important in the speech enhancement system since the accuracy of the noise estimation has a major impact on the performance of the overall system. If the noise is known to be stationary, then it is enough to estimate its PSD estimate only once, from an initial noise only period. In general, voice activity detector (VAD) based methods are used for the recursive estimation of noise PSD during speech pauses and updated during consecutive pause frames. It is useful to control the update rate by using a smoothing factor (α_b) during the detected speech pauses since spectral changes may occur during periods of speech absence.

In general, the noise PSD estimation based on pause detection using VAD is reliable for tracking the stationary noise; however, the highly nonstationary noises can not be sufficiently tracked by recursive noise PSD estimation during speech pauses. Moreover, the VAD based noise estimators are difficult to tune and their application at low SNRs often result in distortion in speech [72, 73].

The general equation used for recursive estimation of the noise PSD during noise-only frames using VAD is as given in (3.27).

$$\hat{\lambda}_b(k,l) = \alpha_b \hat{\lambda}_b(k,l-1) + (1-\alpha_b) R_k^2(l) \tag{3.27}$$

Where, $\hat{\lambda}_b(k,l)$ and $R_k^2(l)$ denote estimated noise PSD estimate and noisy speech signal power spectrum respectively, at *k*th bin in *l*th noise frame, α_b (smoothing factor), is generally set to a value (0.8-0.98), and $(l-1)$ denotes previous frame. The α_b value is chosen as (0.85) in the applications.

Some of the recently proposed methods for noise PSD estimation which does not need speech pause detection and noise estimation can be tracked also during speech activity. Malah at al. [11, 71] proposed a noise estimation method where a modified smoothing parameter has been used for controlling update rate of the noise spectrum estimate. The smoothing parameter (α_b) is modified by using frame based estimation of probability of speech absence as

$$\alpha_b(k,l) = 1 - \Gamma_b \left| \bar{\gamma}_\kappa(l-1) - 1 \right| \hat{q}_k(l), \quad k \in \kappa \tag{3.28}$$

Where Γ_b is a constant (i.e. $\Gamma_b = 0.2$), κ denotes a set of frequency bins for which the update is performed, so that the $\bar{\gamma}_\kappa$ is the mean of γ_k over all $k \in \kappa$. A different VAD which allow controlling the update rate during also speech frames have been used. The VAD operates based on a criterion that, the $\hat{q}_k(l)$ is larger than a threshold value or γ_k has a relatively low value (i.e. $\gamma_k \leq 4$).

Martin [71] has proposed an efficient noise estimator based on minimum statistics to track non-stationary noise. However, the buffer length necessary to bridge peaks of speech activity makes it difficult to tract any rapid variations in noise spectrum.

Cohen [74] has used the minimum statistics as a voice activity detector and estimated the noise by a recursive averaging formula. The performance results are reported only for white Gaussian noise which is a poor approximation for noise containing high bursts.

In our applications satisfactory results have been achieved by using a VAD based noise estimation method given in (3.27) where noise PSD is estimated and updated during the detected noise only frames.

3.2.2.6 A Priori SNR Estimation

It is reported in [9, 12, 67] that a priori SNR estimate is a key parameter in the elimination of musical noise and reduction of speech distortion. The a priori SNR can be estimated from the noisy signal by using two methods, maximum likelihood (ML) estimator and "*Decision Directed*" (DD) method. The priori SNR $\hat{\eta}_k(l)$ is estimated using DD method in [12] for *l*th time frame as given in (3.29).

$$\hat{\eta}_k(l) = \alpha \frac{\hat{A}_k^2(l-1)}{\hat{\lambda}_b(k,l-1)} + (1-\alpha)P\left[\gamma_k(l)-1\right] \tag{3.29}$$

Where $P[.]$, denotes half-wave rectification, $\hat{A}_k(l-1)$ is the estimated speech spectrum at previous frame and α (weighting factor) generally takes the values in the range (0.9-0.99). $\hat{\lambda}_b(k,l)$ denotes the noise variance estimation, $\gamma_k(l)$ and $\hat{\eta}_k(l)$ are a posteriori and a priori SNR estimates respectively. Since,

$$\begin{aligned} P[x] &= x, \ \ if \ \ x > 0 \\ &0, \ \ if \ \ x < 0 \end{aligned} \tag{3.30}$$

And hence,

$$P[x] = \max[x,0] \tag{3.31}$$

The (3.12) can be rearranged as given in (3.15).

$$\hat{\eta}_k(l) = \alpha \frac{\hat{A}_k^2(l-1)}{\hat{\lambda}_b(k,l-1)} + (1-\alpha)\max\left\{\gamma_k(l)-1,0\right\} \tag{3.32}$$

From (3.32) we derive,

$$\frac{\hat{A}_k^2(l-1)}{\hat{\lambda}_b(k,l-1)} = \frac{\hat{A}_k^2(l-1)}{R_k^2(l-1)} \frac{R_k^2(l-1)}{\hat{\lambda}_b(k,l-1)} = G^2(k,l-1)\gamma_k(l-1) \tag{3.33}$$

Substituting (3.33) into (3.32) we have,

$$\hat{\eta}_k(l) = \alpha G^2(k, l-1)\ \gamma_k(l-1) + (1-\alpha)\max\{\gamma_k(l) - 1, 0\} \tag{3.34}$$

Where G in (24) denotes gain function of one of the STSA, LSA, Wiener estimators or their modified versions given in [9-11, 71]. α is taken as 0.98 in the evaluations.

3.2.2.7 Estimation of a Priori Probability of Speech Absence

An important property of both MMSE-STSA and MMSE-LSA [9,10] estimators is that they are able to eliminate "musical noise" in the enhanced signal which plagues most other frequency domain algorithms. This can be attributed to the "decision directed" method which is used for estimation of a priori SNR. It is recommended to use a lower limit $\eta_{\min}$ for the estimated η_k. A weighting factor α, in that estimator controls a tradeoff between noise reduction and signal distortion [11].

The probability of speech absence $q_k(l)$ is generally set to a fixed value (0.2 or 0.5) [9, 12]. It can also be adaptively estimated for each frequency bin (k) during consecutive frames (l) as $q_k(l)$ [11, 71].

To decide whether speech is present in the *k*th bin or not we consider the following composite hypothesis testing problem.

$$K_0: \quad \eta_k \geq \eta_{\min} \qquad \text{(Speech present in } k\text{th bin)}$$
$$K_A: \quad \eta_k < \eta_{\min} \qquad \text{(Speech absent in } k\text{th bin)}$$

Where $\eta_{\min}$ (lower threshold value for priori SNR estimation) was chosen as (-25 dB) in our applications. The K_0 (null hypothesis) is used since; its rejection when true is graver than the alternative error of accepting when false.

A good decision rule for this problem is equivalent to the Neyman-Pearson decision rule for the following hypothesis test between simple hypotheses. K_0' : $\eta_k = n_{\min}$ and K_A' : $\eta_k = \eta_k^a < \eta_{\min}$. This gives the test:

$$\gamma_k \underset{K_A}{\overset{K_0}{\gtrless}} \gamma_{TH} \tag{3.35}$$

Where, γ_{TH} (threshold value of posteriori SNR for hypothesis testing) is generally set to value ($\gamma_{TH} = 0.8$). The following recursive equation is used for adaptive estimation of $q_k(l)$.

$$\hat{q}_k(l) = \alpha_q \hat{q}_k(l-1) + (1-\alpha_q) I_k(l) \tag{3.36}$$

Where, α_q denotes smoothing parameter for the probability of speech absence which is generally set to a value in the range (0.9-0.99) and $I_k(l)$ is an index function that denotes the result of the test given in (3.35). $I_k(l) = 1$ if K_0 is rejected when true and $I_k(l) = 0$ if accepted when false[11, 71].

3.2.2.8 Speech Enhancement Method

We employ the MMSE based estimators modified according to probability of speech absence (Mod-WF, Mod-STSA and MM-LSA estimators). Although, in [11,17] the noise PSD is estimated based on probability of speech absence, we achieved satisfactory results using VAD based pause detection. The noise PSD is estimated during the first ten segments and updated and smoothed during consecutive noise only frames by using (3.27). The priori SNR (which is a key parameter in the MMSE based methods) is estimated from the noisy speech signal using "*Decision Directed*" method as given in (3.34). The gain modification is performed by taking into account the probability of speech absence. The probability of speech absence q_k is generally set to a fixed value (0.2 or 0.5) [9, 12]. However, we adaptively estimated $q_k(l)$ for each frequency bin (k) in a short time frame (l) as given in (44).

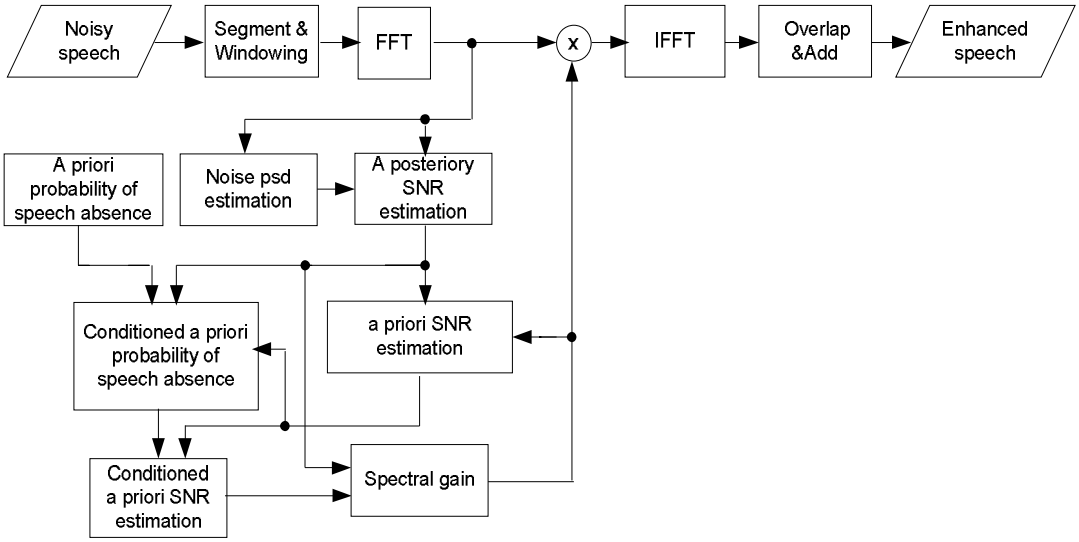

Figure 3.2 Block Diagram of modified MMSE Estimators.

3.3 Proposed 1-D Adaptive Lifting Scheme

In the proposed adaptive lifting scheme, the set of lifting filters CDF (1, N) have been used, based on the average interpolation [68, 69], as also used in the previously proposed work in [37,38]. The original work is based on edge detection and "edge avoiding prediction" method. However, we introduced a new adaptive prediction method in the proposed adaptive lifting scheme.

The motivation behind introducing adaptivity into the lifting scheme is that, choosing better prediction or update filters may give rise to more efficient signal representations and enhancement results.

The idea behind the "edge avoiding prediction" in [37-38] is as follows. The low order predictors well adapt to the edges or break points in the signal while high order predictors well adapt to smooth parts of the signal.

The procedure of "edge avoiding prediction" in [37,38] is as follows. High order predictor CDF(1,7) is used on smooth parts of the signal and low order predictors {CDF(1,1), CDF(1,3) and CDF(1,5) } are used on both sides of an odd indexed edge pixel. The "edge avoiding prediction" method is given below in Figure 3.3.

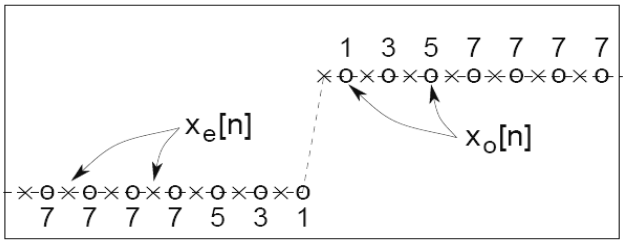

Figure 3.3 Edge avoiding prediction [37, 38]

Since the space adaptive (update-first) lifting scheme is based on adapting the predict stage to the signal structure, the effectiveness of the algorithm is dependent on the sensitivity of prediction stage.

From (3.38-3.41), if the prediction filter (one of the CDF (1, N) group filters) is precisely chosen, the detail coefficient $d[n]$ minimizes since the value of odd pixel $x_o[n]$ will be approximately equal to its predicted value $P(c[n])$. It is clear that, choosing the best prediction filter will result in better adaptation of the prediction filter to the signal structure.

Based on this motivation, we have developed a new prediction algorithm which adaptively chooses the best predictor (among the set of predictors CDF (1, N)) providing the minimum detail coefficient (or minimum prediction error).

Consequently, the prediction filter (or predictor) providing the minimum detail coefficient (for each pixel in the 2-D lifting construction) leads to better signal representation and better enhancement results.

The lifting implementations of the CDF(1,N) group lifting filters with update-first strategy are all given below.

Update stage: The low-pass update coefficients are obtained using Haar filter.

$$c[n] = \frac{1}{2}(x_o[n] + x_e[n]) \tag{3.37}$$

Where, $x_o[n] = x[2n+1]$ denotes odd samples and $x_e[n] = x[2n]$ denotes even samples.

Predict stage: The high pass detail coefficients are obtained as the residues of the prediction of odd samples.

CDF(1,1) lifting scheme (=Haar wavelet transform):

$$d[n] = x_o[n] - c[n] \tag{3.38}$$

CDF(1,3) lifting scheme:

$$d[n] = x_o[n] - (-c[n-1]/8 + c[n] + c[n+1]/8) \tag{3.39}$$

CDF(1,5) lifting scheme:

$$\begin{aligned} d[n] = x_o[n] - (&3c[n-2]/128 - 22c[n-1]/128 + c[n] + \ldots \\ &+ 22c[n+1]/128 - 3c[n+2]/128) \end{aligned} \tag{3.40}$$

CDF(1,7) lifting scheme:

$$\begin{aligned} d[n] = x_o[n] - (&-5c[n-3]/1024 + 44c[n-2]/1024 - 201c[n-1]/1024 + \ldots \\ &+ c[n] + 201c[n+1]/1024 - 44c[n+2]/1024 + 5c[n+3]/1024) \end{aligned} \tag{3.41}$$

The proposed forward adaptive lifting scheme given in Figure 3.4 a) is implemented by rearranging the detail coefficients as follows.

Update Stage: The approximation coefficient $c[n]$ is obtained using (3.37).

Predict Stage: The detail coefficient $d[n]$ is rearranged and given in the form:

$$d[n] = x_o[n] - c[n] - C_N[n] \tag{3.42}$$

Where C_N, $(N = 1,3,5,7)$ in (3.42) is obtained from (3.38–3.41) based on the type of predict filters. The N is obtained for each $(c[n],\ d[n])$ pair for inverse transform.

