WAVELET-BASED

DENOISING OF SPEECH

ARKADY BRON

WAVELET-BASED

DENOISING OF SPEECH

RESEARCH THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

IN ELECTRICAL ENGINEERING

ARKADY BRON

SUBMITTED TO THE SENATE OF

THE TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY

SIVAN , 5760 HAIFA JUNE 2000

Acknowledgements

The research thesis was carried out under the supervision of Professor Shalom Raz and Professor David Malah in the department of Electrical Engineering.

The generous financial help of the Technion is gratefully acknowledged.

I wish to appreciate to my advisors, Professor Shalom Raz and Professor David Malah, for their dedicated supervision, enthusiastic discussions and continued support throughout all stages of this research.

Many thanks to all my friends for their friendship and support along the way.

With love I dedicate this work to my mother, Klara, and to the memory of my father, Mark Bron.

Contents

List of Figures

List of Tables

Abstract

The problem of enhancing speech degraded by uncorrelated additive noise, when the noisy speech alone is available, has received much attention, since it is useful to enhance speech signals prior to coding and identification processes. Wavelet bases are widely used in various applications, such as estimating signals embedded in noise and coding. Wavelet-based methods show good performance for a wide diversity of signals. However, it has been observed that denoising with the conventional wavelet transform and Wavelet Packet Decomposition (WPD) may exhibit visual artifacts, such as pseudo-Gibbs phenomena in the neighborhood of discontinuities. These artifacts were attributed to the lack of shift-invariance.

The main purpose of the thesis was to modify and improve existing denoising algorithms and to study the consequences of shift-invariance on speech enhancement and the resulting artifacts. First, we implemented the state of the art speech denoising algorithms and wavelet-based denoising algorithms. These algorithms served as benchmarks. We then developed the speech denoising algorithms, based on WPD and Local Trigonometric Decomposition (LTD), which utilize the decision directed approach to a priori SNR estimation.

We have shown that artifacts, introduced by wavelet-based denoising algorithms [14, 16, 10, 33, 5], applied to speech enhancement, can be particularly suppressed by increasing

temporal support of the basis functions. Moreover, improvement in frequency localization of the basis functions improves the speech denoising performance. It also has been shown that shift-invariance achieved by Shift-Invariant Wavelet Packet Decomposition (SIWPD) does not contribute to artifacts suppression and does not guarantee an improved denoising performance.

Denoising based on the presumption of prior knowledge of the squared spectral amplitude of the noise is referred to as *ideal* denoising. Quite expectedly, simulations confirm that such ideal denoising attains higher SNR than the practical one. We have proven that ideal speech denoising, based on some real-valued transform, achieve an exact phase reconstruction of the clean speech signal (expressed via the sign of the real-valued transform coefficients). This is an intrinsic advantage of real-valued transforms compared to DFT-based denoising. The results show that the exact phase reconstruction associated with real-valued transforms leads to global SNR improvement by $0.69 \div 1.12$ [dB] while comparing WPD-based vs. DFT-based ideal denoising.

We have compared the proposed speech denoising algorithms to the state of the art speech denoising algorithms [18, 19]. Simulation results indicate that, for each of the tested speech signals, the DFT-based Wiener estimator attains the highest global SNR, segmental SNR and LSD. The quality of the enhanced speech is similar for all the algorithms. The notable difference is the level and type of a residual background noise. All of the algorithms, Ephraim-Malah being the exception, introduce a colored background noise, that was found to be disturbing the listener. The DFT-based Wiener estimator is characterized by the lowest level of the residual noise, and is superior to the proposed algorithms. The Ephraim-Malah algorithm is characterized by a higher level of background noise then DFT and WPD-based Wiener estimator, but, advantageously, the background noise is almost white.

Despite the advantages of WPD and LTD-based algorithms under ideal denoising conditions, in practice (i.e., with an estimated noise variance) the DFT-based denoising algorithms are found to be better. The reasons are:

1) Given the noisy observations, we can't know the exact values of the noise squared spectral components. Hence, using only the estimated averages of the noise squared spectral components we can't exactly reconstruct the clean speech phase.

2) It is shown in Appendix I, that if the additive noise is white and Gaussian, the variance of its squared spectral components, obtained by real-valued transform, is twice (except for the DC coefficient) the variance of the noise squared spectral components, obtained by the DFT. This leads to higher deviations of noise squared spectral amplitude from its estimated value, and subsequently to higher frame to frame gains fluctuations (segment to segment gains fluctuations for LTD-based denoising) thus reducing the resulting global and segmental SNR. The frame to frame gains fluctuations cause the residual background noise to be colored.

Despite the fact that the speech denoising algorithms proposed herein do not possess clear advantage over the DFT-based algorithms, they may have merit in a wider sense. For example, WPD-based denoising can be easily incorporated into a WPD-based speech coding system. Also, LTD can be used as a time-segmentation tool. Thus, the LTD-based denoising algorithm can be conveniently implemented in speech analysis systems, which require adaptive time-segmentation.

List of Symbols and Abbreviations

Symbols

$\mathcal{B}$	Library of orthonormal bases
$\mathcal{L}(\mathbf{y})$	Description length of $\mathbf{y}$
$\mathcal{L}(B\mathbf{y})$	Description length of $\mathbf{y}$ expanded in the basis B
$\mathcal{M}$	Additive information cost function
$\mathcal{M}(B\mathbf{f})$	Information cost of $\mathbf{f}$ expanded in the basis B
$\mathbb{N}$	Set of naturals $\{1, 2, 3, \ldots\}$
$\mathbb{R}$	Set of reals
$\mathbb{Z}$	Set of integers $\{0, \pm 1, \pm 2, \ldots\}$
$\mathbb{Z}_+$	Set of non-negative integers $\{0, +1, +2, \ldots\}$
$\mathbb{Z}_-$	Set of non-positive integers $\{0, -1, -2, \ldots\}$
Ω	Collection of models
Υ	Description model
A	The optimal basis for signal estimation
$A_{\ell,n}, A_{\ell,n}^{(m)}$	The best set of wavelet packets for the subspaces $U_{\ell,n}$ and $U_{\ell,n}^{(m)}$ respectively

$B_{\ell,n}, B_{\ell,n}^{(m)}$	Set of wavelet packets associated with the tree-nodes (ℓ, n) and (ℓ, n, m) respectively
$C^s, C^s(\mathbb{R})$	Class of s-times continuously differentiable functions
E	Set of terminal nodes of an expansion tree (tree-set)
$F(\omega)$	Fourier transform of $f(t)$
G	Quadrature Mirror High-Pass Filter
H	Quadrature Mirror Low-Pass Filter
$I_{\ell,n}$	Dyadic intervals
J	Maximal depth of a decomposition tree
L	Number of decomposition levels $(1 \leq L \leq J)$
L^2	Square-integrable functions
M_α	Mirror Operator (around point α)
N	Length of signal at its highest resolution level
$P(A)$	Probability of event A
$U_{\ell,n}, U_{\ell,n}^{(m)}$	Closure of the linear span of $B_{\ell,n}, B_{\ell,n}^{(m)}$
$\mathbf{f}$	Unknown signal $f(t)$
$\widehat{\mathbf{f}}$	Estimate of $\mathbf{f}$
$\theta_i, \theta_{\ell,n,m}$	Expansion coefficients of the unknown signal
$\Theta, \Theta_{\ell,n}$	Expansion coefficients $\{\theta_i\}$ $(\{\theta_{\ell,n,k}\})$ of the unknown signal $f(t)$
$\hat{\theta}_i, \hat{\theta}_{\ell,n,m}$	Estimate of expansion coefficients of the unknown signal
$\widehat{\Theta}, \widehat{\Theta}_{\ell,n}$	Estimate of expansion coefficients Θ $(\Theta_{\ell,n})$
$g^*(t)$	Complex conjugate of $g(t)$

$\{h_k\}, \{g_k\}$	Wavelet decomposition filter banks
k	Time-domain position index (for WPD) or
	frequency-domain position index (for LTD and DFT)
ℓ	Resolution level index
ℓ^p	ℓ^p norm
(ℓ, n)	Index of a tree-node in WPD tree
(ℓ, n, m)	Index of a tree-node in SIWPD tree
m	Shift index
m_ℓ	Shift index at the resolution level ℓ
m_c	Shift index of children-nodes
m_p	Shift index of a parent node
p	Order of AR model
$p(x)$	Probability density function
r	Roll-off of modified Meyer QMF
$\mathbf{y}$	Noisy data $y(t)$
$\mathbf{w}$	Expansion coefficients w_i of $y(t)$
$\mathbf{e}$	White Gaussian noise $e(t)$
$\mathbf{z}$	Expansion coefficients z_i of $e(t)$

ξ_k	A priori SNR of k-th spectral component
γ_k	A posteriori SNR of k-th spectral component
$\delta_{k,\ell}$	Kronecker delta function
η	Interval on which a window function rises from being

	identically zero to being identically one
η_{max}	Maximal allowed η
η_h	Hard-thresholding operator
η_s	Soft-thresholding operator
λ, λ_d	Thresholds for $RiskShrink$ and $VisuShrink$ estimators
$\phi(t)$	Basis functions
$\varphi(t), \psi_0(t)$	Scaling function
σ^2	Variance of white noise
$\psi(t), \psi_1(t)$	Mother wavelets
$\psi_{\ell,n,k}(t), \psi_{\ell,n,k}^{(m)}(t)$	Wavelet basis functions

$x \bmod y$	Modulus (signed remainder after division)
$\lfloor x \rfloor$	Integer part of x
$\mathrm{Re}\{f\}$	Real part of f
$\#S$, $\|S\|$	The number of elements in the set S
$\|c\|$	Magnitude of a complex number c
$\\|g\\|$	Norm of g
$\mathrm{clos}_{L^2(\mathbb{R})}\{S\}$	Closure of the linear span of S
$\langle f, g \rangle$	Inner product of f and g
$\mathbf{1}_I$	Indicator function for the interval I
$\sim$	Equivalence relation

Abbreviations

AMDL	Approximate Minimum Description Length
AR	Autoregressive (model)
BIOR	Biorthogonal wavelets
CP	Cosine Packet
CPD	Cosine Packet Decomposition
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DMP	Daubechies Minimum Phase wavelets
DNS	Daubechies Nearly Symmetric wavelets
DWT	Discrete Wavelet Transform
FIR	Finite Impulse Response
IDFT	Inverse Discrete Fourier Transform
IFT	Inverse Fourier transform
LSD	Log-Spectral Distance
MDL	Minimum Description Length
ML	Maximum Likelihood
MMSE	Minimum Mean-Square Error
MSE	Minimal Square Error
PSD	Power spectral density
QMF	Quadrature Mirror Filter
SA	Spectral Amplitude
SEGSNR	Segmental SNR

SIAP-LTD	Shift-Invariant Adaptive Polarity Local Trigonometric Decomposition
SIWP	Shift-Invariant Wavelet Packet
SIWPD	Shift-Invariant Wavelet Packet Decomposition
SIWPR	Shift-Invariant Wavelet Packet Reconstruction
SIWT	Shift-Invariant Wavelet Transform
SNR	Signal to noise ratio
STFT	Short-Time Fourier Transform
STSA	Short-Time Spectral Amplitude
SWP	Shifted Wavelet Packet
WGN	White Gaussian noise
WP	Wavelet Packet
WPD	Wavelet Packet Decomposition

Chapter 1 : Introduction

1.1 Motivation

The problem of enhancing speech degraded by uncorrelated additive noise, when the noisy speech alone is available, has received much attention. This is due to a variety of potential applications speech enhancement possesses. Furthermore, technologies enabling the implementation of such intricate algorithms are now available. The main purpose of denoising techniques is to improve the quality and comprehension of speech. It's also useful to enhance the speech prior to the implementation of techniques such as coding and recognition. Unfortunately, while existing speech denoising algorithms appear to improve the quality of speech, they typically do not improve its comprehension.