For CDF(1,1) predict filter $(N = 1)$, from (3.38) and (3.42)

$$C_N[n] = C_1[n] = 0 \tag{3.43}$$

For CDF(1,3) predict filter $(N = 3)$, from (3.39) and (3.42)

$$C_N[n] = C_3[n] = (-c[n-1]/8 + c[n+1]/8) \tag{3.44}$$

For CDF(1,5) predict filter $(N = 5)$, from (3.40) and (3.42)

$$\begin{aligned} C_N[n] = C_5[n] = (3c[n-2]/128 - 22c[n-1]/128 + \ldots \\ + 22c[n+1]/128 - 3c[n+2]/128) \end{aligned} \tag{3.45}$$

For CDF(1,7) predict filter $(N = 7)$, from (3.41) and (3.42)

$$\begin{aligned} C_N[n] = C_7[n] = (-5c[n-3]/1024 + 44c[n-2]/1024 - 201c[n-1]/1024 + \ldots \\ + 201c[n+1]/1024 - 44c[n+2]/1024 + 5c[n+3]/1024) \end{aligned}$$

(3.46)

The detail coefficient $d[n]$ is obtained from (3.42) depending on C_N value given in (3.43) - (3.46) which minimizes the $d[n]$, by using a comparison algorithm.

The inverse adaptive lifting scheme, as given in Figure 3.4 b) can easily be constructed. From (3.42)

$$x_o[n] = d[n] + c[n] + C_N[n] \tag{3.47}$$

where, C_N value is given in (3.43) - (3.46).

Using (3,37) we have

$$x_e[n] = 2c[n] - x_o[n]. \tag{3.48}$$

The enhanced speech signal is obtained based on $x_o[n]$ and $x_e[n]$. The P and U in Figure 3.4 denotes the Haar predict and update filter $(P = 1,\ U = 1/2)$ respectively.

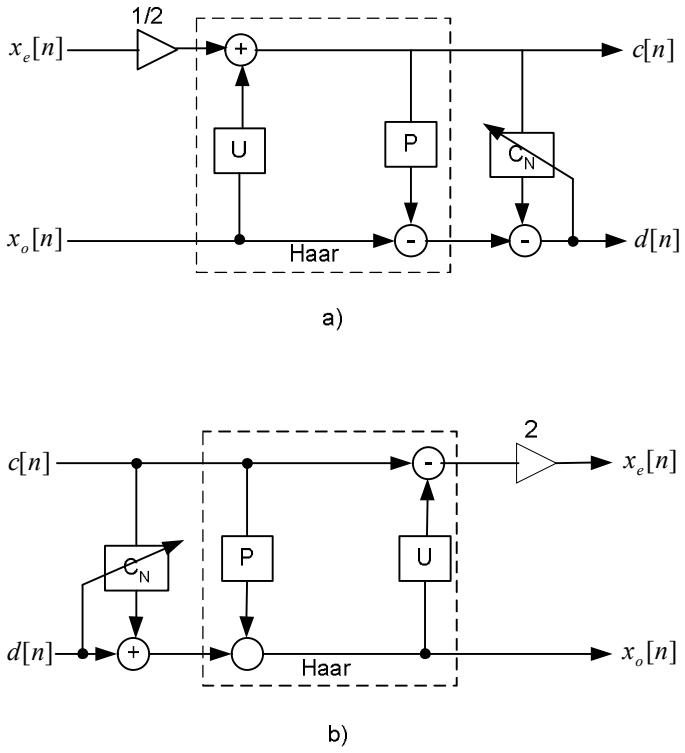

Figure 3.4 Proposed 1-D adaptive lifting scheme, a) Forward adaptive Lifting Scheme, b) Inverse adaptive lifting scheme

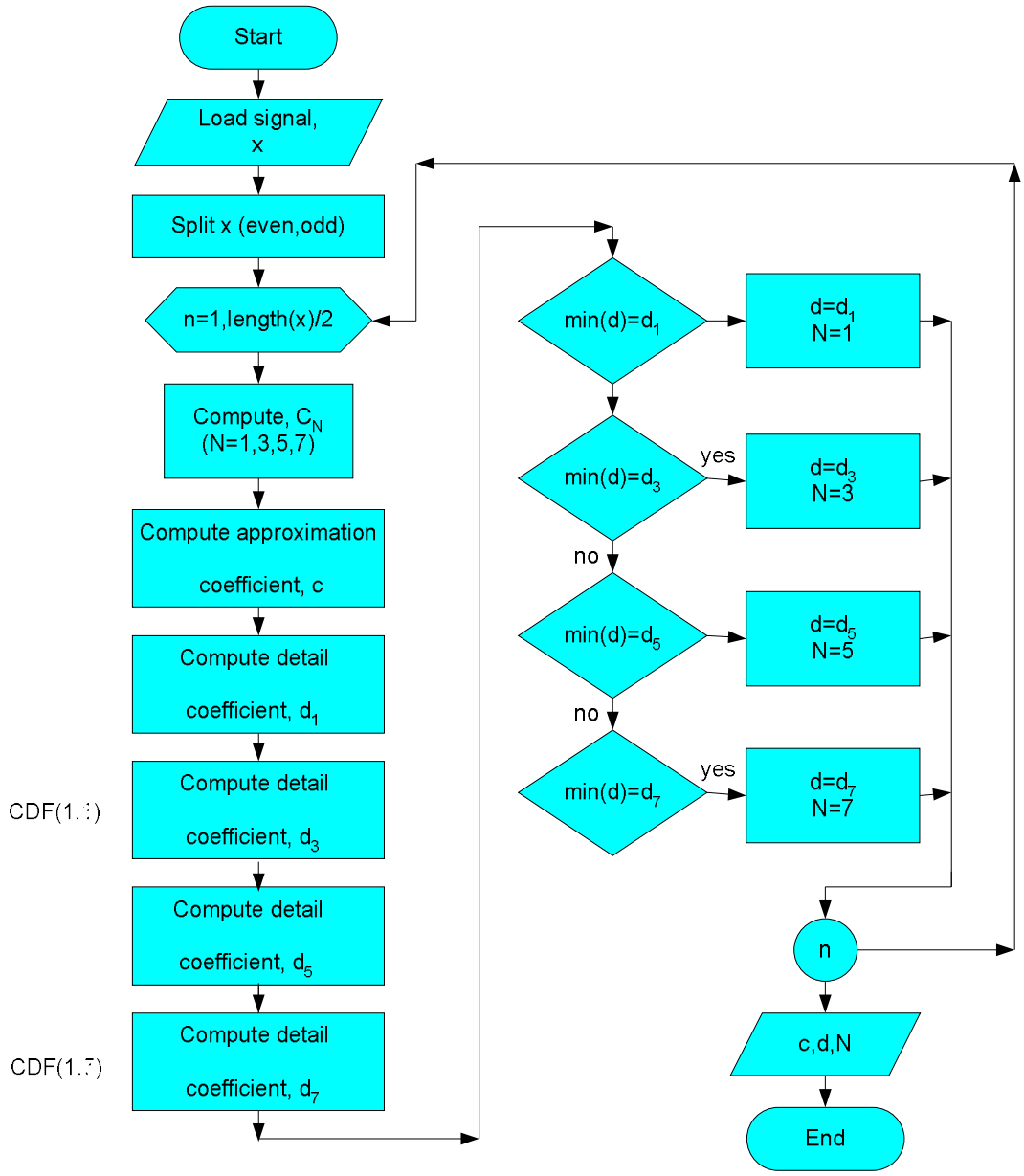

Figure 3.5 Flowchart of proposed 1-D adaptive (forward) lifting scheme

3.4 Proposed Perceptual Filterbank

The perceptual filterbank has been designed as follows: The WPD tree (five levels) has been constructed via the adaptive lifting scheme and adjusted according to critical bands of the human auditory system. Since, 8 kHz sampling frequency has been chosen, 4 kHz bandwidth corresponding to 17 critical bands given in the Table 1 has been used. The Table 1 has been presented based on [75-76].

For 8 kHz sampling rate, the following equation have been used to calculate the frequency bandwidth values corresponding to nodes of the full wavelet packet decomposition (WPD) tree constructed via the proposed adaptive lifting scheme.

Table 3.1 Critical-band frequencies of human auditory system

Critical Band Number	Center Frequency [Hz]	Critical Bandwidth CBW [Hz]	Lower Frequency [Hz]	Upper Frequency [Hz]
1	50	100	0	100
2	150	100	100	200
3	250	100	200	300
4	350	100	300	400
5	450	110	400	510
6	570	120	510	630
7	700	140	630	770
8	840	150	770	920
9	1000	160	920	1080
10	1170	190	1080	1270
11	1370	210	1270	1480
12	1600	240	1480	1720
13	1850	280	1720	2000
14	2150	320	2000	2320
15	2500	380	2320	2700
16	2900	450	2700	3150
17	3400	550	3150	3700
18	4000	700	3700	4400
19	4800	900	4400	5300
20	5800	1100	5300	6400
21	7000	1300	6400	7700
22	8500	1800	7700	9500
23	10500	2500	9500	12000
24	13500	3500	12000	15500

$$bw(j,p) = 2^{-j}(F_s/2) \tag{3.49}$$

Where, $bw(j,p)$ represents the frequency bandwidth [Hz] value corresponding to node (j,n) of the full WPD tree. Note that, p is also a function of j.

$j = (0,1,2,...5)$, (j : number of decomposition levels)
$p = 0,...(2^j - 1)$, (p : position)

The frequency bandwidths $bw(j,p)$ of the full WPD tree has been adjusted, by choosing the closest values. The (five levels) full WPD tree and CB-WPD tree (sub-tree) (bold lines only) constructed via the proposed adaptive lifting scheme is given Figure.5. The full tree corresponds to a uniform filterbank however; the CB-WPD tree (sub-tree) corresponds to a nonuniform filterbank which is called as "perceptual filterbank" in this thesis. Where y denotes the noisy input signal, ($j = 1,2,3,...5$) denotes the number of decomposition levels and $(yc_1.....yc_{17})$ represent critical subbands.

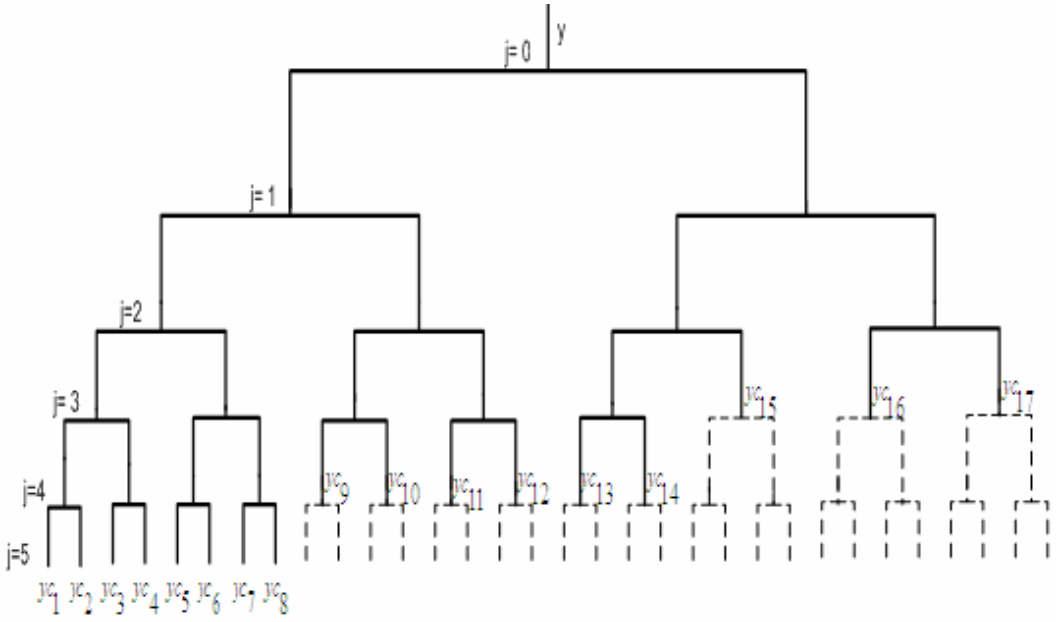

Figure 3.6 WPD tree (bold & dashed lines) and CB-WPD sub-tree (bold lines only) corresponding to proposed perceptual filterbank

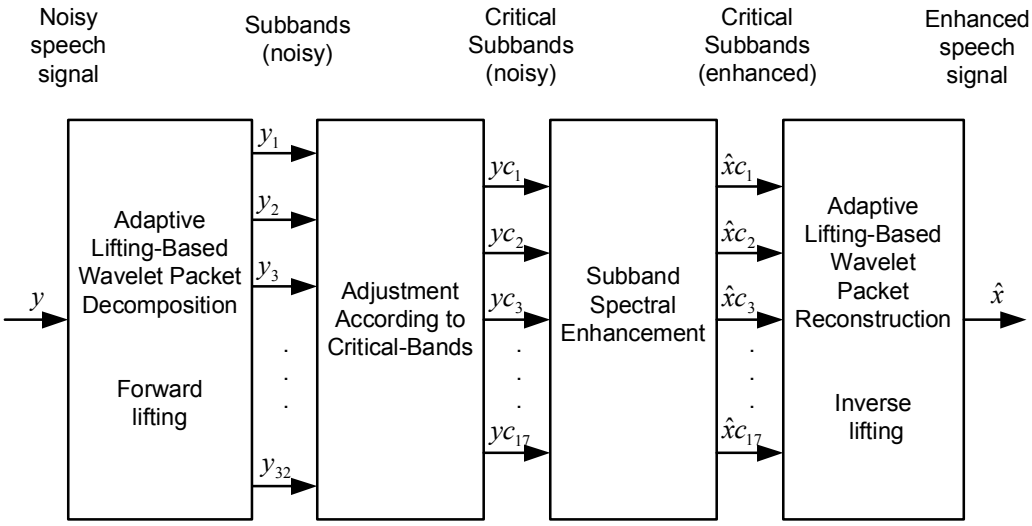

Figure 3.7 Overall block diagram of proposed speech enhancement method with CB-WPD in adaptive lifting based wavelet (packet) domain and MMSE-based estimators.

CHAPTER 4

IMAGE ENHANCEMENT

4.1 Introduction

Images are often corrupted by noises caused by decoding errors or noisy channels. Both have potential of degrading the image quality. The human perception is very sensitive to noise due to its strong amplitude. Removal of such noises without blurring edges and important details of image is still an important issue in image processing.

The image enhancement methods can be broadly divided into two groups. The spatial domain methods based on direct manipulation of the pixels in an image and frequency domain methods based on modification of the Fourier transform of an image [42]. Wavelet based shrinkage or thresholding methods [24-26] have also received considerable attention in image denoising applications in last decades.