Wavelet bases are widely used for estimating signals embedded in noise. While traditional methods often remove noise by low-pass filtering, thus blurring the sharp features in the signal, wavelet-based methods show good performance for a wide diversity of signals. The *wavelet shrinkage* method, developed by Donoho and Johnstone [17], uses a fixed transform of the noisy data into the wavelet-domain, applies soft or hard thresholding to the resulting coefficients, and subsequently transforms the modified wavelet-domain coefficients back into the original space. It was recognized that the success of such a denoising scheme is determined by the extent to which the transform compresses the unknown signal into few significant coefficients [15]. Given a library of bases and a noisy measurement, researchers proposed several different approaches to select a "best" basis and a threshold value, leading to the best signal estimate [13].

Saito [33] proposed to use an information-theoretic criterion, called the *Minimum Description Length* (MDL) principle [30], for noise removal. He claimed that the MDL criterion gives the best compromise between the estimation fidelity (noise suppression) and the efficiency of representation (signal compression).

It has been observed [10, 1, 33] that denoising with the conventional wavelet transform and wavelet packet decomposition (WPD) may exhibit visual artifacts, such as pseudo-Gibbs phenomena in the neighborhood of discontinuities. These artifacts were related to the lack of *shift-invariance*, and proposed to reduce them by averaging over different translations: Applying a range of shifts to the noisy data, denoising the shifted versions with the wavelet transform, then unshifting and averaging the denoised data. This procedure, termed *Cycle-Spinning* [10], generally yields better visual performance on smooth parts of the signal.

Cohen, Raz and Malah [5] presented an extension of WPD into a Shift-Invariant WPD (SIWPD). Moreover, they reformulated the MDL principle as an additive information cost function [8] and presented an adaptive translation-invariant denoising algorithm.

The main purpose of this work is to modify and improve existing denoising algorithms and to study the consequences of shift-invariance on speech enhancement and the resulting artifacts.

1.2 Overview of the Thesis

The organization of this thesis is as follows. In the next chapter we review the state of the art speech denoising algorithms and the so-called "*decision directed*" approach to *a priori* SNR estimation, that was introduced by Ephraim and Malah in [18]. In Chapter 3 we review the basics of joint time frequency representations: Wavelet packet analysis and best-basis expansion, the extension of wavelet packet bases for obtaining shift-invariance, and local trigonometric bases. In Chapter 4 we review different wavelet-based denoising algorithms, including the so-called "*translation-invariant*" denoising algorithm of Coifman and Donoho, and the Cohen-Raz-Malah shift-invariant denoising algorithm, based on shift-invariant WPD.

The main contribution of this thesis begins in Chapter 5, where we present several speech denoising algorithms, based on WPD, Cosine Packet Decomposition and WPD applied to DCT-I coefficients. We utilize the decision directed a priori SNR estimation for each of the mentioned joint time-frequency representations. Importance of shift-invariance, time support and frequency localization are discussed. In Chapter 6 we introduce a comparative performance analysis of different speech denoising algorithms, and present some interesting conclusions corresponding a comparison of DFT-based and real-valued transform-based denoising. Required proofs are given in the Appendices.

Finally, in Chapter 7 we conclude with a summary and discussion on future research directions.

Chapter 2 : State of the Art of Speech Denoising Algorithms

2.1 Speech Characteristics and Modeling

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

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

A short time segment of a speech signal can be regarded as a portion of a stationary stochastic process (or an impulse response of a digital filter), therefore generally a speech signal is said to be *quasi-stationary*, i.e., it can be divided into (almost) stationary segments. Typically segments are up to 30 ms long.

It appears that an *autoregressive* (AR) *model* (or so-called *all-pole model*) is particularly suitable for modelling a speech signal. The motivation stems from a simplified picture of the vocal tract as a lossless tube built of adjoining cylinders of different diameters [35]. An all-pole digital filter, excited by a pulse train, is a basic model for speech production. All-pole modeling of a speech signal refers to extracting the filter coefficients and source power from the given speech signal. The order p of the model can be estimated too (typically $p = 10 \div 16$) [35].

2.2 Overview of Spectral Domain Denoising Algorithms

2.2.1 Spectral Subtraction

Suppose we have noisy data $\mathbf{y} = \{y_i\}_{i=0}^{N-1}$, where

$$y_i = f_i + e_i, \qquad i = 0, ..., N-1, \tag{2.1}$$

$\mathbf{f} = \{f_i\}_{i=0}^{N-1}$ is an unknown real-valued signal which we would like to recover, and $\mathbf{e} = \{e_i\}_{i=0}^{N-1}$ is additive noise. Figure 2.1 shows the basic procedure for spectral domain noise removal. The time-domain signal is first broken up into a series of overlapping (usually 25% or 50% overlapping) frames. Each frame is multiplied by a smooth window function (such as Hanning or Hamming window) in order to reduce spectral artifacts caused by the discontinuities at the edges of the frame. Then, each windowed segment is transformed into the spectral domain and processed individually: the spectral coefficients are multiplied by an appropriate gain. The modified spectral components are then transformed back into the time domain, and the frames are assembled to get the enhanced signal. Transformation into spectral domain can be performed using any orthogonal

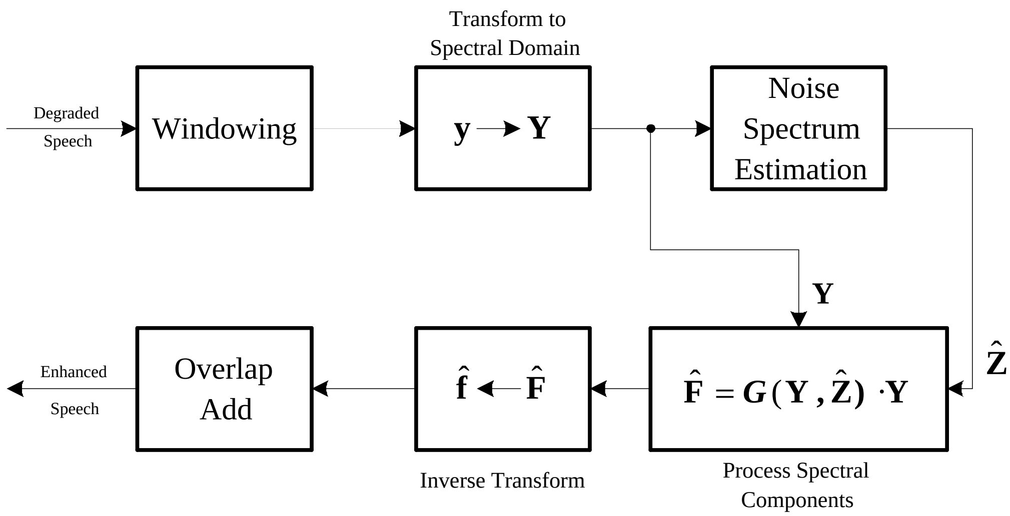


Figure 2.1: Basic Spectral Domain Denoising Procedure.

or biorthogonal transform such as Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), Wavelet Transform (WT), Wavelet Packet Decomposition(WPD) and so on.

All frequency domain algorithms need to estimate the noise spectrum prior to denoising. If one synthesizes a noisy speech signal in order to verify performance of a given denoising algorithm, it's useful to estimate the spectrum from the known noise signal. However, generally the noise spectrum can be estimated from speech-free intervals which are most adjacent in time to the analysis frame. If the noise is known to be stationary, it suffices to estimate its variance and its spectral components once, from an initial speech-free interval averaging the spectrum over speech-free segments. In [23], the noise spectrum estimation is based on *spectral minimum tracking*. Additional information about the estimation of noise spectrum can be found in [2], [24].

For nonstationary noise, some kind of noise spectrum tracking, which can potentially improve performance of a denoising algorithm, is needed. Temporal minima tracking was proposed by Doblinger in [12]. A more advanced algorithm, based on *Voice Activity Detector* (VAD), was presented by Malah, Cox and Accardi in [21].

In Fig. 2.1 $G(\mathbf{Y}, \widehat{\mathbf{Z}})$ is a gain function. The *generalized spectral subtraction* gain function [41] is given by

$$G_k = G(Y_k, \widehat{Z}_k) = \begin{cases} \left(1 - \alpha \left[\frac{|\widehat{Z}_k|}{|Y_k|}\right]^{\gamma_1}\right)^{\gamma_2}, & \left[\frac{|\widehat{Z}_k|}{|Y_k|}\right]^{\gamma_1} < \frac{1}{\alpha+\beta}, \\ \left(\beta \left[\frac{|\widehat{Z}_k|}{|Y_k|}\right]^{\gamma_1}\right)^{\gamma_2}, & otherwise, \end{cases} \tag{2.2}$$

where Y_k is the k-th noisy speech spectral component, $\widehat{Z}_k$ is the estimated k-th noise spectral component, G_k is the k-th spectral component's gain, α is the *oversubtraction factor* ($\alpha > 1$, leads to the reduction of residual noise but also to increased speech distortion), β is *spectral flooring factor* ($0 \leq \beta \ll 1$, allows to leave certain level of background noise), γ_1 and γ_2 are exponent parameters (determine the sharpness of the transition from $G_k = 1$ (the spectral component is not modified) to the $G_k = 0$ (the spectral component is suppressed)).

A number of classical gain functions can be derived from (2.2) by appropriate choice of α, β, γ_1 and γ_2 parameters:

1) *Amplitude spectral subtraction* or *magnitude subtraction* [41]:

$$G_k = G(Y_k, \widehat{Z}_k) = \begin{cases} \left(1 - \frac{|\widehat{Z}_k|}{|Y_k|}\right), & \frac{|\widehat{Z}_k|}{|Y_k|} < 1, \\ 0, & otherwise. \end{cases} \tag{2.3}$$

It corresponds to $\alpha = 1$, $\beta = 0$, $\gamma_1 = \gamma_2 = 1$.