Among the widely used spatial domain linear and nonlinear techniques are Wiener filter, Mean and Median filters. Classical Wiener filter which is a linear technique provides mathematical simplicity but has the disadvantage of blurring edges [43, 77].

The nonlinear mean filter cannot remove positive and negative impulse noises simultaneously. The median filter performs well, but it fails when the probability of impulse noise occurrence is high [78].

Median filter is one of the order-statistic nonlinear filters. In the median filtering, a pixel (whether corrupted by noise or not) is replaced by its local median value within a window. Although the median filtering can be successfully used to suppress impulsive noise while preserving edges, it often fails to provide sufficient smoothing of non-impulsive noise [52,79].

Over the past decade, wavelet transform has received considerable attention between the scientists and researchers since it provides good time-frequency localization. The

most popular wavelet based denoising techniques are wavelet thresholding or shrinkage proposed by Donoho and Johnstone [24-26], where the noisy image is wavelet decomposed into subbands and noise is removed by killing coefficients that are smaller than some threshold. It is a simple and effective method however the choice of thresholding functions and threshold values are critical in enhancement schemes. The threshold value in universal thresholding [25] depends on the number of data samples. If the number of data samples is too small, the enhanced image remains still noisy, on the other hand, if it is too large then the important details of the signal disappears.

The classical wavelet transform is constructed by shifting and scaling of a fixed wavelet function based on Fourier transform. An alternative way of constructing biorthogonal wavelets is the lifting scheme. The lifting scheme is a spatial domain method which has some advantages over the classical lifting scheme [32, 33]. The advantages of the classical lifting scheme are:

- It is a spatial domain method,
- It is easier to implement,
- It allows faster and in-place calculations,
- It allows nonlinear, adaptive, irregularly sampled and integer to integer wavelet transforms,
- It is easier to obtain inverse transform.

Furthermore, any wavelet transform can be factored into lifting stages [34].

The adaptation of lifting filters to the signal structure leads to “adaptive lifting schemes”. The motivation behind introducing adaptivity into the lifting steps is that, choosing better lifting filters (predict or update filters) may lead to more efficient signal representations.

The CDF(1,N) group lifting filters have been exploited in the proposed adaptive lifting scheme as also used in the previously proposed work [37,38]. The prediction stage is adapted to the signal structure based on minimizing the detail coefficient (or

prediction error) . The original work is based on edge detection and "edge avoiding prediction" method. However, we proposed a new adaptive prediction method in the proposed adaptive lifting scheme. The procedure of edge avoiding prediction proposed in [37,38] is given in the Chapter 3, section 3.2.

In this thesis, we proposed an image enhancement method based on space adaptive (update-first) 2-D adaptive lifting scheme. Spatial domain filters (median and adaptive Wiener filters) and wavelet based thresholding methods (Visu Shrink, Bayes Shrink, Normal Shrink and Level-dependent thresholding (LDP)) are used for subband image enhancement.

4.2 Image Enhancement Methods

In this thesis, we have used special domain median and adaptive Wiener filters [43, 77-79] and wavelet-based thresholding methods Visu Shrink, Bayes Shrink, Normal Shrink and (LDP) methods in the space adaptive 2-D lifting-scheme [24-26].

4.3 Spatial Domain Methods

The spatial domain methods are based on direct manipulation of the pixels in an image. The widely used spatial domain linear and nonlinear techniques are Wiener filter, Mean and Median filters [43, 77-79].

4.3.1 Spatial Domain Adaptive Wiener Filter

Let us consider an image $x(i,j)$ is corrupted with Gaussian white noise $b(i,j)$ then the noisy image can be expressed as follows,

$$y(i,j) = x(i,j) + b(i,j) \tag{4.1}$$

where $x(i,j)$ and $y(i,j)$ denote the original image and noisy images respectively. Here we assume that, the noise is stationary with zero mean and σ_b^2 variance and uncorrelated with the original image. If the original image $x(i,j)$ is considered locally stationary within a small region, it can be modeled by:

$$x(i,j) = m_x + \sigma_x b(i,j) \tag{4.2}$$

Where, m_x and σ_x are local mean and standard deviation respectively and $b(i,j)$ is white noise with zero mean and unit variance. Within the local region, spatial domain wiener filter (SDWF) that minimizes the mean square error (MSE) between the original image $x(i,j)$ and the enhanced image $\hat{x}(i,j)$ is given as follows.

$$\hat{x}(i,j) = m_x + \frac{{\sigma_x}^2}{{\sigma_x}^2 + {\sigma_b}^2}(y(i,j) - m_x) \tag{4.3}$$

$m_x(i,j)$ and $\sigma_x(i,j)$ are estimated from the noisy observation signal and updated at each pixel in (4.4).

$$\begin{aligned} \hat{m}_x(i,j) &= \frac{1}{(2m+1)(2n+1)} \sum_{k=i-m}^{i+m} \sum_{l=j-n}^{j+n} y(k,l) \\ {\hat{\sigma}_y}^2(i,j) &= \frac{1}{(2m+1)(2n+1)} \sum_{k=i-m}^{i+m} \sum_{l=i-n}^{i+n} [y(k,l) - \hat{m}_x(i,j)]^2 \\ {\hat{\sigma}_x}^2(i,j) &= \max\left\{0, {\hat{\sigma}_y}(i,j) - {\sigma_b}^2\right\} \end{aligned} \tag{4.4}$$

Substituting $\hat{m}_x(i,j)$ and $\hat{\sigma}_x(i,j)$ into (3) we get,

$$\hat{x}(i,j) = \hat{m}_x(i,j) + \frac{{\hat{\sigma}_x}^2(i,j)}{{\hat{\sigma}_x}^2(i,j) + {\sigma_b}^2}[y(i,j) - \hat{m}_x(i,j)] \tag{4.5}$$

The filter size $(2m+1)(2n+1)$ is fixed over the entire observed image and is generally chosen as 5x5 [43].

4.3.2 Spatial Domain Median Filter

The standard spatial domain median filter is based on sliding a window of odd length over an image, ranking of the pixels in a neighborhood of size (3x3, 5x5, 7x7...) according to brightness within the input window. Center pixel in the window is then replaced by the median of the pixels within the window [52]. In case of the spatial

domain mean filter [77], the center pixel in the window is replaced by the mean of the pixels within the window.

4.4 Image Denoising Based on Wavelet Thresholding

One of the important features of the wavelet based denoising is that, it maps the white Gaussian noise in the signal domain and in the transform domain. Since the noise energy is heavily concentrated in the high band detail coefficients (while signal energy is concentrated in the low band approximate coefficients), applying a thresholding technique to the detail coefficients enables removing noise from the image. There are two types of thresholding techniques called as "soft thresholding" and "hard thresholding". The soft thresholding is more preferred than the hard thresholding since it causes fewer artifacts [47]. The shrinkage methods used in our applications are all based on soft thresholding.

The soft thresholding (shrinkage) function of the wavelet coefficient matrix w with threshold T can be expressed as follows:

$$\begin{aligned} \lambda_{soft} &= w - T \quad if \quad w \geq T \\ &= 0 \quad if \quad |w| < T \\ &= w + T \quad if \quad w \leq -T \end{aligned} \tag{4.6}$$

Similarly, the hard thresholding function is expressed as follows:

$$\begin{aligned} \lambda_{hard} &= w \quad if \quad w \geq T \\ &= 0 \quad otherwise \end{aligned} \tag{4.7}$$

The most commonly used shrinkage functions exist in the literature are as follows

4.4.1 Visu Shrink

Visu Shrink (Visually calibrated adaptive smoothing) [25] is the most popular shrinkage method used for image denoising. The 'universal' threshold selection is given as $\sigma_b \sqrt{2 \log M}$, where σ_b is noise variance and M is number of sample/pixel.

Visu Shrink is known to be providing overly smoothed images. The noise standard deviation is estimated as the median absolute deviation (MAD) of the finest scale wavelet coefficient matrix (HH1).

$$\hat{\sigma}_b = \frac{median(|Y_i|)}{0.6745} \quad where\ Y_i \in subband\ HH1 \tag{4.8}$$

4.4.2 Bayes Shrink

Bayes Shrink is an adaptive data-driven threshold for image denoising via wavelet soft thresholding [50]. It uses the threshold,

$$T = \frac{\hat{\sigma}_b^{\,2}}{\hat{\sigma}_x} \tag{4.9}$$

$\hat{\sigma}_b^{\,2}$ is the estimated noise variance and $\hat{\sigma}_x^{\,2}$ is the estimated signal variance under consideration.

$$\hat{\sigma}_x = \sqrt{\max(\hat{\sigma}_Y^{\,2} - \hat{\sigma}_b^{\,2}, 0)} \tag{4.10}$$

Where

$$\hat{\sigma}_Y = \frac{1}{N}\sum_{k=1}^{N} Y_k^{\,2} \tag{4.11}$$

$\hat{\sigma}_Y$, is the estimation of the standard deviation of the noisy observation coefficient matrix Y_k on the subband under consideration and N is the number of pixels in the Y_k.

4.4.3 Normal Shrink

Normal Shrink [48] performs soft thresholding with the data driven subband dependent threshold T, which is given as,

$$T = \frac{\beta \hat{\sigma}_b^{\,2}}{\hat{\sigma}_Y} \tag{4.12}$$

Where, the scale parameter β is computed once for each scale using the following equation.

$$\beta = \sqrt{\log(\frac{L_k}{J})} \tag{4.13}$$

Where, L_k is the length of the subband at k*th* scale and J is the total number of scales $(k = 1,2,3....J)$. The standard deviation estimates of noise and noisy observations at k*th* scale ($\hat{\sigma}_b$ and $\hat{\sigma}_Y$) were previously given in (4.8) and (4. 11).

4.5 2-D lifting Construction

The general 2-D separable lifting implementation of an image is performed by first applying 1-D lifting wavelet transform horizontally on rows and then vertically on columns (or vice versa). Four subbands (LL1, HL1, LH1, and HH1) can be acquired after 1-level, 2-D lifting decomposition of a given image W. The Figure.4.1 demonstrates the 2-D lifting construction procedure exploited in our applications.

Decomposition can be further continued by applying the same procedure on sub-image (LL1) similarly as in the case of original image W. Thus, after 2-level lifting decomposition of sub-image (LL1), four subband coefficients (LL2, HL2, LH2, and

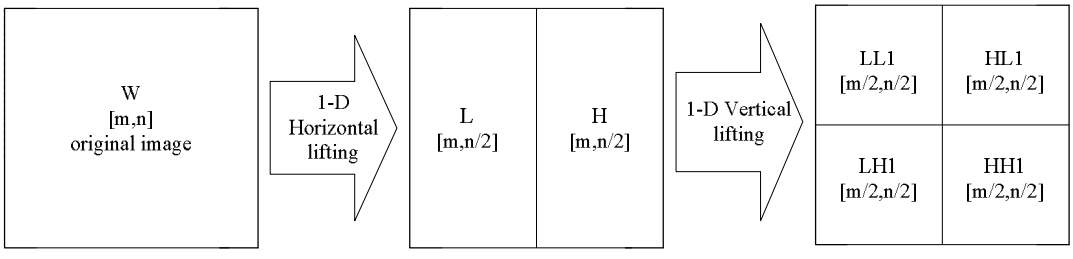

Figure 4.1 Procedure of the 2-D lifting scheme

HH2) are obtained. 2-levels, 2-D lifting wavelet decomposition stages for an image (W) are given in Figures 4.2- 4.3.

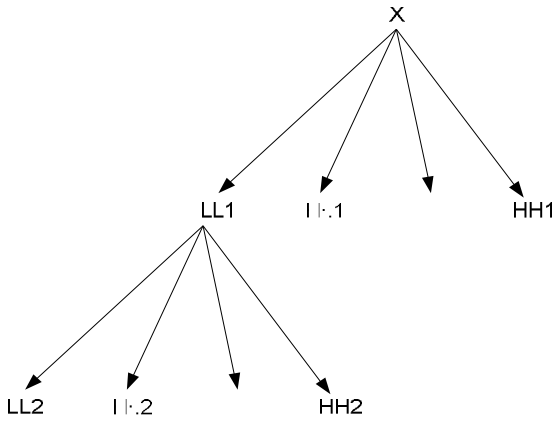

Figure 4.2 Two levels, 2-D lifting wavelet decomposition tree

In the 1-level decomposition, LL1 is the approximation coefficients sub-matrix; HL1, LH1 and HH1 are horizontal, vertical and diagonal detail coefficients sub-matrices.

After subband image enhancement, the inverse 2-D lifting scheme can easily be constructed by applying a reverse procedure to the one given in Figure.4.1.

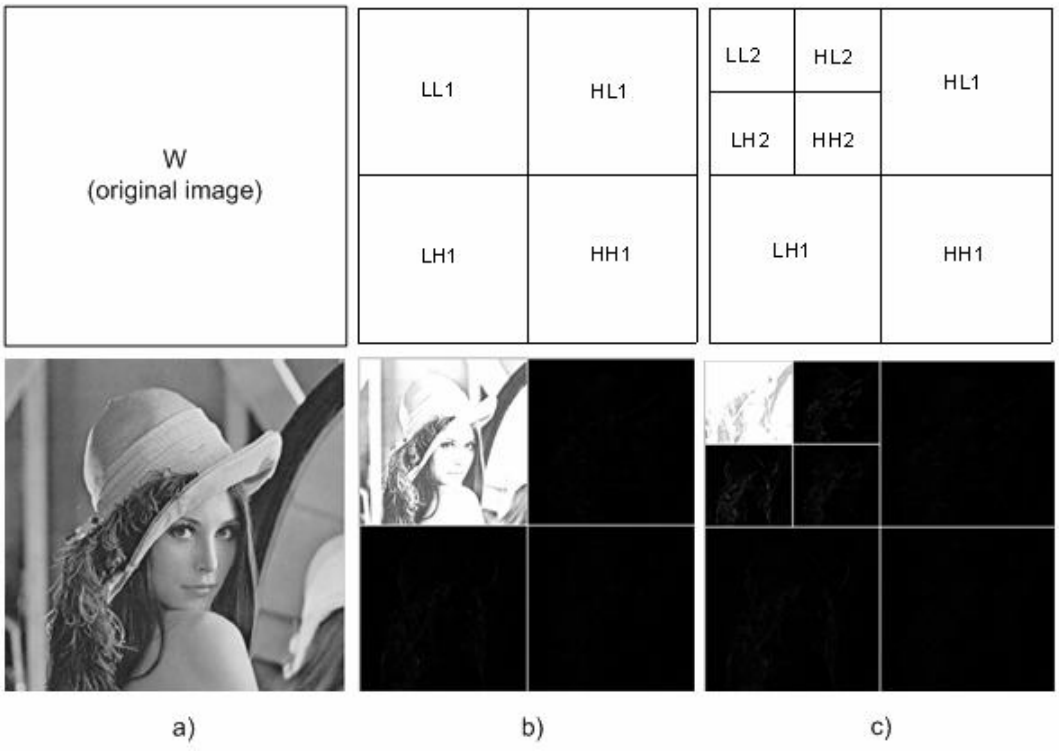

Figure 4.3 2-D lifting wavelet decomposition a) original image
b) 1-level decomposition c) 2-levels decomposition

4.6 Proposed Adaptive 2-D Lifting Scheme

The adaptive 2-D lifting scheme proposed in this thesis is based on adaptive 1-D lifting scheme (given in Chapter 3, Section 3.2). The adaptive 1-D lifting scheme is applied horizontally (on rows) and then vertically (on columns) of a given noisy test image using the procedure given in Section 4.5. The procedure and motivations behind the adaptive prediction method "edge avoiding prediction" proposed in [37,38] and our proposed adaptive prediction method are given in the Chapter 3, Section 3.2 in detail.

CHAPTER 5

PERFORMANCE EVALUATION AND EXPERIMENTAL RESULTS

5.1 Introduction

This chapter describes the performance evaluation of the proposed speech and image enhancement algorithms. Objective quality evaluation tests signal to noise ratio (SNR), segmental signal to noise ratio (SegSNR), Itakura-Saito distance (IS) tests and perceptual evaluation of speech quality (PESQ-MOS) (which predicts the subjective MOS results) are employed for performance assessments of the developed speech enhancement algorithms. All the noise signals (white Gaussian noise (WGN), F16 cockpit noise, car interior noise and babble noise) have been taken from Noisex-92 database. The noisy signals have been obtained by adding noise the original (clean) signals at noise levels in the range [-5, 10] dB SNR. The performance results are obtained by averaging the performance results of four different speech sentences (half spoken by females and half by males) taken from Harward speech database. The experimental results (speech signal waveforms and spectrograms) obtained for the English sentence "*A pot of tea helps to pass the evening*" spoken by a male speaker are demonstrated in this chapter. The estimators magnitude spectral subtraction (MSS) and MMSE based methods modified short-time spectral amplitude (Mod-STSA) estimator, multiplicatively modified-log spectral amplitude (MM-LSA) estimator, Wiener filter (WF) and modified Wiener filter (Mod-WF) have been tested and their results have also been demonstrated. All the tested methods are used in the adaptive lifting-wavelet domain using (CB-WPD) or perceptual filterbank as given in detail in Chapter 3, Section 3.3.

The well known objective quality evaluation test (PSNR) is used for performance evaluation of the image enhancement algorithms proposed in this thesis. The proposed image enhancement methods are spatial domain methods (median and adaptive wiener filters and wavelet-based thresholding methods (Visu Shrink, Bayes Shrink, Normal Shrink and Level-dependent thresholding).

5.2 Performance Evaluation for Speech Enhancement Algorithms

The performance of proposed speech enhancement method is tested by using the objective and subjective evaluation tests. The objective tests are the signal to noise ratio (SNR), segmental signal to noise ratio (SegSNR) and Itakura-Saito (IS) distance. The subjective evaluation is performed by using (PESQ-MOS) test. The original, noisy and enhanced speech signal waveforms and corresponding spectrograms are also demonstrated in this chapter.

5.2.1 Objective Evaluation Methods

5.2.1.1 Signal to Noise Ratio (SNR)

The following equation is computed for evaluation of SNR results of noisy and enhanced speech signals. The SNR is calculated in decibels (dB).

$$SNR_{dB} = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1}[x(n)]^2}{\sum_{n=0}^{N-1}[\hat{x}(n)-x(n)]^2}\right) \tag{5.1}$$

Where $x(n)$ denotes original signal, $\hat{x}(n)$ denotes processed (enhanced) signal, and N denotes number of samples in original speech signal. The SNR (sometimes called as global or instantaneous SNR) is the most common measure of speech quality; however it is not well correlated with human auditory system.

5.2.1.2 Segmental Signal to Noise Ratio (SegSNR)

The frame based segmental SNR is a reasonable measure of speech quality. It is formed by averaging frame level estimates as follows**.**

$$SegSNR = \frac{1}{M}\sum_{m=0}^{M-1}10\log_{10}\left[\frac{\sum_{n=Nm}^{Nm+N-1}x^2(n)}{\sum_{n=Nm}^{Nm+N-1}[\hat{x}(n)-x(n)]^2}\right] \tag{5.2}$$

Where $x(n)$ denotes original signal, $\hat{x}(n)$ denotes enhanced signal, M denotes number of frames and N denotes number of samples in each short time frame. Since the frames with SNRs above (35 dB) do not reflect the human perceptual differences, they are generally replaced with (35 dB). Similarly, during periods of silence, SNR values may become very negative since signal energies are small. These frames also do not truly reflect the perceptual contributions of the signal. Therefore a lower threshold is often set for much realistic frame based SNR calculation. Here**,** we have chosen (-10 dB) SNR as the lower threshold.

5.2.1.3 Itakura-Saito Distance (IS)

For $\vec{a}_x$ and $\vec{a}_{\hat{x}}$ being linear prediction coefficients (LPC) vector of clean and processed speech signal respectively and R_x denotes (R+1) x (R+1) (Toeplitz) autocorrelation matrix. The R is the order of LPC analysis. The Itakura-Saito distance (IS) measure is given by:

$$IS(\vec{a}_{\hat{x}}, \vec{a}_x) = \left[\frac{\sigma_x^2}{\sigma_{\hat{x}}^2}\right]\left[\frac{\vec{a}_{\hat{x}} R_x \vec{a}_{\hat{x}}^T}{\vec{a}_x R_x \vec{a}_x^T}\right] + \log\left(\frac{\sigma_{\hat{x}}^2}{\sigma_x^2}\right) - 1 \tag{5.3}$$

Where, $\sigma_{\hat{x}}^2$ and σ_x^2 represent the all-pole gains for the processed and original speech frame respectively. The IS measure is well correlated to subjective results. Note that, the lower the IS measure (close to zero) the better the is perceived quality of the enhanced speech.

5.2.2 Subjective Evaluation Methods

The subjective evaluation methods include the methods which focus on speech intelligibility and those which focus on overall quality [80]. The most often used intelligibility tests are Modified Rhyme Test (MRT) and Diagnostic Rhyme Test (DRT). The quality tests aim to obtain an overall idea on the perceptual characteristics of speech such as intelligibility, acceptability, naturalness, etc. The most frequently used quality test is Mean Opinion Score (MOS) where a five-point score as given in the Table. 5.1. The MOS test gives an idea of mean impression of

listeners. In order to obtain a reliable result, many listeners must be present since opinion of single listeners may be much different.

Table: 5.1 MOS quality score

Quality of Speech	Scale
Bad	1
Poor	2
Fair	3
Good	4
Excellent	5

The subjective tests provide very reliable results however, they are time and money consuming and it is difficult to reproduce in the same conditions. Hence, it is desirable to develop objective measures based on characteristics of the speech signal for prediction of subjective results. This kind of measures such as Log-Area Ratio (LAR) and Perceptual Speech Quality measure (PSQM) are the most correlated ones with subjective results [81]. The ITU standard PESQ (ITU-T Recommendation P.862, 2001) [82] is an advanced version of the PSQM which predicts subjective MOS for a wide range of speech distortions in transmission systems.

5.3 Objective Evaluation Results

In this section we obtained segmental SNR (SegSNR) improvement and Itakura-Saito distance (IS) measure for various noise types (white Gaussian noise (WGN), F16 cockpit noise, car interior noise and babble noise at various noise levels in the range [-5, 10] dB SNR. All noise materials have been taken from Noisex-92 database and speech materials from Harward sentences database. The proposed algorithms have been tested by four different speech sentences (two spoken by a female and two spoken by a male speaker). Objective evaluation results have been obtained by using following enhancement algorithms: Magnitude spectral subtraction (MSS), Wiener filter (WF), modified-Wiener filter (Mod-WF), modified-short-time spectral amplitude (Mod-STSA) estimator, multiplicatively modified-log spectral amplitude (MM-LSA) estimator. The modifications have been performed by taking into

account the probability of speech absence, where the probability of speech absence $q(k,l)$ is adaptively estimated). Figures 5.1 to 5.4 demonstrate the (SegSNR) improvement and (IS) measures for various noise types and enhancement algorithms. Noisy input signals have been obtained by adding the noise signals to the original speech signals at [-5, 10] dB SNR. The speech signal waveforms and spectrograms of the original, noisy and the enhanced speech signals are demonstrated in the Figures 5.4-5.8.

For all the experiments and objective quality tests, 32 ms Hamming window with % 50 overlap and 256 point DFT is applied to a speech signal with 8 kHz sampling rate. The first ten segments were assumed to be noise only frames. The noise PSD is estimated through these segments and updated during the consecutive noise frames.

All the computations are performed using a PC Intel Pentium IV, 1.73 GHz processor and 1.99 GB RAM.

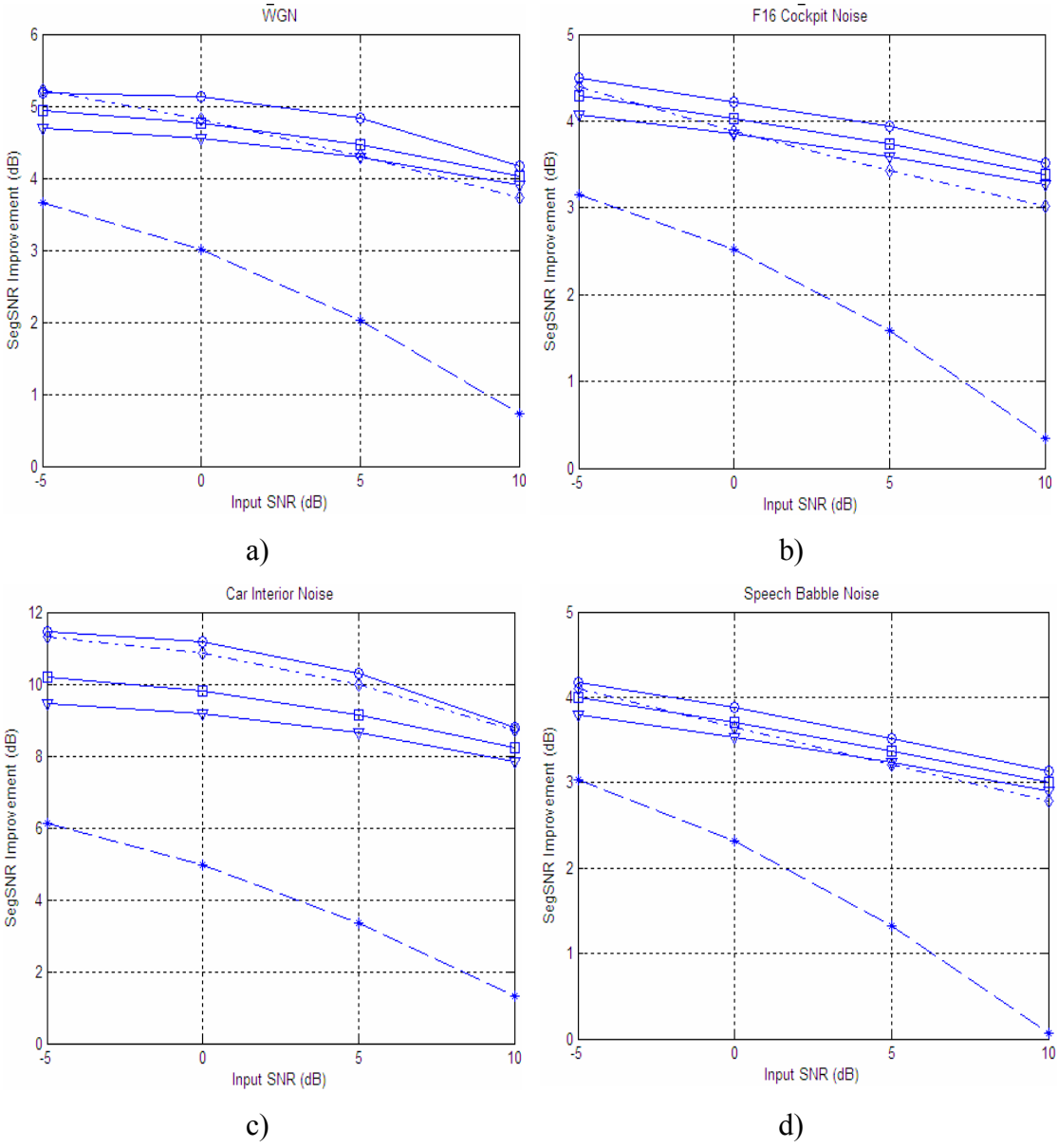

Figure 5.1 SegSNR improvement vs. input SNR for the noise types: a) WGN, b) F16 cockpit noise, c) Car interior noise, d) Speech babble noise and the estimators: MSS (dashed, *), WF (dash-dot, ◊), Mod-STSA (bold, ▽), MM-LSA (bold, □) and Mod-WF (bold, o) using an English sentence "*A pot of tea helps to pass the evening*" spoken by a male speaker.