2) *Power spectral subtraction* [41]:

$$G_k = G(Y_k, \widehat{Z}_k) = \begin{cases} \left(1 - \frac{|\widehat{Z}_k|^2}{|Y_k|^2}\right)^{1/2}, & \frac{|\widehat{Z}_k|^2}{|Y_k|^2} < 1, \\ 0, & otherwise. \end{cases} \tag{2.4}$$

It corresponds to $\alpha = 1$, $\beta = 0$, $\gamma_1 = 2$, $\gamma_2 = 1/2$.

3) *Wiener estimator* [41]:

$$G_k = G(Y_k, \widehat{Z}_k) = \begin{cases} \left(1 - \frac{|\widehat{Z}_k|^2}{|Y_k|^2}\right), & \frac{|\widehat{Z}_k|^2}{|Y_k|^2} < 1, \\ 0, & otherwise, \end{cases} \tag{2.5}$$

It corresponds to $\alpha = 1$, $\beta = 0$, $\gamma_1 = 2$, $\gamma_2 = 1$.

The above spectral subtraction gain functions can also be thought of as thresholding algorithms: If the spectral amplitude at a given frequency bin equals or is below the expected noise content for that bin, then it is assumed that the original signal had no significant contribution in that bin and it is set to zero. However, if the spectral content is above the expected noise content for a given bin, then in each of the above gain functions, the bin content is scaled according to some function of the estimated noise power and measured signal spectrum.

2.2.2 Ephraim-Malah Denoising Algorithm

This algorithm capitalizes on the major importance of the short-time spectral amplitude (STSA) of the speech signal to its perception [18]. It is well known that a distortion measure which is based on the mean squared error of the log-spectra is more suitable for speech processing. Thus the proposed algorithm for enhancing is the STSA estimator which minimizes the mean squared error of the log-spectra [19].

The estimation problem of the STSA is formulated as that of estimating the amplitude of each Fourier expansion coefficient of the speech signal $\mathbf{f} = \{f(t), 0 \le t \le T\}$, given the noisy process $\mathbf{y} = \{y(t), 0 \le t \le T\}$.

Let the $F_k = A_k e^{j\alpha_k}$, Z_k and $Y_k = R_k e^{j\vartheta_k}$, denote the k-th Fourier expansion coefficients of the speech signal, the noise process and the noisy observations in the analysis interval $[0, T]$, respectively. Following the above formulation, we are looking for an estimator $\widehat{A}_k$, which minimizes the distortion measure:

$$E\{(\log A_k - \log \widehat{A}_k)^2\}. \tag{2.6}$$

This estimator [19]

$$\widehat{A}_k = \frac{\xi_k}{1+\xi_k} \exp\left\{\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\right\} R_k \tag{2.7}$$

is briefly derived in Appendix III.1. Here, ξ_k is the so-called *a priori* signal to noise ratio (SNR):

$$\xi_k \equiv \frac{\lambda_f(k)}{\lambda_z(k)}, \tag{2.8}$$

γ_k is the so-called *a posteriori* SNR:

$$\gamma_k \equiv \frac{R_k^2}{\lambda_z(k)} \tag{2.9}$$

and v_k is defined by

$$v_k \equiv \frac{\xi_k}{1+\xi_k} \gamma_k \ , \tag{2.10}$$

where $\lambda_z(k) \equiv E\{|Z_k|^2\}$ and $\lambda_f(k) \equiv E\{|F_k|^2\}$ are the variances of the noise and the signal k-th spectral components.

It is seen from (2.7) that $\widehat{A}_k$ is obtained from R_k by a multiplicative nonlinear gain function which depends only on the *a priori* and the *a posteriori* SNRs. The gain function is defined by

$$G_k(\xi_k, \gamma_k) \equiv \frac{\widehat{A}_k}{R_k} = \frac{\xi_k}{1+\xi_k} \exp\left\{\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\right\}. \tag{2.11}$$

The log-spectral amplitude estimator is superior to the Minimum Mean-Square Error (MMSE) STSA estimator, derived in [18], since it results in a much lower residual noise level without further affecting the speech itself.

In [18] it was shown that the MMSE complex exponential estimator is the complex exponential of the noisy phase ($e^{j\widehat{\alpha}_k} = e^{j\vartheta_k}$). Moreover, it was shown that the optimal phase estimator of the speech's phase is the noisy phase itself ($\alpha_k = \vartheta_k$). Thus, the estimate $\widehat{\mathbf{f}}$ of the speech $\mathbf{f}$ is given by

$$\widehat{\mathbf{f}} = IDFT\{\widehat{A}_k e^{j\vartheta_k}\} = IDFT\{G_k(\xi_k, \gamma_k) R_k e^{j\vartheta_k}\} \tag{2.12}$$

It's clear that this estimator assumes knowledge of *a priori* SNR ξ_k and *a posteriori* SNR γ_k. Thus they have to be estimated from the observations $\mathbf{y}$. In [18] the authors proposed two methods for estimating ξ_k defined above. These approaches assume knowledge of the noise spectral component variance λ_z. In practice this variance should be estimated as well.

Maximum Likelihood Estimation Approach

The ML estimate $\widehat{\xi}_k$ of ξ_k in the n-th analysis frame is obtained in [18, 19] by

$$\widehat{\xi}_k = \begin{cases} \overline{\gamma}_k(n) - 1 \,, & \overline{\gamma}_k(n) - 1 \geq 0, \\ 0 \,, & \text{otherwise,} \end{cases} \tag{2.13}$$

where

$$\overline{\gamma}_k(n) = \alpha \overline{\gamma}_k(n-1) + (1-\alpha)\frac{\gamma_k(n)}{\beta} \,, \qquad 0 \leq \alpha < 1 \,, \quad \beta \geq 1. \tag{2.14}$$

The estimation is based on the Gaussian statistical model and the statistical independence assumed for the spectral components of a noisy speech. Here β is correction factor, and α is the smoothing parameter. They can be determined by informal listening. According to [18], the best quality of enhanced speech was achieved for $\alpha = 0.725$, $\beta = 2$.

"Decision Directed" Estimation Approach

This estimator was found to be very useful when it is combined with any amplitude or log-amplitude estimators [18, 19]. The proposed estimator is given by

$$\widehat{\xi}_k(n) = \alpha \frac{\widehat{A}_k^2(n-1)}{\lambda_z(k, n-1)} + (1-\alpha)P[\gamma_k(n) - 1] \ , \qquad 0 \leq \alpha < 1 \tag{2.15}$$

where $P[\cdot]$ is a hard thresholding operator which is defined by

$$P[x] = \begin{cases} x \ , & \text{if } x \geq 0, \\ 0 \ , & \text{otherwise.} \end{cases} \tag{2.16}$$

Using the definition of $\widehat{A}_k(n)$ we get:

$$\widehat{\xi}_k(n) = \alpha G_k^2(\widehat{\xi}_k(n-1), \gamma_k(n-1))\gamma_k(n-1) + (1-\alpha)P[\gamma_k(n) - 1]. \tag{2.17}$$

The initial condition $\widehat{\xi}_k(0) = \alpha + (1-\alpha)P[\gamma_k(0) - 1]$ was found to be appropriate, since it minimizes the initial transition effects in the enhanced speech. The theoretical investigation of (2.17) is very complicated due to its highly nonlinear nature. Therefore, the "best" value of α is to be determined by simulation. In [18], the best quality of enhanced speech was achieved for $\alpha = 0.98$.

Chapter 3 : Joint Time-Frequency Representations

3.1 Wavelet Analysis

3.1.1 Introduction

Wavelet and local trigonometric bases are two classes of bases for representing signals behavior. Under certain circumstances they may provide a better insight and a clearer interpretation of the signal (see e.g. [5, 11, 20, 22, 43]).

This chapter presents the underlying mathematics of the Wavelet Packet and Local Trigonometric libraries of bases, and how these are exploited to yield "best basis" representations. Instead of restricting ourselves to a specified basis, we consider a *library* of orthonormal bases, that may include either wavelet or local trigonometric bases.

3.1.2 Discrete Wavelet Transforms and Wavelet Packet Decompositions

Wavelet Packet Decomposition (WPD) constitute a direct extension of multiresolution analysis [11, 42]. The approach is based on the generation of a library $\mathcal{B}$ of *Wavelet Packets* (WP) which is defined by

$$\mathcal{B} = \{\psi_{\ell,n,k}(t) = 2^{\ell/2}\psi_n(2^\ell t - k) : \ell \in \mathbb{Z}_-, n \in \mathbb{Z}_+, k \in \mathbb{Z}\}, \tag{3.1}$$

where ℓ is the scaling parameter, n is the oscillation (modulation) parameter and k is the time-domain position parameter [26]. The functions $\{2^{\ell/2}\psi_n(2^\ell t - k)\}$ are the scaled and translated versions of orthonormal wavelet bases functions $\psi_n(t)$ defined by

$$\psi_{2n}(t) \equiv \sqrt{2}\sum_k h_k \psi_n(2t - k) \equiv H\psi_n(t), \tag{3.2}$$

$$\psi_{2n+1}(t) \equiv \sqrt{2}\sum_k g_k \psi_n(2t - k) \equiv G\psi_n(t), \tag{3.3}$$

where H and G are operators of digital low-pass and high-pass filtering, characterized respectively by the impulse-response sequences $\{h_k\}$ and $\{g_k\}$, followed by decimation (2:1). The filters $\{h_k\}$ and $\{g_k\}$ are the so-called quadrature mirror filters (QMFs) that have to satisfy the orthogonality, perfect reconstruction and the admissibility condition:

$$\sum_k h^*_{k-2n} g_{k-2l} = 0, \tag{3.4}$$

$$\sum_k h_k = \sqrt{2}, \tag{3.5}$$

$$\sum_l h_{k-2l} h^*_{m-2l} + g_{k-2l} h^*_{m-2l} = \delta_{m,k}, \tag{3.6}$$

$$\sum_k g_k = 0. \tag{3.7}$$

Moreover, the filter $\{g_k\}$ is defined by

$$g_k = (-1)^k h_{1-k}. \tag{3.8}$$

The basis functions $\{\psi_{\ell,n,k}\}$ represent different time and frequency resolution levels, thus this type of analysis (decomposition) is commonly referred to as *"multiresolution analysis"*.

Construction of the library $\mathcal{B}$ starts by selecting a so-called *characteristic* (*scaling*) function $\psi_0(t) \equiv \varphi(t)$, that has to satisfy the *two scale relation*:

$$\varphi(t) = \sum_k h_k \varphi_{1,k}(t) = \sqrt{2} \sum_k h_k \varphi(2t - k), \tag{3.9}$$

Here

$$\varphi_{\ell,k}(t) = 2^{\ell/2} \varphi(2^\ell t - k) \tag{3.10}$$

is the scaled and translated version of the characteristic function. The low-pass filter coefficients $\{h_k\}$ can be easily calculated from (3.9) via

$$h_k = \langle \varphi_{1,k}, \varphi \rangle = \sqrt{2} \langle \varphi(2t - k), \varphi(t) \rangle, \tag{3.11}$$

where $\langle \cdot \rangle$, the inner product, is defined by

$$\langle f, g \rangle = \int f(t) g^*(t) dt, \tag{3.12}$$

and

$$\psi_1(t) = \sqrt{2} \sum_k g_k \psi_0(2t - k) = \sqrt{2} \sum_k g_k \varphi(2t - k) = \psi(t) \tag{3.13}$$

is the *mother wavelet* $\psi(t)$. A mother wavelet function, defined by Eq. (3.13), must satisfy the *admissibility condition*: $\Psi(w)|_{w=0} = 0$. The merits of the library $\mathcal{B}$ depend on the selected mother wavelet and its properties, such as the number of vanishing moments (smoothness), compactness of support and symmetry.

The construction of an orthonormal basis for $L^2(\mathbb{R})$ is done as follows. Let E_m be the tree-set of indices $\{\ell, n\}$ that corresponds to the terminal nodes of some binary tree. If a set of disjoint intervals

$$\{I_{\ell,n}\} = \left\{ \left[2^\ell n, 2^\ell (n+1)\right) : (\ell, n) \in E_m \right\} \tag{3.14}$$

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

$$B_m = \{\psi_{\ell,n,k} = 2^{\ell/2}\psi_n(2^\ell t - k) : (\ell, n) \in E_m, k \in \mathbb{Z}\} \tag{3.15}$$

(the wavelet packet B_m) constitutes an orthonormal basis for $L^2(\mathbb{R})$. Therefore, a wavelet packet (basis) B_m in the library $\mathcal{B}$ is defined by specifying the range of ℓ, n and k parameters [26].

A particular case of *Wavelet Packet Decomposition* (WPD) (or so-called *Subband Decomposition*) of a signal $\mathbf{x} = \{x(i)\}_{i=0}^{7}$ is given in the Fig. 3.1. Here the tree-set E_m is the subset of the set of all terminal nodes in the binary tree ($E_m \subset$ {(-1,0), (-1,1), (-2,0), (-2,1), (-2,2), (-2,3)}) and has to satisfy condition, defined by Eq. (3.14).

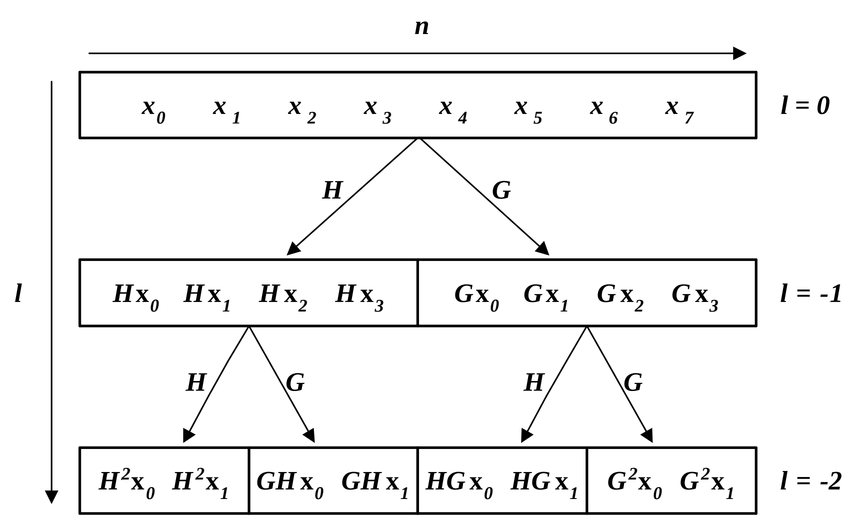


Figure 3.1: Discrete wavelet packet decomposition coefficients on $\mathbb{R}^8$: 2 decomposition levels.

The *discrete wavelet transform* (DWT) (Fig. 3.2) is a particular case of WPD, where the coefficients obtained by high-pass filtering are not transformed further.

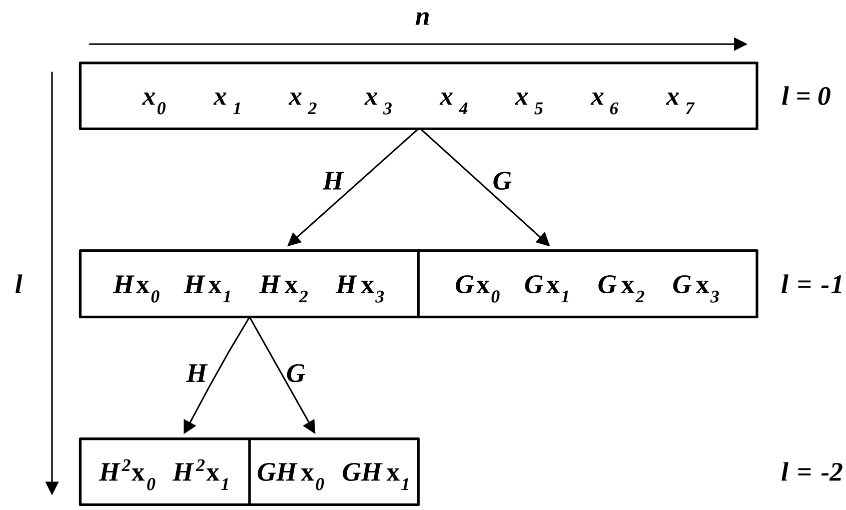


Figure 3.2: Discrete wavelet transform coefficients on $\mathbb{R}^8$: 2 decomposition levels.

The expansion coefficients associated with a prescribed signal $\mathbf{x}$ can be computed efficiently, using the operators H and G (Fig. 3.1), according to the well known decomposition into subband signals [39].

3.1.3 Best Basis Selection and Cost Functions

Of all bases $\{B_m\}$ contained in the library $\mathcal{B}$ one would like to pick up the *best basis* according to some criterion. A search algorithm is clearly needed.

Let $\mathcal{M}$ denote some *"information cost"* function. The cost function is said to be *additive* (a feature that directly effects the efficiency of the search procedure) if

$$\mathcal{M}(0) = 0 \text{ and } \mathcal{M}\left(\Theta)\right) = \sum_{j \in \mathbb{Z}} \mathcal{M}(\theta_j), \tag{3.16}$$

where $\Theta = B\mathbf{f} = \{\theta_j : j \in \mathbb{Z}\}$ is the expansion vector of $\mathbf{f} = \{f(i)\}_{i=0}^{N-1}$ on the basis B. Examples of such an additive cost functions are [38]:

1) The Shannon entropy:

$$\mathcal{H}(\Theta) = \left(- \sum_j \frac{\theta_j^2}{||\mathbf{f}||_2^2} \ln \frac{\theta_j^2}{||\mathbf{f}||_2^2} \right) \tag{3.17}$$

2) The log energy:

$$\mathcal{E}(\Theta) = \left(\sum_j \ln \frac{\theta_j^2}{||\mathbf{f}||_2^2}\right) \tag{3.18}$$

3) The concentration in ℓ^1 norm:

$$\ell^1(\Theta) = \left(\sum_j \left|\frac{\theta_j}{||\mathbf{f}||_2}\right|\right) \tag{3.19}$$

Minimization of $\mathcal{M}(B\mathbf{f})$ will yield to the "best basis".

We have seen that the library $\mathcal{B}$ is organized as a binary tree (Fig. 3.1) whose maximal depth is $J = \log_2 N$ resolution levels, where N is the length of the signal. Thus, the best basis can be found by computing the information cost at each node and comparing "children"-"parent" costs. Starting at the lowest decomposition level ($\ell = -L,\ 1 \leq L \leq J$), we move upward and include the "children"-nodes in the best basis tree if their cost is less then the cost of the "parent"-node, otherwise, we chose the "parent"-node. The basis, that is characterized by the minimal total information cost, will be identified as the best basis.

Owing to the presumed additive nature of the information cost function, each node needs to be examined twice (once as a "child" and once as a "parent"), leading to a low-complexity scheme [9, 43]. Decomposing the signal of length N with FIR filters of length r requires $O(rNL) \leq O(rN \log_2 N)$ real-valued products, computing the cost-function - $O(NL)$, search for best basis is of $O(2N)$. Thus, the complexity of the *adaptive* WPD process is of $O(rNL)$. For $L = J$ the complexity is of $O(rN \log_2 N)$.

3.1.4 Multiresolution Analysis

The time-frequency analysis using WPD and related techniques is referred to as "multiresolution analysis". The "multiresolution analysis" consists of a characteristic function

$\varphi(t)$ and a family of closed subspaces $U_{\ell,n} \subset L^2(\mathbb{R})$, defined by

$$U_{\ell,n} = \text{span}[\{\psi_{\ell,n,k}(t) = 2^{\ell/2}\psi_n(2^\ell t - k) : k \in \mathbb{Z}\}]. \tag{3.20}$$

The family of closed subspaces $U_{\ell,n}$ has to satisfy [26]:

1) $U_{\ell,n}$ can be decomposed into the following orthonormal direct sum

$$U_{\ell,n} = U_{\ell-1,2n} \bigoplus U_{\ell-1,2n+1}, \tag{3.21}$$

(implying $U_{\ell-1,2n} \perp U_{\ell-1,2n+1}$),

2) $\bigcap_{\ell=const, n\in\mathbb{Z}_+} U_{\ell,n} = \{0\}$ (*downward completeness*)),

3) $\bigcup_{\ell\in\mathbb{Z}, n\in\mathbb{Z}_+} U_{\ell,n} = L^2(\mathbb{R})$ (*upward completeness*),

4) $x(t) \in U_{\ell,0}$ implies $x(2t) \in U_{\ell-1,0}$,

5) $x(t) \in U_{0,0}$ implies $x(t-\tau) \in U_{0,0}$ for all $\tau \in \mathbb{R}$ ("shift invariance"),

6) There exists $\varphi(t) \in U_{0,0}$, such that for all $\ell \in \mathbb{Z}$, $\{\varphi_{\ell,k}(t)\}$ constitutes an orthonormal basis for

$$U_{\ell,0} = \text{span}[\{\varphi_{\ell,k}, k \in \mathbb{Z}\}]. \tag{3.22}$$

The previously mentioned filtering operators H and G can be viewed as orthogonal projection operators from $U_{\ell,n}$ onto the subspaces $U_{\ell-1,2n}$ and $U_{\ell-1,2n+1}$ respectively.