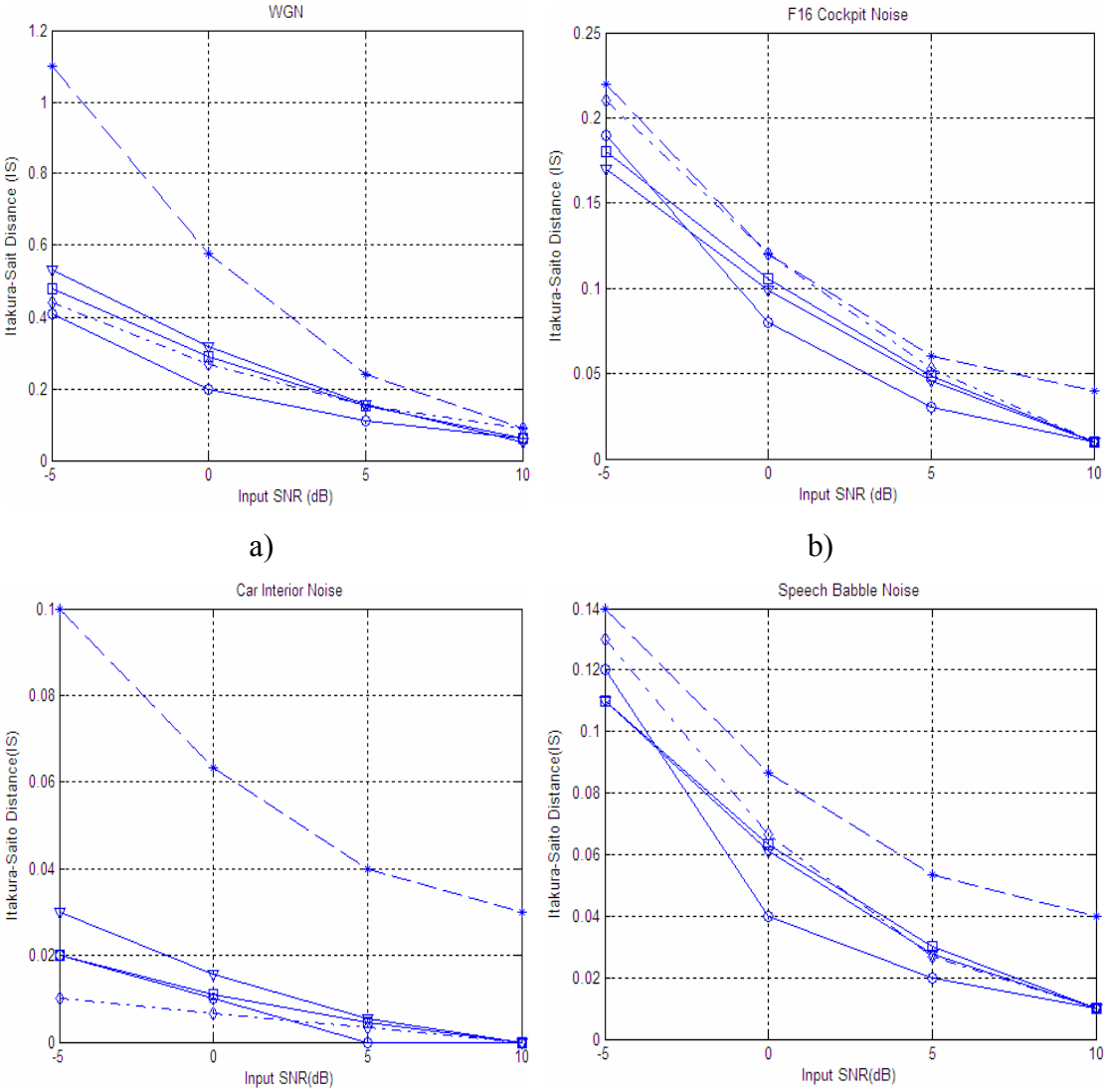

Figure 5.2 IS measure vs. input SNR for the noise types: a) WGN, b) F16 cockpit noise, c) Car interior noise, d) Speech babble noise and the estimators: MSS (dashed, *), WF (dash-dot, ◊), Mod-STSA (bold, ▽), MM-LSA (bold, □) and Mod-WF (bold, o) using an English sentence "*A pot of tea helps to pass the evening*" spoken by a male speaker.

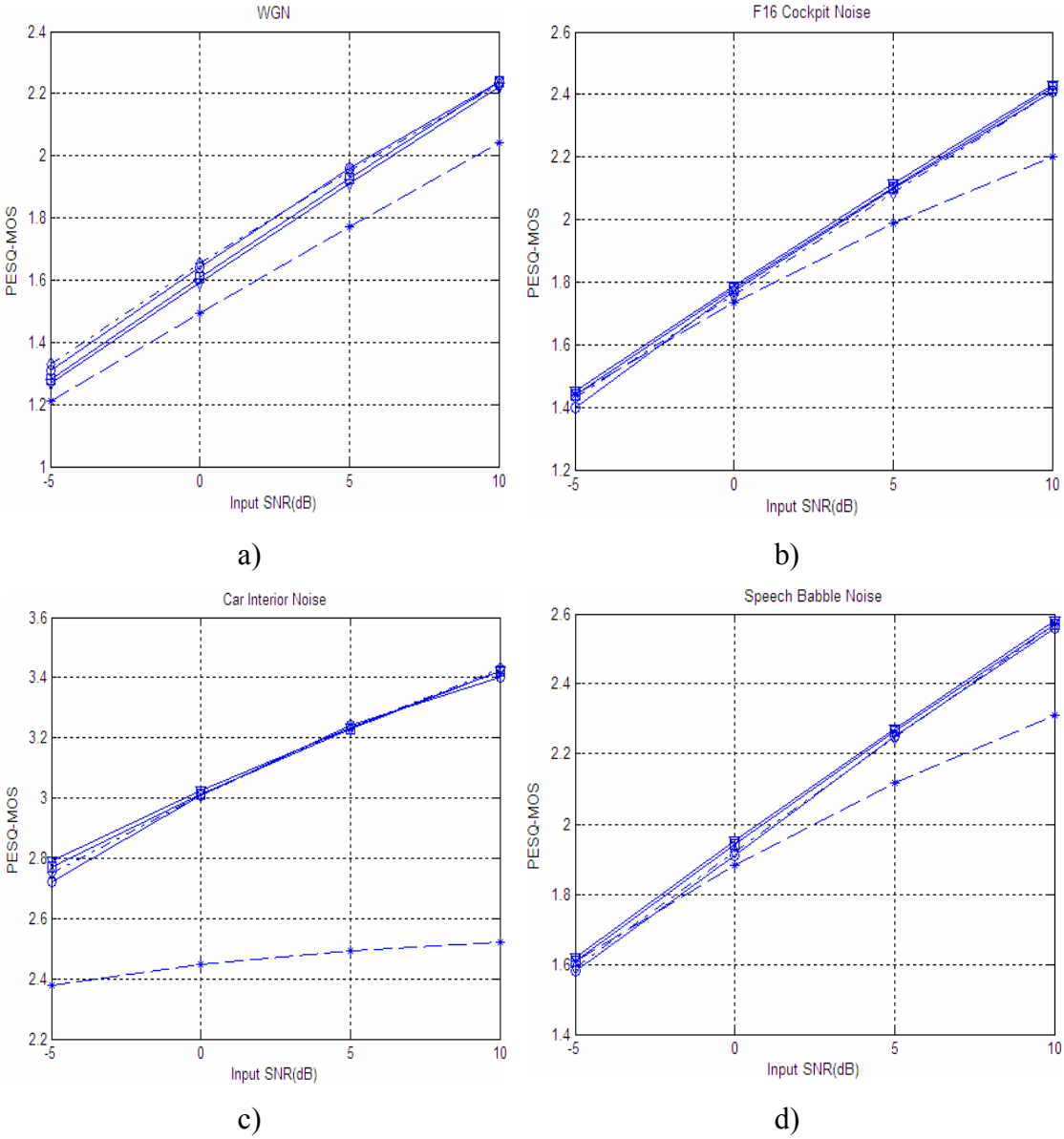

Figure 5.3 PESQ-MOS vs. input SNR for the noise types: a) WGN, b) F16 cockpit noise, c) Car interior noise, d) Speech babble noise and the estimators: MSS (dashed, *) , WF (dash-dot, ◊), Mod-STSA (bold, ▽), MM-LSA (bold, □) and Mod-WF (bold, o) using an English sentence "*A pot of tea helps to pass the evening*" spoken by a male speaker.

5.4 Experimental Results of Speech Enhancement Algorithms

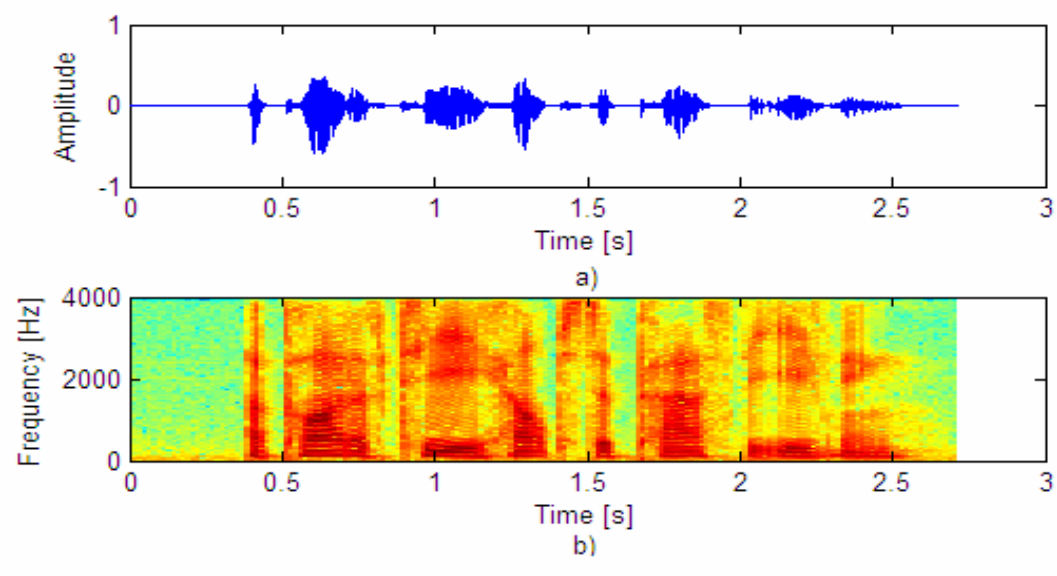

Figure 5.4 Original (clean) speech signal "*A pot of tea helps to pass the evening*" spoken by a male speaker a) signal waveform b) signal spectrogram

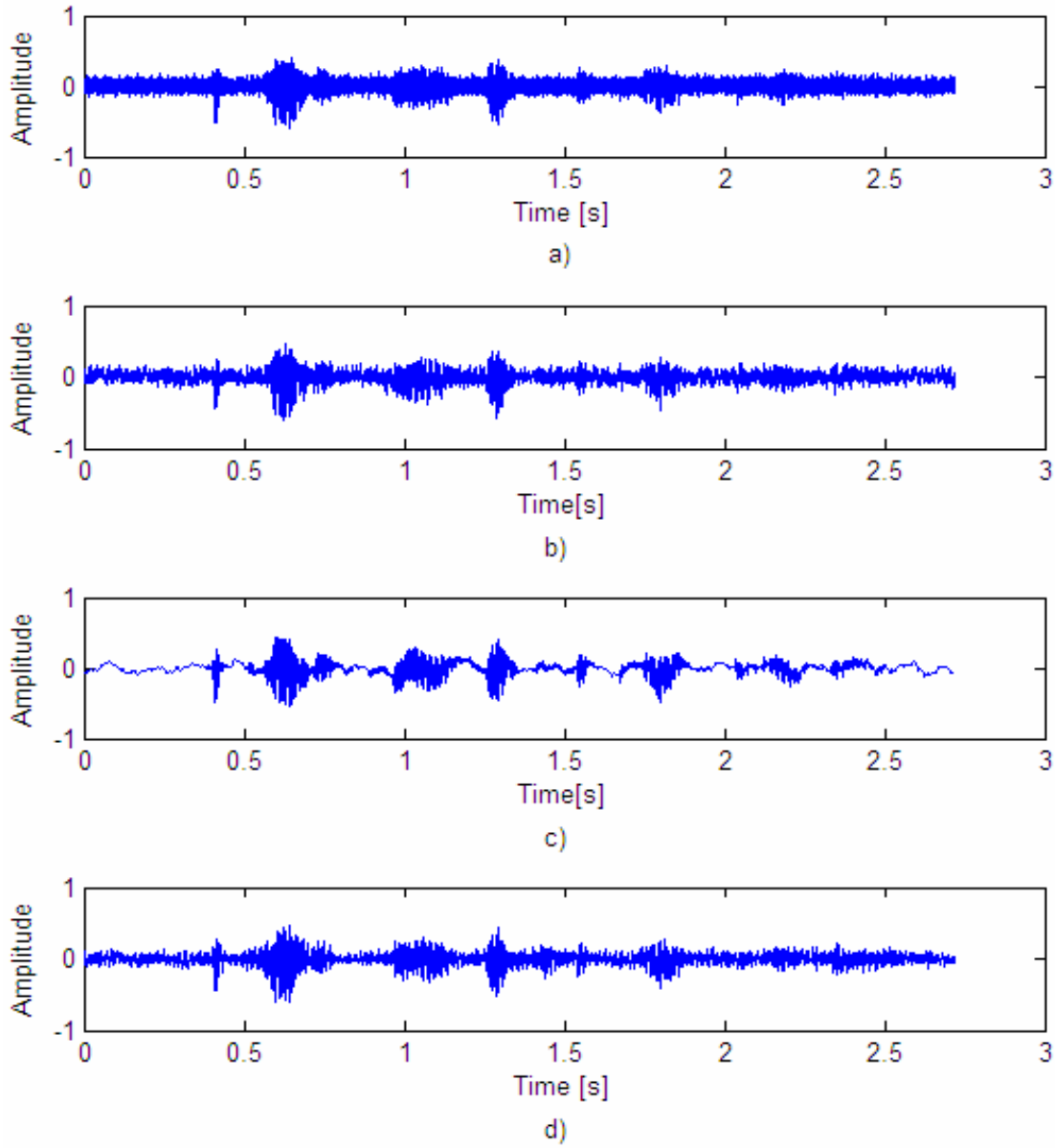

Figure 5.5 Noisy speech signals obtained at 0 dB SNR by adding a) WGN (SegSNR = -5.29 dB, IS=1.85, PESQ-MOS= 1.43), b) F16 cockpit noise (SegSNR = -5.16 dB, IS=0.25, PESQ-MOS=1.78), c) Car interior noise (SegSNR = -3.82 dB, IS=0.09, PESQ-MOS=3.82), d) Speech babble noise (SegSNR = -5.18 dB, IS=0.06, PESQ-MOS=2.05). The original speech signal is "*A pot of tea helps to pass the evening*" spoken by a male speaker.