3.2 Shift-Invariant Wavelet Packet Decompositions

3.2.1 Introduction

It can be shown that as $f(t)$ is decomposed into orthonormal wavelet packets best basis, the property of shift-invariance is no longer valid [5].

A strategy for re-introducing shift-invariance has been proposed in [3]. It is based on extending the wavelet packet library to include all shifted versions of the bases (hence the name Shifted Wavelet Packet library), organizing it into a tree structure and providing an efficient "best-basis" search algorithm.

In implementing the *Shift-Invariant Wavelet Packet Decomposition* (SIWPD) algorithm, shift-invariance is achieved by the introduction of an additional degree of freedom. The added dimension is a *relative shift* between a given parent-node and its respective children-nodes. Specifically, upon expanding a prescribed node, with minimization of the information cost in mind, we test as to whether or not the information cost indeed decreases. For any given parent-node it is sufficient to examine and select one of two alternative decompositions, made feasible by the Shifted Wavelet Packet (SWP) library. These decompositions correspond to a zero shift and a $2^{-\ell}$ shift where ℓ $(-L \leq \ell \leq 0)$ denotes the resolution level. An alternative view of SIWPD is facilitated via filter-bank terminology. Accordingly, each parent-node is expanded by high-pass and low-pass filters, followed by a 2:1 down-sampling. In executing WPD, down-sampling is achieved by ignoring all even-indexed (or all odd-indexed) terms. In contrast, when pursuing SIWPD, the down-sampling is carried out *adaptively* for the prescribed signal.

3.2.2 Shifted Wavelet Packet Library

Let us consider the following extended forms of shifted and translated scaling and mother wavelet functions:

$$\varphi_{\ell,k}^{(m)}(t) = 2^{\ell/2}\varphi(2^\ell t - 2^\ell m - k), \tag{3.23}$$

$$\psi_{\ell,n,k}^{(m)}(t) = 2^{\ell/2}\psi_n(2^\ell t - 2^\ell m - k), \tag{3.24}$$

where m is the *relative shift* index: $m \in \{0,1\}$.

The SWP library is defined as the collective set

$$\mathcal{B} = \{\psi_{\ell,n,k}^{(m)}(t) = 2^{\ell/2}\psi_n(2^\ell t - 2^\ell m - k) : \ell \in \mathbb{Z}_-, n \in \mathbb{Z}_+, k \in \mathbb{Z}, m \in \{0,1\}\}. \tag{3.25}$$

This library is larger than the WP library by a square power, but it can still be cast into a tree configuration facilitating fast search algorithms [5].

Similarly to Eq. (3.11) we can define

$$h_k^{(m)} = \langle \varphi_{1,k}^{(m)}, \varphi \rangle = \sqrt{2}\langle \varphi(2t - k - 2m), \varphi(t)\rangle = h_{k+2m}. \tag{3.26}$$

Moreover, using the previously defined filtering operators H and G, we can define the operators $H^{(0)}$ and $G^{(0)}$ of low and high-pass filtering respectively, followed by a 2:1 downsampling, and the operators $H^{(1)}$ and $G^{(1)}$ of low and high-pass filtering respectively, followed by a unit sample delay and a 2:1 downsampling (Fig. 3.3).

Similar to WPD, SIWPD can be associated with a set of closed subspaces $U_{\ell,n}^{(m)} \subset L^2(\mathbb{R})$, each defined by

$$U_{\ell,n}^{(m)} = \operatorname{span}[B_{\ell,n}^{(m)} = \{\psi_{\ell,n,k}^{(m)}(t) : k \in \mathbb{Z}\}]. \tag{3.27}$$

Each of the subspaces $U_{\ell,n}^{(m)}$ can be decomposed into

$$U_{\ell,n}^{(m)} = U_{\ell-1,2n}^{(m)} \bigoplus U_{\ell-1,2n+1}^{(m)}$$

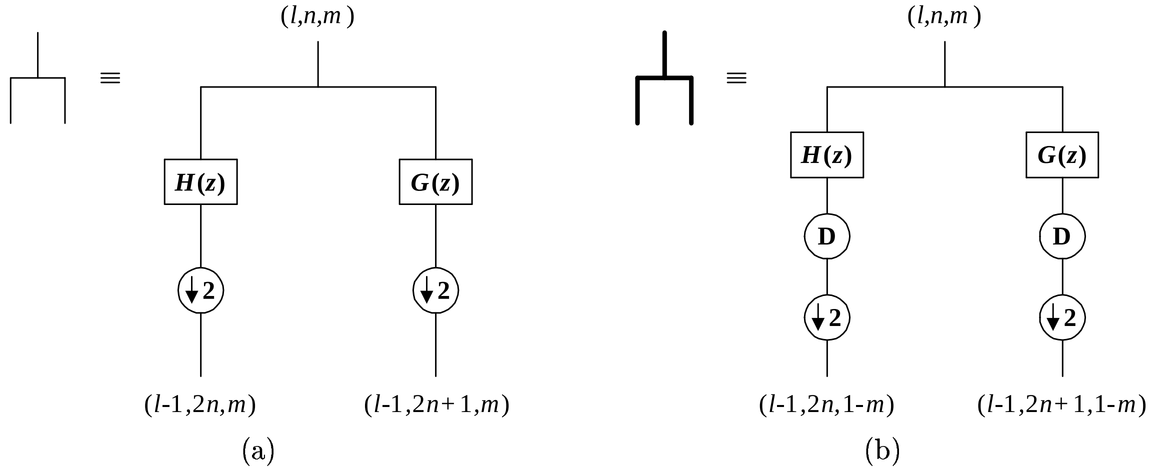


Figure 3.3: A "parent" node binary expansion according to SIWPD: (a) $H^{(0)}$ and $G^{(0)}$ filtering operators: low and high-pass filtering followed by a 2:1 downsampling, (b) $H^{(1)}$ and $G^{(1)}$ filtering operators: low and high-pass filtering followed by a unit sample delay (D) and subsequently by a 2:1 downsampling. Each node is defined by the triplet (ℓ, n, m).

or into

$$U_{\ell,n}^{(m)} = U_{\ell-1,2n}^{(1-m)} \bigoplus U_{\ell-1,2n+1}^{(1-m)}.$$

The tree structure of SWP is depicted in Fig. 3.4. Each node in the tree is indexed by the triplet (ℓ, n, m) and represents the subspace $U_{\ell,n}^{(m)}$. Similarly to the WP binary trees [9], the nodes are identified with dyadic intervals of the form $I_{\ell,n} = [2^\ell n,\ 2^\ell(n+1))$. The additional parameter m facilitates a time-shift adjustment of the basis functions. The generated branches are respectively depicted by thin (no delay) or heavy (a unite delay) lines (Fig. 3.5).

Let $E = \{(\ell, n, m)\} \subset \{\mathbb{Z}_- \times \mathbb{Z}_+ \times \{0,1\}\}$ denote a collection of indices. If the segments $I_{\ell,n} = [2^\ell n, 2^\ell(n+1))$ are a disjoint cover of $[0,1)$, then E generates an orthonormal basis for $U_{0,0}^{(m)}$ [5]. The expansion tree associated with a given signal describes the signal's representation on an orthonormal basis selected from the SWP library. The index set E is interpreted as the collection of all terminal nodes. That is, all nodes beyond which no further expansions are to be carried out.

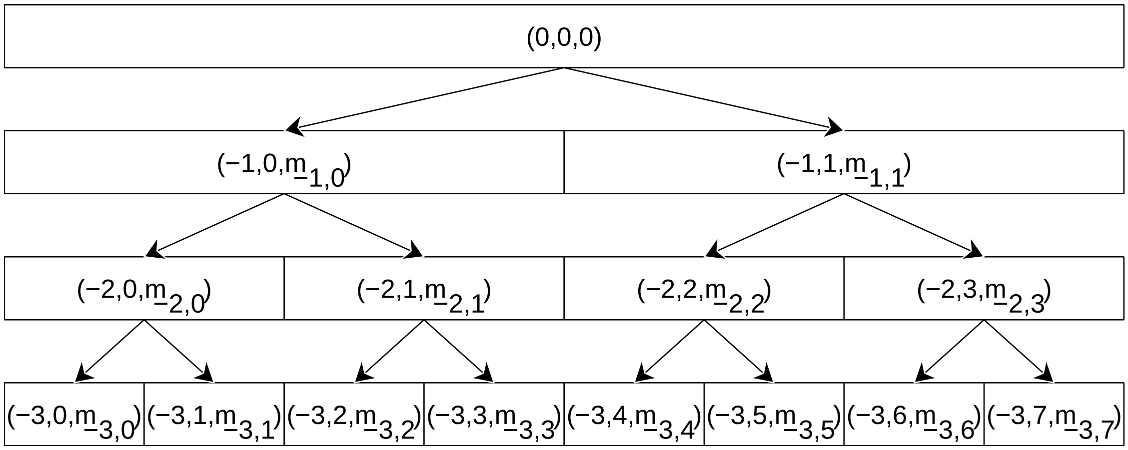


Figure 3.4: The extended set of wavelet packets organized in a binary tree structure.

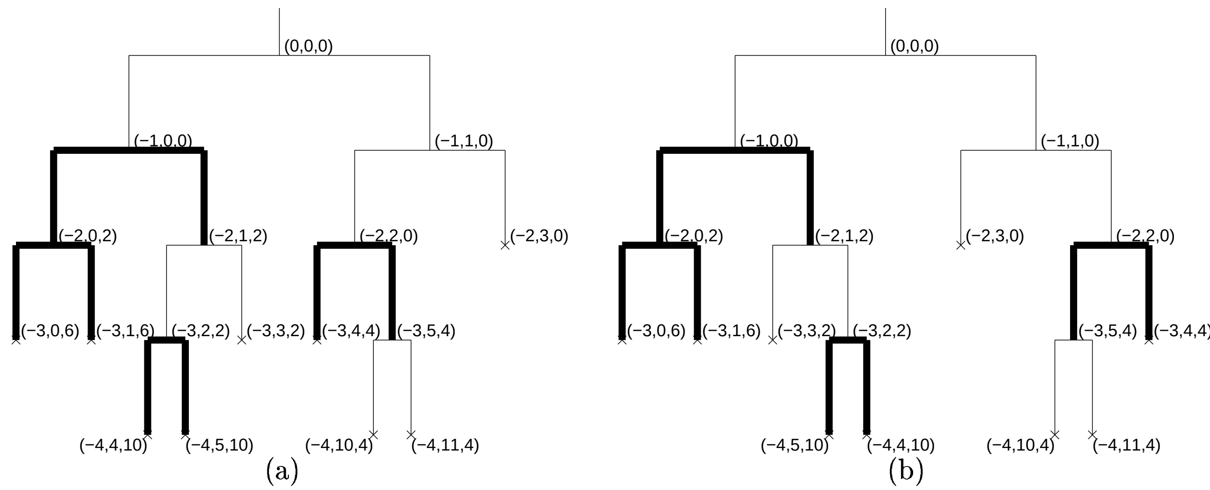


Figure 3.5: An example of a SIWPD binary tree. (a) The children-nodes corresponding to (ℓ, n, m) are $(\ell - 1, 2n, \tilde{m})$ and $(\ell - 1, 2n + 1, \tilde{m})$, where $\tilde{m} = m$ (depicted by thin lines) or $\tilde{m} = 1 - m$ (depicted by heavy lines). (b) Rearrangement of the nodes in a *sequency* order.