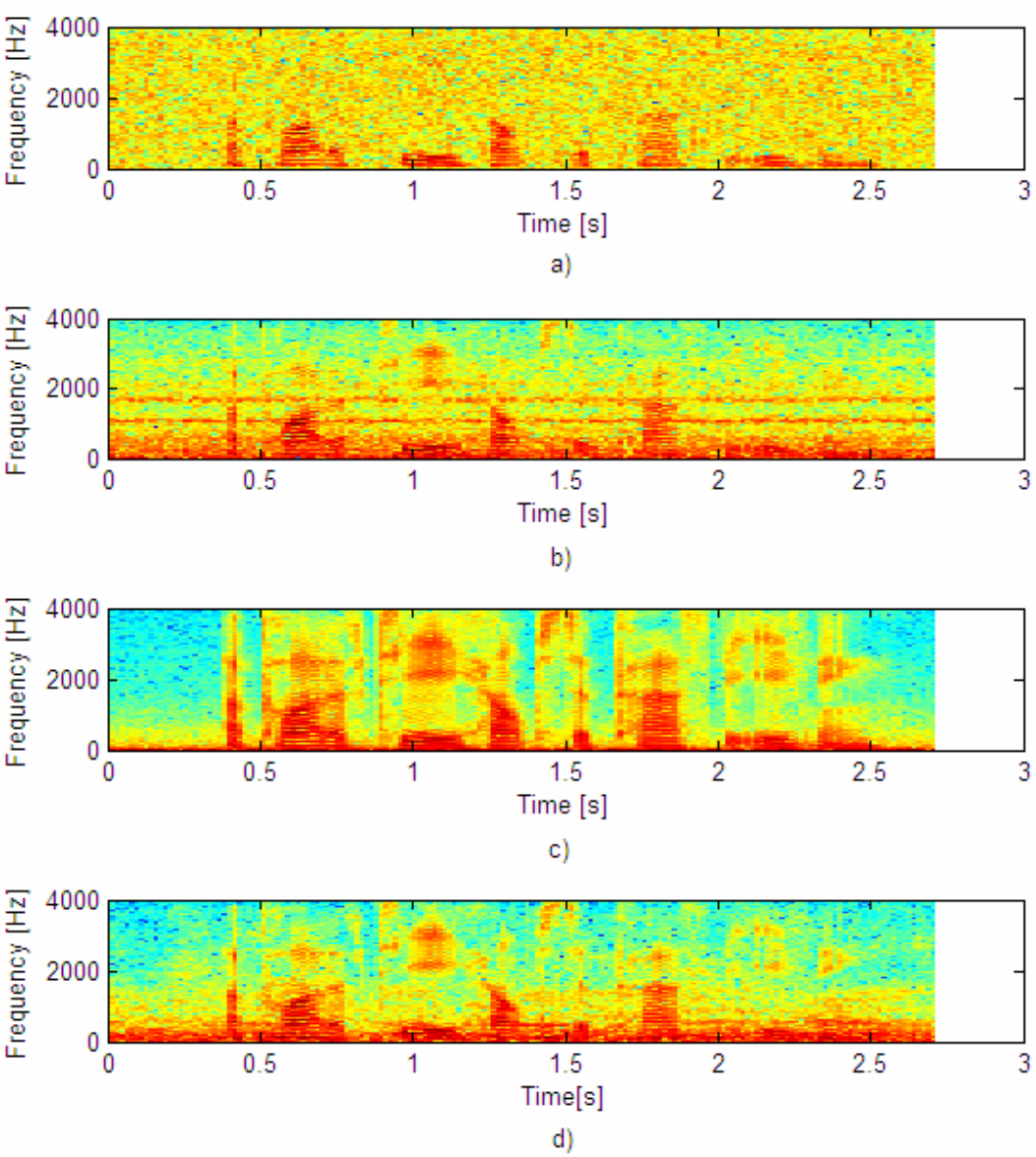

Figure 5.6 Spectrograms of the noisy speech signals obtained at 0 dB SNR by adding: a) WGN (SegSNR = -5.29 dB, IS=1.85, PESQ-MOS= 1.43), b) F16 cockpit noise (SegSNR = -5.16 dB, IS=0.25, PESQ-MOS=1.78), c) Car interior noise (SegSNR = -3.82 dB, IS=0.09, PESQ-MOS=3.82), d) Speech babble noise (SegSNR = -5.18 dB, IS=0.06, PESQ-MOS=2.05). The original speech signal is "*A pot of tea helps to pass the evening*" spoken by a male speaker.

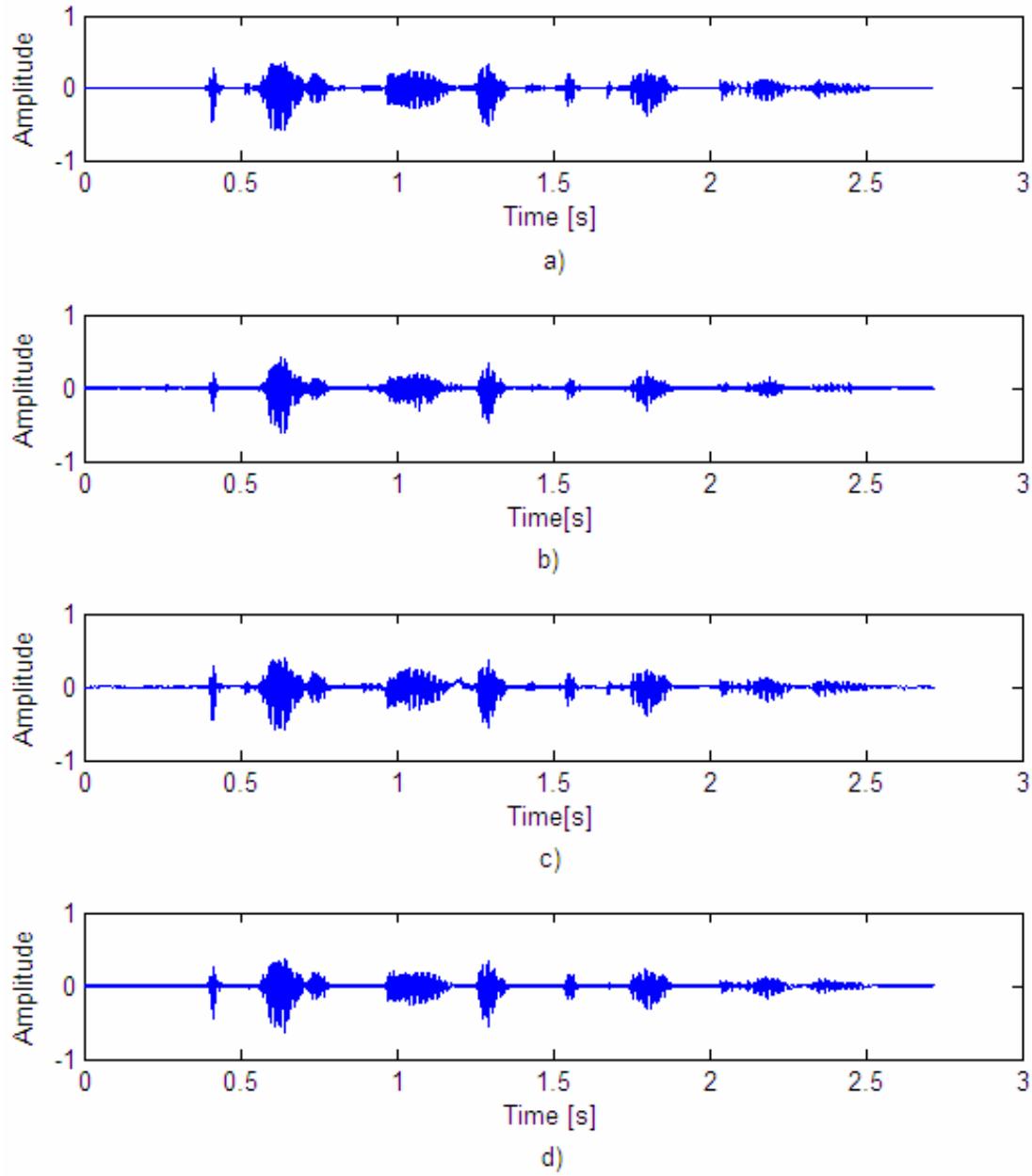

Figure 5.7 Enhanced speech signals using Mod-WF via the proposed adaptive lifting scheme: a) WGN (SegSNR improvement=5.32 dB, IS=0.14, PESQ-MOS= 1.80), b) F16 cockpit noise (SegSNR improvement=4.78 dB, IS=0.06, PESQ-MOS=2.00), c) Car interior noise (SegSNR improvement = 11.69 dB, IS=0.003, PESQ-MOS=3.34), d) Speech babble noise (SegSNR improvement = 3.82 dB, IS=0.03, PESQ-MOS=2.10). The original speech signal is "*A pot of tea helps to pass the evening*" spoken by a male speaker.

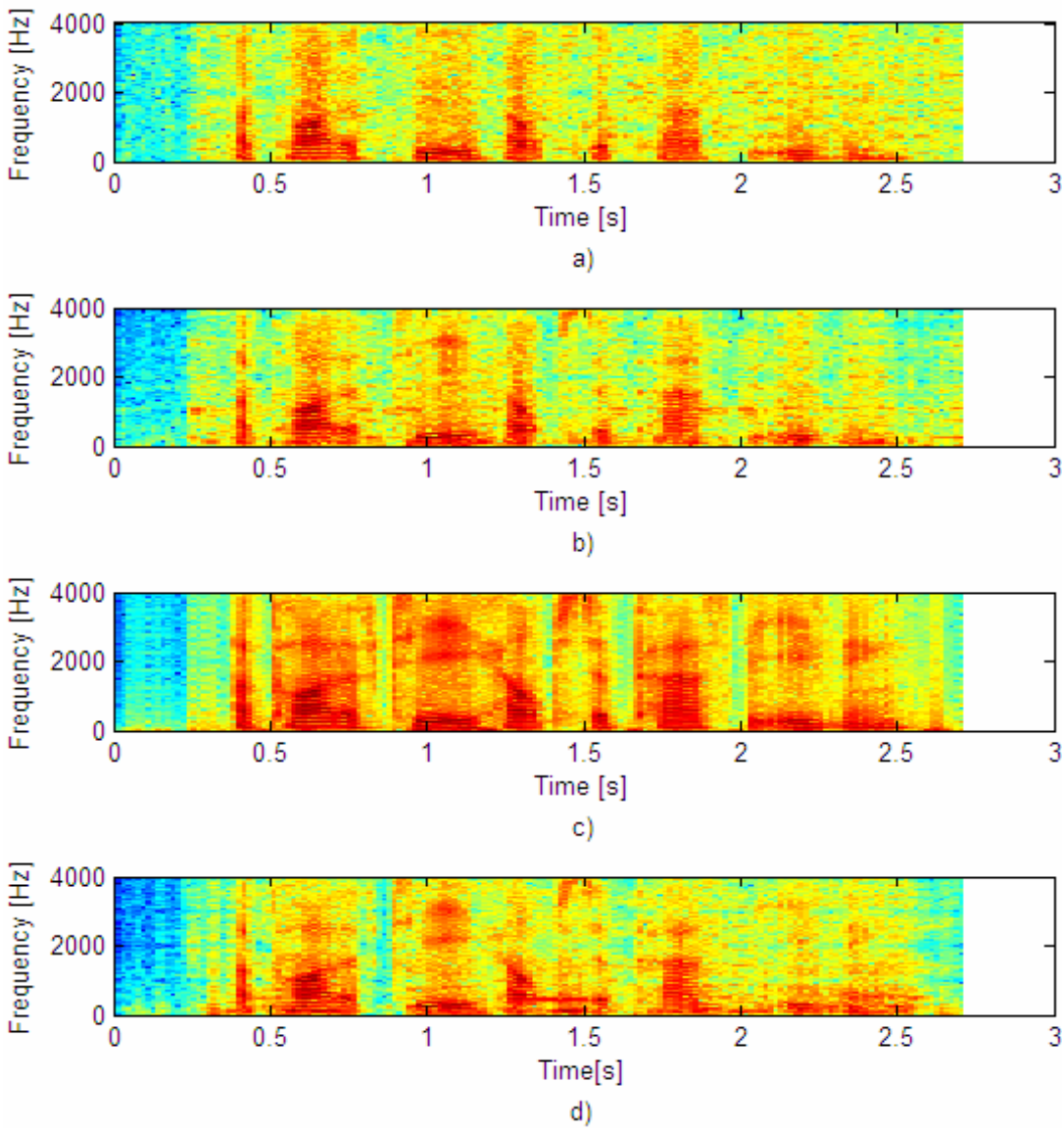

Figure 5. 8 Enhanced speech signals using Mod-WF via the proposed adaptive lifting scheme: a) WGN (SegSNR improvement=5.32 dB, IS=0.14, PESQ-MOS= 1.80), b) F16 cockpit noise (SegSNR improvement=4.78 dB, IS=0.06, PESQ-MOS=2.00), c) Car interior noise (SegSNR improvement = 11.69 dB, IS=0.003, PESQ-MOS=3.34), d) Speech babble noise (SegSNR improvement = 3.82 dB, IS=0.03, PESQ-MOS=2.10). The original speech signal is "*A pot of tea helps to pass the evening*" spoken by a male speaker.

5.5 Performance Evaluation for Image Enhancement Algorithms

The performance of proposed speech enhancement algorithms is tested by using Peak-Signal-to-Ratio (PSNR) test. The PSNR test is a general and reliable objective quality measurement test for the image enhancement applications. The spatial domain methods, Median filter and adaptive Wiener filter and wavelet thresholding

based methods, Bayes-Shrink, Normal-Shrink, Visu-Shrink and Level-Dependent thresholding methods are used based on soft thresholding criteria. The methods are all used in the lifting-wavelet domain. The original test image is degraded by adding uncorrelated white Gaussian noise at standard deviations (STDs) of (10, 15, 20, 25, 30). The noisy image is decomposed into subbands using the proposed adaptive 2-D lifting scheme. The detail noisy subbands (sub-images) LH, HL and HH are enhanced or (denoised) using the proposed image enhancement methods. The enhanced (processed) image is reconstructed via the inverse 2-D lifting scheme. The mathematical derivation of the PSNR test is given in the following section. The PSNR results obtained at various standard deviations (10, 15, 20, 25, 30) for the test images (Lena, Boat and Barbara, 512x512) are given in Figures 5.9-5.11. Furthermore, the experimental results (original, noisy and enhanced images) obtained for the test images (Lena, Boat and Barbara, 512x512) at STD of 25 are given in the Figures 5.12-5.13.

5.5.1 Peak Signal to Noise Ratio Test

The performance of the proposed image enhancement algorithm is tested by applying the peak signal to noise ratio (PSNR) test to the results of the enhancement algorithm. The PSNR test is a well known and reliable objective evaluation test which is used for the image enhancement applications Let and $w(x, y)$ be the restored and the original images where the image size is MxN. The mean square (MSE) error between two images can be given as:

$$MSE = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left[\hat{w}(x, y) - w(x, y)\right]^2 \tag{5.7}$$

Peak Signal-to-Noise Ratio (PSNR) is given in (dB) as:

$$PSNR = 10\log\left(\frac{S^2}{MSE}\right) \tag{5.8}$$

Where ‘S’ is the maximum pixel value which is equal to 255 for 8 bits/pixel image

5.6 Performance Evaluation Results for Image Enhancement Algorithms

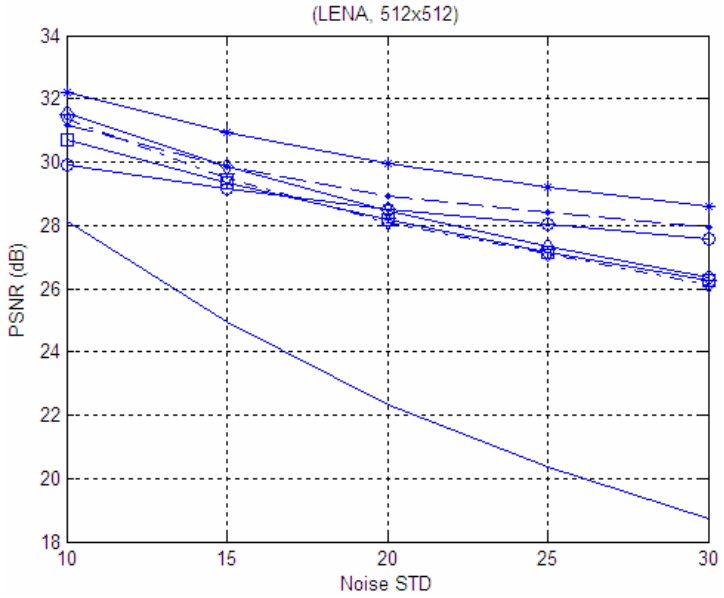

Figure 5.9 Performance curves for image enhancement algorithms for image (Lena, 512x512) for various standard deviations. Noisy image, (bold), Bayes Shrink (dashed, ●) , Normal Shrink(bold, ∇), Visu Shrink (bold, □), LDT (bold ,◊), Median Filter (bold, o), Ad-WF2 (bold, *)

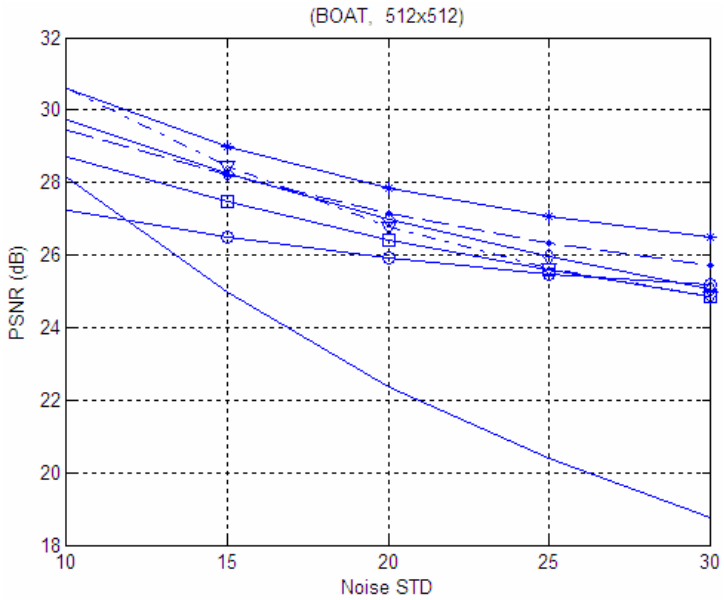

Figure 5.10 Performance curves for image enhancement algorithms for image (Boat, 512x512) for various standard deviations. Noisy image, (bold), Bayes Shrink (dashed, ●) , Normal Shrink(bold, ∇), Visu Shrink (bold, □), LDT (bold ,◊), Median Filter (bold, o), Ad-WF2 (bold, *)

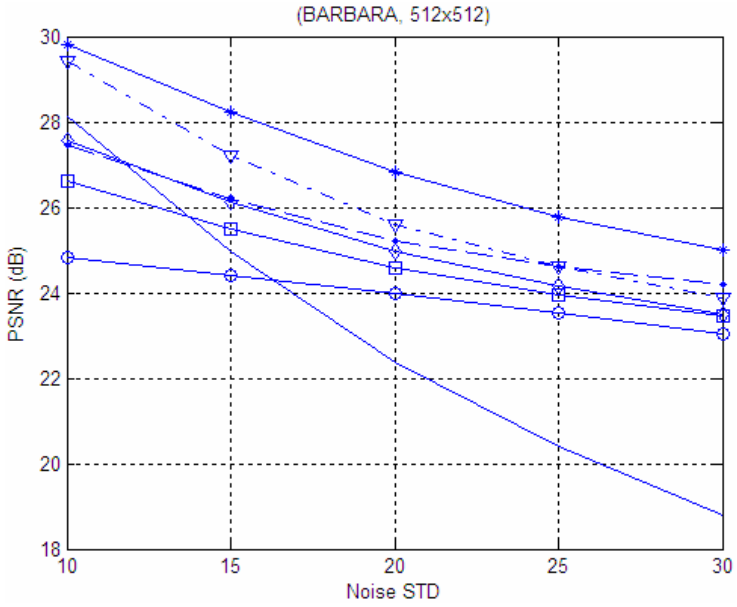

Figure 5.11 Performance curves for image enhancement algorithms for image (Barbara, 512x512) for various standard deviations. Noisy image, (bold), Bayes Shrink (dashed, ●) , Normal Shrink(bold, ∇), Visu Shrink (bold, □), LDT (bold ,◊), Median Filter (bold, o), Ad-WF2 (bold, *)

5.7 Experimental Results of Image Enhancement Methods

LENA (512x512) **BOAT (512x512)** **BARBARA (512x512)**

a)

PSNR=20.22 PSNR= 20.28 PSNR=20.30

b)

Figure 5.12. Original and noisy images, from left to right (Lena, Boat and Barbara, 512x512)
a) Original images

b) Noisy images (STD=25)

LENA (512x512)	BOAT (512x512)	BARBARA (512x512)
		
PSNR =29.14	PSNR= 27.07	PSNR=25.52

a)

		
PSNR =28.04	PSNR= 25.48	PSNR=23.50

b)

		
PSNR =28.42	PSNR=26.15	PSNR=24.56

c)

		
PSNR =27.06	PSNR= 25.59	PSNR=24.59

d)

LENA (512x512) **BOAT (512x512)** **BARBARA (512x512)**

PSNR =27.10 PSNR=25.56 PSNR=23.91

e)

PSNR = 27.35 PSNR= 25.87 PSNR=24.10

f)

Figure 5.13 Enhanced images, from left to right (Lena, Boat and Barbara, 512x512), (STD=25) using various enhancements methods:

a) Enhanced images using Ad-WF2
b) Enhanced images using Median Filter
c) Enhanced image using Bayes Shrink
d) Enhanced images using Normal Shrink
e) Enhanced images using Visu Shrink
f) Enhanced images using LDT

CHAPTER 6

RESULTS AND CONCLUSIONS

This chapter includes summary of conclusions on the performance evaluation results presented in Chapter 5 and major contributions. The possible directions for the future work are also given in the scope of this chapter.

6.1 Conclusions on the Results of the Proposed Speech Enhancement Method

The proposed speech enhancement method is based on space adaptive (1-D) lifting scheme using wavelet packet decomposition. The wavelet packet decomposition is advantageous over the wavelet decomposition since it allows better signal handling and analyzing the subbands of interest (i.e. allows critical-band decomposition). The advantages of the lifting scheme have been given in Chapter 1 and Chapter 3 in detail. The adaptive lifting schemes give rise to more efficient signal representations since it uses different lifting filters different sample points during the transform. However, the classical lifting scheme always uses the same lifting filter throughout the transform which does not provide a good signal representation since a high order filter does not well adapt to the signal structure on an edge or discontinuity point in the signal. Similarly, a low order filter does not well adapt to the signal structure on smooth parts.

It is known that the classical speech enhancement methods aim to improve the SNR of speech and they have no effect on the perceptual quality or intelligibility of the enhanced speech. In order to improve the quality or intelligibility of enhanced speech, we have taken into account the aspects of human auditory system. The subbands of the signal have been adjusted according to the Critical Bands of the human auditory system. Such a decomposition is called as CB-WPD or "perceptual filterbank" since the CB-WPD tree structure represents the human auditory system.

For subband speech enhancement, we basically focus on the MMSE-based estimators. These estimators (STSA, LSA, and STSA-Wiener) are well known

popular estimators. The estimator is based on a priori SNR estimation, noise PSD estimation and estimation of priori probability of speech absence. The priori SNR is a key parameter in these estimators and it is estimated directly from the noisy observation signal using "decision directed method". The noise PSD is estimated from the noisy speech based on pause detection using VAD. Another important parameter in the estimators is the priori probability of speech absence. The estimators are modified taking into account the priori probability of speech absence. The MMSE-based estimators provide a trade-off between noise suppression and signal distortion when the parameters of the estimators are well tuned.

Following conclusions are extracted from the evaluation results of the proposed speech enhancement algorithms and itemized as follows:

1- The Mod-WF provides the best SegSNR, IS distance and PESQ-MOS results (especially for WGN). The MM-LSA and the Mod-STSA methods provides the second and third best results. The MSS provides the worst results.

2- For WGN the SegSNR improvement IS and PESQ-MOS results for all the estimators are compatible to each other.

(This may be due to the fact that, the estimators employed for subband speech enhancement are based on Gaussian statistical model)

3- The classical WF also provides reasonable results. The only disadvantage is the clipping the sound at the beginning of the speech.

4- Space-adaptive lifting scheme provides good signal representation and enhancement results.

5- Integrating a perceptual filterbank with the proposed speech enhancement method leads to improved quality and intelligibility.

6- Proposed speech enhancement method causes insignificant speech distortion and a reasonable computational cost (14 s.)

6.2 Conclusions on the Results of the Proposed Image Enhancement Method

The proposed image enhancement method is based on 2-D adaptive lifting scheme. The noisy test image (corrupted by WGN) is decomposed into sub-images using the proposed adaptive 2-D lifting scheme with 2-levels decomposition. The noisy subbands are enhanced or (denoised) using spatial domain modified 2-D Wiener and median filters and wavelet based thresholding methods (soft thresholding) Bayes Shrink, Normal Shrink, Visu Shrink and LDT.

The following conclusions are drawn from the results of the proposed image enhancement method.

1- The adaptive 2-D Wiener filter (Ad.WF2) provides the best enhancement results among all the spatial domain and thresholding methods.

2- Bayes Shrink provides the best image enhancement results among the wavelet thresholding methods.

3- The proposed 2-D space-adaptive lifting scheme provides good edge preserving ability than the classical lifting scheme with CDF(1.3) filter.

4 The experimental and evaluation (PSNR) results show that the proposed image enhancement method provides good visual quality and satisfactory evaluation results.

6.3 Main Contributions

The following original contributions are made to the subject.

1- Development of a new adaptive prediction method (in the proposed space adaptive (1-D) lifting scheme).

2- Development of a perceptual filterbank using Critical-band decomposition.

3- Modification of MMSE-STSA and MMSE-LSA and STSA-Wiener Filters by taking into account a priori probability of speech absence, where

probability of speech absence is estimated adaptively for each spectral bin in each short time frame as $q_k(l)$.

(As we know, the Mod-STSA and Mod-WF with adaptively estimated a priori probability of speech absence $q_k(l)$, are not used in the literature before).

4- Development of a 2-D "space-adaptive" lifting scheme algorithm (forward and inverse lifting schemes), by applying proposed 1-D space-adaptive lifting scheme algorithm to the 2-D images.

6.4 Suggestions for Future Work

1- The proposed space-adaptive (1-D and 2-D) lifting scheme algorithms carry the potential of further improvement by exploiting the other wavelet filters into the lifting scheme.

2- The scale-adaptive lifting scheme may further reduce the computational cost.

3- One disadvantage of the proposed space adaptive lifting scheme algorithm is that, it requires bookkeeping for inverse transform which causes extra computational load. An adaptive lifting scheme needing no bookkeeping may be more efficient.

4- The other statistical models (Gamma, Laplacian) may cause further improved noise suppression (for speech enhancement).

5- The proposed space-adaptive 2-D lifting scheme based image enhancement method can be applied to the other fields such as enhancement of medical, satellite and geographical images.

APPENDICES

A1. SHORT-TIME FOURIER TRANSFORM (STFT)

The Fourier transforms DFT (or FFT) do not clearly show how the frequency content of a signal changes over time. That information is hidden in the phase and it is not revealed by the plot of the magnitude of the spectrum. To see how the frequency content of a signal changes over time, we can segment the signal and compute the spectrum of each segment. The result can be improved if:

a) Segments are overlapping,
b) Each segment is multiplied by a window that is tapered at its endpoints.

Several parameters must be chosen:

- Window length.
- The type of window.(Hamming, Hanning, …)
- Amount of overlap between windows.
- Amount of zero padding, if any.

Mathematical definition of Discrete Short Time Fourier Transform (STFT):

$$X(m,\omega_k) = \sum_{n=0}^{L-1} x[n]\omega[n-mL]e^{-j\omega_k n} \qquad \text{(A.1)}$$

Where,

$\omega_k = \frac{2\pi}{N}k$: Discrete frequency, N: FFT length, $x[n]$: Input signal at time n, $\omega[n]$:

Window function of length L, X(m,ω): DFT(or FFT) window data centered about time mL, L: Window length, R: Distance between two consecutive windows (window hop size), L-R: Overlap. (R-L = 0,…0 zero padding).

A2. PERFECT RECONSTRUCTION (PR) CRITERIA

Let us consider the critically sampled two channel (single level) wavelet filterbank because it is the simplest and most important case in practice and leads to wavelets.

Let $x[n]$ and $\hat{x}[n]$ are the original and reconstructed signals; $H(z)$ and $G(z)$ are low-pass and high-pass analysis filters and $\widetilde{H}(z)$ and $\widetilde{G}(z)$ are low-pass and high-pass synthesis filters.

The overall system response of the filterbank in frequency domain can be given as:

$$\begin{aligned}\hat{X}(\omega) &= \frac{1}{2}\left[H(\omega)\widetilde{H}(\omega) + G(\omega)\widetilde{G}(\omega)\right]X(\omega) \\ &\quad + \frac{1}{2}\left[H(\omega+\pi)\widetilde{H}(\omega) + G(\omega+\pi)\widetilde{G}(\omega)\right]X(\omega+\pi) \\ &= X(\omega)e^{-j\omega l}\end{aligned} \tag{A.2}$$

Where the first term is a linear shift invariant (LSI) system response related to *distortion* and the second term with $(\omega+\pi)$ reflects the system aliasing. For the perfect reconstruction (PR) filterbank the filters have to satisfy the following two PR conditions.

1- Distortion-free (DF) condition:

$$\begin{aligned}\left[H(\omega)\widetilde{H}(\omega) + G(\omega)\widetilde{G}(\omega)\right] &= 2e^{-j\omega l} \\ \left[H(z)\widetilde{H}(z) + G(z)\widetilde{G}(z)\right] &= 2z^{-l}\end{aligned} \tag{A.3}$$

2- Aliasing-free (AF) condition:

$$\begin{aligned}\left[H(\omega+\pi)\widetilde{H}(\omega) + G(\omega+\pi)\widetilde{G}(\omega)\right] &= 0 \\ \left[H(-z)\widetilde{H}(z) + G(-z)\widetilde{G}(z)\right] &= 0\end{aligned} \tag{A.4}$$

REFERENCES

[1] Martin, R. (2003). Statistical methods for the enhancement of noisy speech. *IWAENC*, Kyoto, Japan.

[2] Reddy, D.R. (1975). *Speech recognition.* New York: Academic Press.