A specific example of an expansion tree is shown in Fig. 3.5(a). The nodes at each decomposition level in this example have a natural or *Paley* order. It is normally useful to rearrange them in a *sequency* order [43], so that the nominal frequency of the associated wavelet packets increases monotonically as we move from left to right along a given tree level. The exchange is carried out efficiently using the inverse Gray code permutation [43]. The resultant tree is depicted in Fig. 3.5(b).

Like the wavelet packet library [9], the tree configuration of the extended library facilitates an efficient best basis selection process. However, in contrast to the WPD, the best-basis representation is now shift-invariant.

3.2.3 The Best Basis Selection

This section follows closely the sequence and notations presented in [5]. Let $\mathbf{f} = f(t) \in U_{0,0}^{(0)}$; let $\mathcal{M}$ denote an additive cost function, and let $\mathcal{B}$ represent a SWP library. As for the WPD, the best basis for $\mathbf{f}$ in $\mathcal{B}$ with respect to $\mathcal{M}$ is $B \in \mathcal{B}$ for which $\mathcal{M}(B\mathbf{f})$ is minimal. Here, $\mathcal{M}(B\mathbf{f})$ is the information cost of representing $\mathbf{f}$ in the basis $B \in \mathcal{B}$.

Let $A_{\ell,n}^{(m)}$ denote the best basis for the subspace $U_{\ell,n}^{(m)}$. The desired best basis can be determined recursively by setting

$$A_{\ell,n}^{(m)} = \begin{cases} B_{\ell,n}^{(m)} \text{ if } \mathcal{M}(B_{\ell,n}^{(m)}\mathbf{f}) \leq \mathcal{M}(A_{\ell-1,2n}^{(m_c)}\mathbf{f}) + \mathcal{M}(A_{\ell-1,2n+1}^{(m_c)}\mathbf{f}), \\ \\ A_{\ell-1,2n}^{(m_c)} \oplus A_{\ell-1,2n+1}^{(m_c)}, \text{ otherwise,} \end{cases} \tag{3.28}$$

where the shift indices of the respective children-nodes are given by

$$m_c = \begin{cases} m, \text{ if } \sum_{i=0}^{1} \mathcal{M}(A_{\ell-1,2n+i}^{(m)}\mathbf{f}) \leq \sum_{i=0}^{1} \mathcal{M}(A_{\ell-1,2n+i}^{(1-m)}\mathbf{f}) \\ \\ 1-m, \quad \text{otherwise.} \end{cases} \tag{3.29}$$

The recursive sequence proceeds down to a specified level $\ell = -L$ $(1 \leq L \leq \log_2 N)$, where

$$A_{-L,n}^{(m)} = B_{-L,n}^{(m)}. \tag{3.30}$$

The recursive algorithm proposed in [9] for a best basis search in WP library may be viewed as a special case where $(m_c - m)$ is arbitrarily set to zero. Thus the algorithm searches only through non-shifted bases and the selected basis will be a WPD basis that does not possess shift-invariance. Moreover, the property of shift-invariance can also be achieved within the framework of the wavelet transform (WT) and a prescribed information cost function ($\mathcal{M}$). It may be viewed as a special case whereby the tree configuration is constrained to expanding exclusively the *low frequency* nodes [3].

So far we have observed that WPD lacks shift-invariance but is characterized by an attractive complexity level $O(rNL)$, where L denotes the lowest resolution level in the expansion tree. Comparatively, the quadratic complexity level, $O(rN2^{L+1})$ [5], associated with SIWPD is substantially higher. In return, one may achieve a potentially large reduction of the information cost, in addition to gaining the all important *shift-invariance.* However, whenever the SIWPD complexity is viewed as intolerable, one may resort to a sub-optimal SIWPD procedure entailing a reduced complexity, and higher information cost while still retaining the desirable shift-invariance [3]. In this case, the depth of a subtree, used at a given parent-node to determine its shift index, is restricted to d resolution levels $(1 \leq d \leq L)$, and the computational complexity reduces to $O[rN2^d(L-d+2)]$. In the extreme case $d=1$, the complexity, $O(rNL)$, is similar to that associated with the conventional WPD. The larger d and L, the larger the complexity, however, the determined optimal basis generally yields a lower information cost.

3.3 Local Trigonometric Decompositions

3.3.1 Introduction

In a sense, local trigonometric decompositions (LTD) [1, 9, 20] can be considered as conjugates of the wavelet packet decompositions, where the partitioning of the frequency axis is replaced by the partitioning of the time axis. With this decomposition, a prescribed signal is first split into overlapping intervals. Then a folding operator [43] "folds" overlapping parts into segments, and a standard cosine or sine transform is applied to each segment. In this case, the basis functions are cosines or sines multiplied by smooth window functions.

Local trigonometric set can be organized into a binary-tree structured library of orthonormal bases. The best basis, minimizing a prescribed information cost function, is searched using the divide-and-conquer algorithm [9].

A library of local trigonometric bases can be extended to a library of *shift-invariant adaptive polarity* local trigonometric bases [4].

3.3.2 Smooth Local Trigonometric Bases

Let's consider a partition of the line with a set of disjoint intervals $I_j = [a_j, a_{j+1})$, such that the width of the intervals is never less than a fixed positive number ($a_{j+1} - a_j \geq 2\epsilon > 0$) for all $j \in \mathbb{Z}$:

$$\mathbb{R} = \bigcup_{j \in \mathbb{Z}} I_j. \tag{3.31}$$

Let $r(t)$ be a function in the class C^s for some $s > 0$, satisfying the following conditions:

$$r(t) = \begin{cases} 0, & \text{if } t \leq -1, \\ 1, & \text{if } t > 1, \end{cases} \tag{3.32}$$

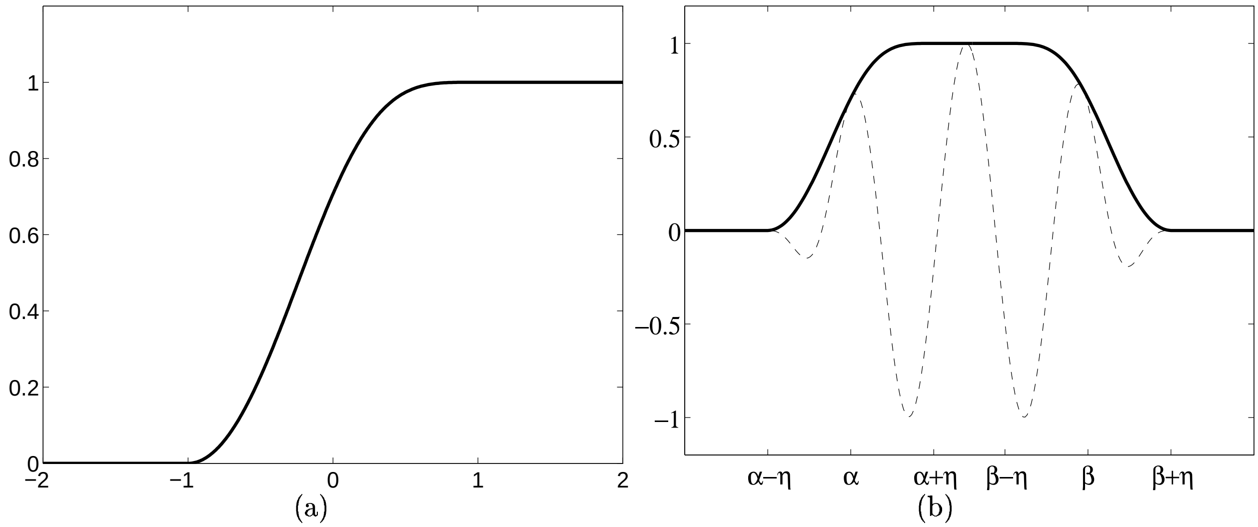


Figure 3.6: (a) An example of a right cut-off function in C^1. (b) The corresponding window function on $[\alpha, \beta]$ for $\eta < (\beta - \alpha)/2$ (solid), and a modulated function (dashed).

$$|r(t)|^2 + |r(-t)|^2 = 1 \quad \text{for all } t \in \mathbb{R}. \tag{3.33}$$

It is the so-called "*right cut-off function*". An example of a continuously differentiable real-valued right cut-off function $r_1(t) \in C^1$ is given by

$$r_1(t) = \begin{cases} 0, & \text{if } t \leq -1, \\ \sin[\frac{\pi}{4}(1 + \sin\frac{\pi}{2}t)], & \text{if } -1 < t < 1, \\ 1, & \text{if } t \geq 1, \end{cases} \tag{3.34}$$

and is depicted in Fig. 3.6(a).

A window function $b_j(t)$, which is supported on the interval $[a_j - \eta, a_{j+1} + \eta]$ is defined by

$$b_j(t) = r\left(\frac{t - a_j}{\eta}\right) r\left(\frac{a_{j+1} - t}{\eta}\right), \tag{3.35}$$

where $0 < \eta \leq \epsilon$. Here η is interval on which a window function rises from being identically zero to being identically one, it allows overlap of windows and controls the smoothness of the window function. Multiplying a window function by some modulating trigonometric function we obtain a smooth local trigonometric function:

$$\Psi_{j,k}(t) = b_j(t) F_{j,k}(t), \tag{3.36}$$

where $F_{j,k}(t)$ is a modulating function, and $|I_j| = (a_{j+1} - a_j)$ is the length of the j-th disjoint interval. Each local trigonometric function $\Psi_{j,k}$ is well localized in both time and frequency. The basis function $\Psi_{j,k}(t)$ is supported on the same interval as the window function $b_j(t)$. The set of the functions $\Psi_{j,k}(t)$ with $j \in \mathbb{Z}$ and $k \in \mathrm{N}$ is an orthonormal basis for $L^2(\mathbb{R})$. Consequently, each signal $f(t) \in L^2(\mathbb{R})$ can be written in terms of the functions $\Psi_{j,k}(t)$:

$$f(t) = \sum_{j \in \mathbb{Z},\ k \in \mathrm{N}} c_{j,k} \Psi_{j,k}(t) \tag{3.37}$$

with

$$c_{j,k} = \langle f(t), \Psi_{j,k}(t) \rangle \,. \tag{3.38}$$

An example of modulating functions can be given by the basis functions of the so-called DST-IV transform - discrete sampled sines at half-integer frequencies:

$$F_{j,k}(t) = \frac{\sqrt{2}}{\sqrt{|I_j|}} \sin\left(\pi \left(k + \frac{1}{2}\right) \frac{t - a_j}{|I_j|}\right). \tag{3.39}$$

Here $\Psi_{j,k}$ is supported in time on $[a_j - \eta, a_{j+1} + \eta]$ and thus has a position uncertainty, that is equal, at most, to the width of the compact interval I_j. In the frequency domain, it consists of two bumps centered at $\pm(2k+1)/(a_{j+1} - a_j)$, with an uncertainty determined by the support of the Fourier transform of the window function.