[3] Rabiner, L.R., Schafer, R.W. (1978). *Digital signal processing,* Englewood Cliffs, New Jersy: Prentice-Hall Inc.

[4] Widrow, B. et al. (1975). Adaptive noise cancelling principles and applications, *Proceedings of IEEE*, 63(12), 1692-1716.

[5] Erçelebi, E. (2004). Speech enhancement based on the discrete Gabor transform and multi-notch adaptive digital filters. *Applied Acoustics,* 65, 739-762.

[6] Boll, S.F (1979). Suppression of acoustic noise using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2), 113-120.

[7] Berouti, M., Schwartz, R., Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise, *ICASSP,* 208-211.

[8] Boubakir, C., Berkani, Z. D., Grenez, F. (2007). A frequency-dependent speech enhancement methods, *Journal of Mobile Communication* 1(3), 97-100.

[9] Ephraim Y., Malah, D. (1984). Speech enhancement using a minimum mean square error short time spectral amplitude estimator, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP*-32(6), 1109-1121.

[10] Ephraim, Y., Malah, D. (1985). Speech enhancement using a minimum mean square error log-spectral amplitude estimator, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP*-32(2), 449-445.

[11] Malah, D., Cox, R.V., Accardi, A.J. (1999). Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments, *Proceedings of IEEE ICASSP*, 789–792.

[12] McAulay, R.J., Malpass, M.L. (1980). Speech enhancement using a soft-decision noise suppression filter, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(2), 137-145.

[13] Soon, I.Y., Koh, S.N. and Yeo, C.K. (1999). Improved noise suppression filter using self adaptive estimator of probability of speech absence, *Signal Processing,* 75(2) 151-159.

[14] Ephraim, Y., Malah, D., Juang, B.H. (1989). On the application of hidden Markov models for enhancing noisy speech, *IEEE Transactions on Acoustics, Speech, and Signal Process, ASSP*-37(12) , 1846-1856.

[15] Merhav, N. and Ephraim, Y. (1991). Hidden Markov modeling using a dominant state sequence with application to speech recognition, *Computer Speech and Language*, 5, 327-339.

[16] Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system, *IEEE Transactions on Speech and Audio Processing*, 7(2), 126-137.

[17] Tsoukalas, D., Paraskevas, M. and Mourjopoulos, J. (1993). Speech enhancement using psycho acoustic criteria, in *Proceedings of IEEE ICASSP*, 359-362, Minneapolis.

[18] Usagawa, T., Iwata M. and Ebata, M. (1998). Speech parameter extraction in noisy environments using a masking model, in *Proceedings of IEEE ICASSP*, II, 81-84, Adelaide, Australia.

[19] Cohen, A., Kovacevic, J. (1996). Wavelets: The mathematical background, *Proceedings of the IEEE*, 84(4).

[20] Vetterli, M., Herley, C. (1992) Wavelets and filter banks: theory and design, *IEEE Transactions on Signal Processing,* 40(9), 2207-2232.

[21] Polikar, R. (1999). The story of wavelets," in *Proceedings of IMACS IEEE CSCC'99*, 5481-5486.

[22] Soon, I.Y., Koh, S.N. Yeo, C.K. (1997). Wavelets for speech denoising, *IEEE TENCON*, 2, 479-482.

[23] Abbate, A. (2002). *Wavelets and subbands: Fundamentals and applications* Boston: Birkhauser.

[24] Donoho, D.L., Iain M., Johnstone B. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika,* 81(3) , 425-455.

[25] Donoho, D.L. (1995). De-noising via soft-thresholding, *IEEE Transactions on Information Theory* 41 (3), 613-627.

[26] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association* 90, 1200-1224.

[27] Fan, N., Balan R., Roska, J. (2004). Comparison of wavelet and FFT based single channel speech signal noise reduction techniques, *Wavelet Applications in Industrial Processing II. Proceedings of the SPIE,* 5607, 127-138.

[28] Ganchev, T., Siafarikas, M., Fakotakis, N. (2004, September, 8-11). Speaker verification based on wavelet packets, *International conference on Text, Speech and Dialogue (TSD*), 3206, 299-306, Brno, Tcheque Republique.

[29] Taşmaz, H. and Erçelebi, E. (2008). Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE-STSA estimation in various noise environments, *Digital Signal Processing*, 18(5), 797-812.

[30] Cohen, I. (2001, Sept.). Enhancement of speech using bark scaled wavelet package decomposition, *in EUROSPEECH–2001,* Denmark.

[31] Wang, J.-F., Yang, C.-H. and Chang, K.-H. (2004, May). Subspace tracking for speech enhancement in car noise environments, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, 2, 789–792, Montreal, Quebec, Canada.

[32] Sweldens, W. (1995). The lifting scheme: A new philosophy in biorthogonal wavelet constructions, *Wavelet Application in Signal and Image Process*ing III, 2569, 68-79.

[33] Sweldens, W. (1996). The lifting scheme: A custom design construction of biorthogonal wavelets, *Applied and Computational. Harmonic Analysis*, 3(2), 186-200.

[34] Daubechies, I., Sweldens, W. (1998). Factoring wavelet transform into lifting steps, *Journal of Fourier Analysis and Applications*, 4(3), 245-267.

[35] Piella G., Heijmans, H.J.A.M. (2002). Adaptive lifting schemes with perfect reconstruction," *IEEE Transactions on Image Processing,* 50(7), 1620-1630.

[36] Pan, Y. Wu, Q., Zhang, H., Zhang, S. (2004, Aug. 31- Sept. 4). Adaptive denoising based on lifting scheme, In Proceedings *ICSP'04*, *7th. International Conference on Signal Processing,* 1, 352-355, Beijing, China.

[37] Claypoole, R., Baraniuk, R., Novak, R. (1999). Adaptive wavelet transform via lifting, *Submitted to IEEE Transactions on Signal Processing, May 1999. (also Rice University ECE Technical Report, #9304,)*

[38] Claypoole, R.L., Davis, G.M., Sweldens, W., Baraniuk, R.G. (2003). Nonlinear wavelet transforms for image coding via lifting," *IEEE Transactions on Image Processing,* vol. 12(12), 1449-1459.

[39] Zhang, E.H., Huang, S.Y. (2004, Aug. 26-29). A new image denoising method based on the dependency wavelet coefficients, *Proceedings of International Conference on Machine Learning and Cybernetics,* 6, 3841-3844.

[40] Gonzalez, R.C., Woods, R.E., Eddins, S.L. (2004). *Digital image processing using MATLAB*, New Jersey: Pearson, Prentice-Hall.

[41] Dorf, R.C. (2006). *Circuits, signals, and speech and image processing*, Boca Raton FL : CRC/Taylor&Francis.

[42] Gonzalez, R.C., Woods, R.E. (2002). *Digital image processing*, N. J. : Prentice Hall.

[43] Erçelebi, E. and Koç, S. (2006). Lifting-based wavelet domain adaptive Wiener filter for image enhancement, *IEE Proceedings – Vision, Image, and Signal Processing*, 153(1).

[44] Pesquet, J.C. and Leporini, D. (1997, July 14-17). A new wavelet estimator for image demising, *IEE Sixth International Conference on Image Processing and its Applications*, Dublin, Ireland, 1, 249-253.

[45] Chen, Y. and Han, C. (2005, May 12). Adaptive wavelet threshold for image denoising, *IEE Electronics Letters,* 41(10), 586-587.

[46] Achim, A. and Kuruoğlu, E.E. (2005, January 17-20). Image denoising using bivariate α -stable distributions in the complex wavelet domain, *IEEE Signal Processing Letters*, 12(1), 17-20.

[47] Hussain, I. and Yin, H. (2008, March, 12-14.). A novel wavelet thresholding method for adaptive image denoising, *3rd International Symposium on Communications, Control, and Signal Processing, ISCCSP 2008*, 1252-1256, Malta.

[48] Kaur, L., Gupta, S., Chauhan, R.C. (2002). Image denoising using wavelet thresholding, *Indian Conference on Computer Vision, Graphics and Image Processing,* Ahmedabad.

[49] Suda, S., Suresh, G. R., Sukanesh, R. (2007, Dec 13-15). Wavelet based image denoising using adaptive thresholding, *International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 2007*, 3, 296-300.

[50] Chang, S.G., Yu, B. and Vattereli, M. (2000) Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, 9(9), 1532-1546.

[51] El-sayed, W (2007). Image enhancement using second generation wavelet super resolution, *International Journal of Physical Sciences,* 2(6), 149-158.

[52] Koç, S. and Erçelebi, E. (2006). Image restoration by lifting-based wavelet domain e-median filter. *ETRI Journal*, 28(1)

[53] Stepien, J., Zielinski, T., Rumian, R. (2000). Image denoising using scale-adaptive lifting schemes, in *Proceedings of the IEEE International Conference on Image Processing*, Vacouver, 288–290.

[54] Burrus, S., Gopinath, R.A. and Guo, H. (1998). *Introduction to wavelets and wavelet transforms: A primer*. New Jersey : Prentice Hall.

[55] Goswami, J.C., Andrew K.C., (1999). *Fundamentals of wavelets: Theory, algorithms, and applications*, New York : Wiley.

[56] Agostino, A., DeCusatis, C.M., Dasb. P.K. (2002). *Wavelets and subbands: Fundamentals and applications* Boston : Birkhauser.

[57] Daubechies, I. (1996). Where do wavelets come from?, *Proceedings of IEEE* 84(4), 510–513.

[58] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis Machine Intelligence,* 11(7), 674–693.

[59] Coifman, R.R. and Wickerhauser, M.V. (1992). Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory* 38(2), 713–718.

[60] Wickerhauser, M.V. (1994). *Adapted wavelet analysis from theory to software.* Wellesley, MA : A. K. Peters.

[61] Tian, J., Wells, R.O. (1998) *Vanishing moments and biorthogonal wavelet systems* In Mathematics in Signal Processing IV. Oxford University Press.

[62] Maarten, H.J. and Patrick J.O., (2005). *Second generation wavelets and applications*, London: Springer.

[63] Uytterhoeven, G., Roose, D., Bultheel, A. (1997, April, 28). Wavelet transforms using the lifting scheme, *REPORT ITA-WAVELETS-WP1.1 (revised version)*.

[64] Cohen, A., Daubechies, I. and Feauveau, J. (1992). Bi-orthogonal bases of compactly supported wavelets, *Communications on Pure and Applied Math*ematics, 45, 485-560.

[65] Cohen I. and Berdugo, B. (2003). Two-channel speech enhancement based on the transient beam-to-reference ratio, *Proceedings of ICASSP 2003*, V-233-236.

[66] Cappé, O. (1994). Elimination of musical noise phenomenon with the Ephraim and Malah noise suppressor, *IEEE Transactions on Speech and Audio Processing*, 2, 345-349.

[67] Scalart, P. and Filho, J. V. (1996, May. 7-10). Speech enhancement based on a priori signal to noise estimation, *IEEE ICASSP'96*, 2, 629-632.

[68] Donoho, D.L. (1993). *Smooth wavelet decompositions with blocky coefficient kernels,* in Recent Advances in Wavelet Analysis, 259-308, Boston: Academic Press.

[69] Sweldens, W., Schröder, P. (1996). Building your own wavelets at home, in: *Wavelets in Computer Graphics, ACM SIGGRAPH Course Notes, ACM*, 15-87.

[70] Kamath, S. and Loizou, P. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, *Proceedings of ICASSP-2002*.

[71] Martin, R., Malah, D., Cox, R. V. and Accardi, A. J. (2004). A noise reduction preprocessor for mobile voice communication, *EURASIP Journal on Applied Signal Processing*, 8, 1046-1058.

[72] Manohar, K., Rao, P. (2006). Speech enhancement in nonstationary noise environments using noise properties. *Speech Communication* 48, 96-109.

[73] Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics, *IEEE Transactions on Speech and Audio Processing,* 9(5), 504-512.

[74] Cohen, I., Berdugo, B. (2001, April, 9–11). Spectral enhancement by tracking speech presence probability in subbands. *Proceedings of IEEE Workshop on Hands Free Speech Communication*, HSC'01, Kyoto, Japan, 95–98.

[75] Smith, J. O. and, Abel, J. S. (1999). Bark and ERB Bilinear Transforms,, *IEEE Transactions on Speech and Audio Processing*, 7(6), 697-708.

[76] Chen S.H., Wang, J.F. (2004). Speech enhancement using perceptual wavelet packet decomposition and teager energy operator, *Journal of VLSI Signal Processing Systems*, 36, 125-139.

[77] Pitas, I. and Venetsanopoulos, A. N. (1986) Nonlinear mean filters in image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP–34,* 573-584.

[78] Hwang, H. and Haddad, R. A. (1995). Adaptive median filters: New algorithms and results. *IEEE Transactions on Image Processing*, 4(4).

[79] Ibrahim, H., Kong, N. S. P. and Ng, T. F. (2008). Simple adaptive median filter for the removal of impulse noise from highly corrupted images. *IEEE Transactions on Consumer Electronics*, 54(4).

[80] Hansen, J.H.L., Pellom, B.L. (1998). An effective quality evaluation protocol for speech enhancement algorithms, *ICSLP–98: International Conference on Spoken Language Processing, 7*, 2819–2822, Sydney, Australia.

[81] Manohar, K., Rao. P. (2006). Speech enhancement in nonstationary noise environments using noise properties, *Speech Communication*, 48, 96-109.

[82] International Telecommunication Union, (2001). Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T P.86

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: TAŞMAZ, Haci
Nationality: Turkish (T.C.)

Date and Place of Birth: July 15th, 1969, Adıyaman
Marital status: Single
Phone: 0 505 640 7427
Fax:
e-mail: tasmaz@gantep.edu.tr

EDUCATION

Degree	**Institution**	**Year of Graduation**
MS	University of Gaziantep Elec.-Electr. Engineering	2001
BS	METU Elec.-Electr. Engineering	1993
High school	Adıyaman High school	1986

WORK EXPERIENCE

Year	**Place**	**Enrollment**
1996-Present	University of Gaziantep Vocational high school	Instructor

FOREIGN LANGUAGE

English

PUBLICATIONS

Taşmaz, H. and Erçelebi, E. Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE-STSA estimation in various noise environments, Digital Signal Processing, 18(5), pp.797–812, 2008.

Taşmaz, H. ve Erçelebi, E. Örnek Seyreltilmemiş Dalgacık Paket Dönüşümü-Duyusal Filtre Öbeği ve ÇM_LSG Kestirimi ile Konuşma Pekiştirme, IEEE 16. Sinyal işleme, İletişim ve Uygulamaları Kurultayı, SİU 2008, s. 1–4, 20–22 Nisan 2008.

Taşmaz, H. and Erçelebi, E. Image enhancement via space-adaptive lifting scheme using spatial domain adaptive Wiener filter, 6th International Symposium on Image and Signal Processing and Analysis, ISPA 2009. (Accepted paper)