3.3.3 Fast Implementations

The inner products in (3.38) can be efficiently computed using a standard fast discrete trigonometric transforms (such as DCT-I, DCT-II, DCT-IV, DST-II, DST-IV) [43], after a preliminary "folding" step. We consider the DST-IV transform, as an example.

Let's define by M_α [20] the *mirror operator* around point α:

$$M_\alpha f(t) = f(2\alpha - t). \tag{3.40}$$

It's unitary and essentially flips the function around α. We denote by $r_\alpha(t)$ the right cut-off function centered at $t = \alpha$:

$$r_\alpha(t) = r\left(\frac{t-\alpha}{\eta}\right). \tag{3.41}$$

The left cut-off function $l_\alpha(t)$ will then be given by

$$l_\alpha = M_\alpha r_\alpha. \tag{3.42}$$

Let $\mathbf{1}_I$ be an indicator function for the interval I - it is equal to 1 if $t \in I$ and is equal to 0 otherwise. Accordingly let $\mathbf{1}^l_\alpha(t) = \mathbf{1}_{(-\infty,\alpha]}$ and $\mathbf{1}^r_\alpha(t) = \mathbf{1}_{[\alpha,\infty)}$.

The *folding operator* around a point α is defined by

$$Q_\alpha = \mathbf{1}^l_\alpha(1 + M_\alpha)l_\alpha + \mathbf{1}^r_\alpha(1 - M_\alpha)r_\alpha \tag{3.43}$$

and it is unitary if r_α satisfies the condition (3.33). The adjoint of the Q_α operator is given by

$$Q^*_\alpha = l_\alpha(1 + M_\alpha)\mathbf{1}^l_\alpha + r_\alpha(1 - M_\alpha)\mathbf{1}^r_\alpha. \tag{3.44}$$

Below we discuss several properties of the folding operator (3.43). Multiplication with l_α ensures a smooth decay to the left of $(\alpha + \eta)$. The operator $(1 + M_\alpha)$ then adds this function to its mirrored version. This generates an even around α function. This function is now cut off by $\mathbf{1}^l_\alpha$. The right part is similar and generates an odd function. Consequently, if $f(t)$ is smooth, then $\mathbf{1}^l_\alpha Q_\alpha f$ is a smooth function when extended in an "even" fashion to the right and $\mathbf{1}^r_\alpha Q_\alpha f$ is a smooth function when extended in an "odd" fashion to the left (by extending in an "even" and "odd" fashions we mean applying the operators (1+M) and (1-M), respectively). The adjoint operator (3.44) (which is also the inverse) does essentially the same but switches the parity properties: even to odd and vice versa.

Our basic goal is to split $L^2(\mathbb{R})$ into subspaces such that each subspace will contain functions that are localized around one of the intervals I_j. Moreover, we want a basis that is suitable for representing smooth functions on that interval. The splitting of $L^2(\mathbb{R})$ into subspaces is done using the orthogonal projection operator:

$$P_{I_j} = T^* \mathbf{1}_{I_j} T, \tag{3.45}$$

where T is the *total folding operator* which is defined over $\mathbb{R}$ and is given by

$$T = \prod_j Q_{a_j}. \tag{3.46}$$

Since T is a product of unitary operators, it is necessarily unitary. The total folding operator transforms a smooth function into a function with specific parity properties at the endpoints of each interval I_j. Then, in order to get good approximation properties we should use a trigonometric basis which reflects these parities (that's the reason why previously defined folding operator Q_α and consequently the total folding operator T are adopted to DST-IV bases functions).

We decompose $L^2(\mathbb{R})$ into orthogonal subspaces as

$$L^2(\mathbb{R}) = \bigoplus_j V_{I_j} \quad \text{with} \quad V_{I_j} = P_{I_j} L^2(\mathbb{R}). \tag{3.47}$$

An orthonormal basis for V_{I_j} is given by $\{\Psi_{j,k}(t)\}_{k\in\mathrm{N}}$, and consequently

$$f(t) = \sum_{j\in\mathbb{Z},\ k\in\mathrm{N}} P_{I_j} f(t) = \sum_{j\in\mathbb{Z},\ k\in\mathrm{N}} c_{j,k}\Psi_{j,k}(t), \tag{3.48}$$

where the coefficients are given by

$$c_{j,k} = \langle f(t), \Psi_{j,k}(t)\rangle = \left\langle Tf(t), \mathbf{1}_{I_j}\Psi_{j,k}(t)\right\rangle. \tag{3.49}$$

3.3.4 Tree-Structured Library of Bases

The basis functions on interval $[\alpha, \beta)$ are the orthogonal direct sum of the basis functions on its left and right halves ($[\alpha, \frac{\beta-\alpha}{2})$ and $[\frac{\beta-\alpha}{2}, \beta)$ respectively). Thus recursive subdivision of the intervals I_j into halves will propagate this orthogonality through the multiple levels and this subdivision will build a binary tree. The complete local trigonometric basis will be the orthogonal direct sum of the basis functions on each subinterval of $\mathbb{R}$.

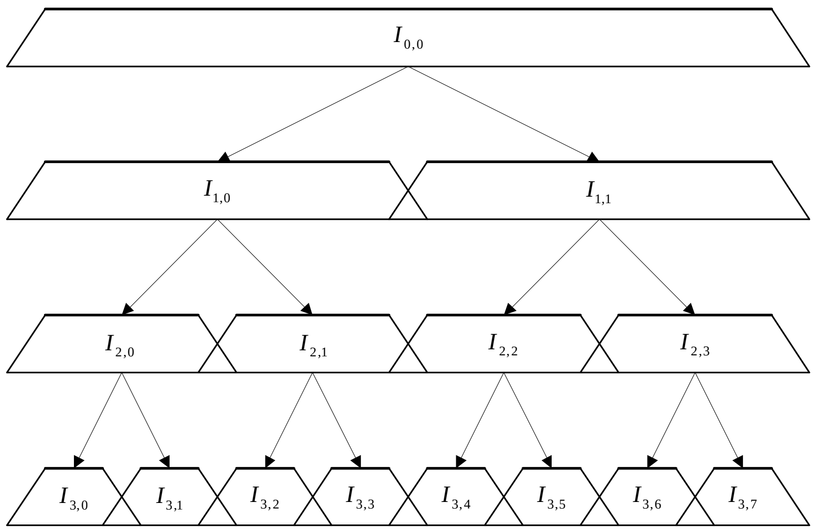


Figure 3.7: Organization of the smooth local trigonometric bases in a binary tree structure.

The organization of local trigonometric library into a binary tree facilitates an efficient search for a suitable best basis. The best local trigonometric basis is again selected by searching for the minimum of an additive cost function. The computational complexity of the LTD with the best-basis selection algorithm is $O(LN \log_2 N)$, where L is the number of decomposition levels. For $L = J$, the total complexity is of $O(N \log_2 N^2)$.

Chapter 4 : Wavelet-Based Denoising Techniques

4.1 Wavelet Domain Denoising: The Donoho-Johnstone Algorithm

4.1.1 Problem Formulation

Suppose we are given noisy data $\mathbf{y} = \{y_i\}_{i=0}^{N-1}$ with $N = 2^J$, where

$$y_i = f_i + e_i, \quad i = 0, ..., N-1, \tag{4.1}$$

$\mathbf{f} = \{f_i\}_{i=0}^{N-1}$ is an unknown discrete real-valued signal which we would like to recover, and $\mathbf{e} = \{e_i\}_{i=0}^{N-1}$ is a white Gaussian noise (WGN) with zero mean and a presumably known power spectral density (PSD) σ^2. Let $\hat{\mathbf{f}} = \{\hat{f}_i\}_{i=0}^{N-1}$ denote the vector of estimated sample values. We now introduce basic denoising techniques using the example of a finite discrete wavelet transform.

The vector $\mathbf{w} = \{w_{\ell,n,k}\}$ of wavelet expansion coefficients of the noisy data $\mathbf{y}$ is defined by

$$\mathbf{w} = W\mathbf{y}, \tag{4.2}$$

where ℓ is the resolution level index (the scaling parameter), n is the oscillation (modulation) index, k is the time-domain position index and W is the finite orthonormal wavelet transform matrix. The orthonormality of W yields the following reconstruction formula: $\mathbf{y} = W^T\mathbf{w}$.

Under the noise model underlying (4.1), noise contaminates all wavelet coefficients equally: The noise vector $\mathbf{e}$ is assumed to represent WGN so that its orthogonal transform $\mathbf{z} = W\mathbf{e} = \{z_{\ell,n,k}\}$ is also WGN. Consequently, applying the wavelet transform to $\mathbf{y}$ yields

$$w_{\ell,n,k} = \theta_{\ell,n,k} + z_{\ell,n,k}, \tag{4.3}$$

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

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

$$\hat{\mathbf{f}} = W^T\hat{\Theta} = W^T\{T\mathbf{w}\} = W^T\{T\{W\mathbf{y}\}\}, \tag{4.4}$$

where T is generally a nonlinear operator, that determines the set Γ and performs noise subtraction in the wavelet domain. Clearly, the quality of the resulting estimation $\hat{\mathbf{f}}$ depends on the algorithm, that determines Γ and the gain function.

We measure the quality of $\hat{\mathbf{f}}$ in terms of quadratic loss at the sampling points. Let

$$||v||_{2,N}^2 = \sum_{i=0}^{N-1} v_i^2 \tag{4.5}$$

denote the usual squared l_N^2 norm. Thus, we measure performance by the risk

$$R(\hat{\mathbf{f}}, \mathbf{f}) = N^{-1}E||\hat{\mathbf{f}} - \mathbf{f}||_{2,N}^2, \tag{4.6}$$

which we would like to minimize. Owing to its orthonormality, W transforms estimators in one domain into estimators in the other domain with isometry of risks:

$$||\hat{\Theta} - \Theta||_2 = ||\hat{\mathbf{f}} - \mathbf{f}||_2. \tag{4.7}$$

Thus, minimization of $R(\hat{\Theta}, \Theta)$ implies minimization of $R(\hat{\mathbf{f}}, \mathbf{f})$.

4.1.2 Thresholding Types

Thresholding techniques are effective whenever few wavelet coefficients contribute to the noiseless signal. Let's consider *threshold rules*, that retain only observed data, which exceeds a predetermined multiple of the noise level. Often used thresholding techniques are [16]:

- *Hard Thresholding* :

$$\eta_h(x,t) = x \cdot \mathbf{1}_{(|x|-t)}, \tag{4.8}$$

where

$$\mathbf{1}_{(u)} = \begin{cases} 1, u > 0 \\ 0, u \leq 0 \end{cases} \tag{4.9}$$

- *Soft Thresholding* :

$$\eta_s(x,t) = (|x| - t) \cdot \mathrm{sign}(x) \cdot \mathbf{1}_{(|x|-t)}. \tag{4.10}$$

4.1.3 The RiskShrink Estimator

The *RiskShrink* and *VisuShrink* estimators were introduced by Donoho and Johnstone in [14]. The name *RiskShrink* emphasizes that modification of wavelet coefficients is performed by soft thresholding, and that a mean squared error, or "risk" approach has been taken to specify the threshold.

The *RiskShrink* estimator is defined by

$$\widehat{\mathbf{f}}^{(Risk)} \equiv W^T\{\widehat{\Theta}^{(Risk)}(\mathbf{w}, \lambda, \sigma)\}, \tag{4.11}$$

where

$$\widehat{\Theta}^{(Risk)}(\mathbf{w}, \lambda, \sigma) = \{\widehat{\theta}^{(Risk)}_{\ell,n,k}(w_{\ell,n,k}, \lambda, \sigma)\}_{\ell,n,k} \tag{4.12}$$

and

$$\widehat{\theta}^{(Risk)}_{\ell,n,k}(w_{\ell,n,k}, \lambda, \sigma) = \begin{cases} w_{\ell,n,k}, & \ell + J < j_0, \\ \eta_s(w_{\ell,n,k}, \lambda\sigma), & j_0 \leq \ell + J < J. \end{cases} \tag{4.13}$$

Here, σ^2 is the variance of the additive noise $\mathbf{e}$ and $\lambda = \lambda(\ell, J)$ is the threshold value that depends on the resolution level ℓ and the maximal depth of a decomposition tree $J = \log_2 N$. j_0 is the so-called low-resolution cutoff. Note that at levels $\ell < j_0 - J$ the basis functions $\{\varphi_{\ell,k}(t)\}$ do not have vanishing means, thus wavelet coefficients at these resolution levels should not be shrunken towards zero. The values of λ for different ℓ and J were computed by Donoho and Johnstone [14] and can be embedded as a look-up table:

$\ell + J$	6	7	8	9	10	11	12	13	14	15	16
λ	1.474	1.669	1.860	2.048	2.232	2.414	2.594	2.773	2.952	3.131	3.310

Table 4.1: Look-up table of λ dependent on resolution level ℓ.

Donoho and Johnstone point out that whenever we rely exclusively on the data, the *RiskShrink* estimator mimics the risk in the best possible way. Moreover, they prove that the RiskShrink performs better than alternative estimators that are based on the selective wavelet reconstruction principle.

4.1.4 The VisuShrink Estimator

The *VisuShrink* estimator [14] resembles *RiskShrink*. The difference stems from the use of a universal threshold $\lambda_d = (2 \ln d)^{1/2}$ (the asymptotic value of the λ) instead of λ:

$$\widehat{\mathbf{f}}^{(Visu)} \equiv W^T \{\widehat{\Theta}^{(Visu)}(\mathbf{w}, \lambda_d, \sigma)\}, \tag{4.14}$$

where

$$\widehat{\theta}^{(Visu)}_{\ell,n,k}(w_{\ell,n,k}, \lambda_d, \sigma) = \begin{cases} w_{\ell,n,k}, & \ell + J < j_0, \\ \eta_s(w_{\ell,n,k}, \sigma\sqrt{2\ln d}), & j_0 \leq \ell + J < J. \end{cases} \tag{4.15}$$

Here $d = 2^{(\ell+J)}$ is the number of wavelet coefficients in each subband at the ℓ-th resolution level. No look-up table is needed and the threshold can be easily calculated as $\sigma\sqrt{2(\ell+J)\ln 2}$ [14]. Moreover, this estimator has an important *visual advantage* resulting from the almost "noise-free" character of reconstructions. This can be explained as follows. When $\{z_i\}$ is a white noise sequence i.i.d. $\mathcal{N}(0, \sigma^2)$, then

$$pr\{\max_i |z_i| > \sigma(2\ln d)^{1/2}\} \to 0, \qquad d \to \infty.$$

So that, with high probability, every wavelet transform sample, where the underlying signal is exactly zero, will indeed be estimated as zero.

The drawback of this simple threshold formula is that the MSE performance of adaptive thresholds (like *SureShrink*) is noticeably better.

4.1.5 The SureShrink Estimator

The *SureShrink* estimator [16] suppresses noise by thresholding the empirical wavelet coefficients adaptively. A threshold value is assigned to each resolution level using the minimization principle of *Stein's Unbiased Estimate of Risk* (*Sure*).

The *SureShrink* estimator is defined by

$$\widehat{\mathbf{f}}^{(Sure)} \equiv W^T\{\widehat{\Theta}^{(Sure)}(\mathbf{w}, \{t_{\ell,n}^{(Sure)}\}_{(\ell,n)\in E}, \sigma)\}, \tag{4.16}$$

where

$$t_{\ell,n}^{(Sure)}(\mathbf{w}_{\ell,n}, \sigma) = \arg \min_{0\leq t\leq \lambda_d} SURE(\mathbf{w}_{\ell,n}, t, \sigma) \tag{4.17}$$

is the *SureShrink* threshold value for shrinking the wavelet coefficients $\mathbf{w}_{\ell,n} = \{w_{\ell,n,k}\}_{k=0}^{2^{(\ell+J)}-1}$ that belong to the WPD tree node (subband), indexed by the pair (ℓ, n). E denotes the tree-set of indices (ℓ, n), $\lambda_d = \sqrt{2(\ell + J)\ln 2}$,

$$SURE(\mathbf{w}_{\ell,n}, t, \sigma) = 2^{(\ell+J)} - 2\cdot\#\{k : |w_{\ell,n,k}| \leq t\sigma\} + \sum_{k=0}^{2^{(\ell+J)}-1}\{\min(|w_{\ell,n,k}|, t\sigma)\}^2 \tag{4.18}$$

is the unbiased estimate of risk $E_\theta||\widehat{\Theta}_{\ell,n}^{(Sure)} - \Theta_{\ell,n}||_{2,2^{(\ell+J)}}^2$, and

$$\widehat{\Theta}_{\ell,n}^{(Sure)} = \{\widehat{\theta}_{\ell,n,k}^{(Sure)}(\mathbf{w}_{\ell,n}, t_{\ell,n}^{(Sure)}, \sigma)\}_{k=0}^{2^{(\ell+J)}-1}. \tag{4.19}$$

The estimate $\widehat{\theta}_{\ell,n,k}^{(Sure)}$ of the unknown signal $\mathbf{f}$ expansion coefficient $\theta_{\ell,n,k}$ is given by

$$\widehat{\theta}_{\ell,n,k}^{(Sure)}(\mathbf{w}_{\ell,n}, t_{\ell,n}^{(Sure)}, \sigma) = \begin{cases} w_{\ell,n,k}, & \ell + J < j_0, \\ \eta_s(w_{\ell,n,k}, \sigma\lambda_d), & j_0 \leq \ell + J < J \text{ and } s_{\ell,n}^2 \leq \eta_{\ell,n}/\sqrt{2^{(\ell+J)}}, \\ \eta_s(w_{\ell,n,k}, t_{\ell,n}^{(Sure)}\sigma), & j_0 \leq \ell + J < J \text{ and } s_{\ell,n}^2 > \eta_{\ell,n}/\sqrt{2^{(\ell+J)}}, \end{cases} \tag{4.20}$$

where $\eta_{\ell,n} = (\ell + J)^{3/2}$ and

$$s_{\ell,n}^2 \equiv \frac{1}{2^{(\ell+J)}} \sum_{k=0}^{2^{(\ell+J)}-1} \left((w_{\ell,n,k}/\sigma)^2 - 1\right). \tag{4.21}$$

Detailed explanations can be found in Appendix III.2.

4.2 Coifman-Donoho Translation-Invariant Denoising

Coifman et al [10, 1, 33] observed that denoising with the conventional wavelet transform and WPD may exhibit visual artifacts, such as pseudo-Gibbs phenomena in the neighborhood of discontinuities. They related these artifacts to the lack of *shift-invariance*, and proposed to reduce them by the following averaging procedure [10]: Applying a range of shifts to the noisy data, subsequently denoising the shifted versions using the wavelet transform, and finally unshifting and averaging the denoised data. This procedure, termed *Cycle-Spinning* [10], often yields a better visual performance on smooth parts of the signal.

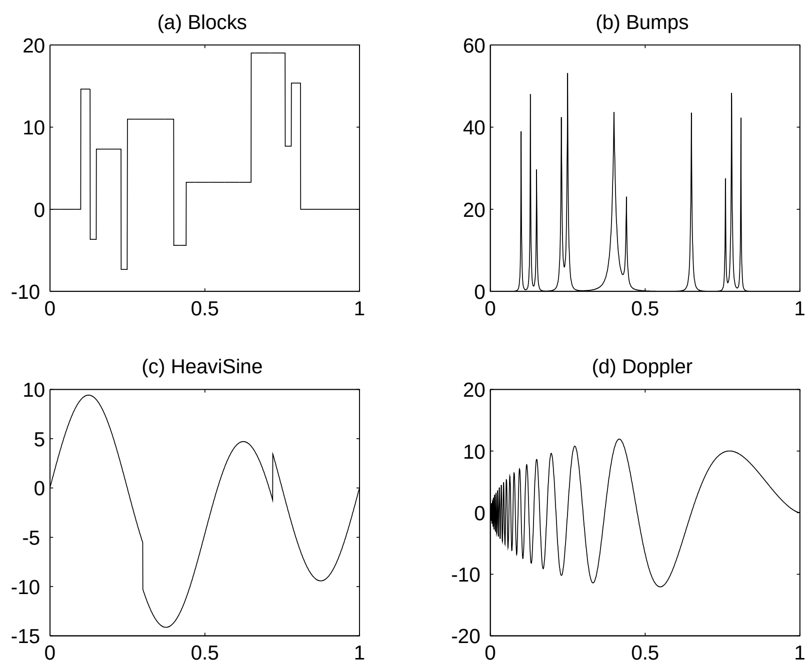


Figure 4.1: Test signals.

Figure 4.1 shows the test signals of length $N = 2048$, chosen to represent various signal classes. Figure 4.2 represents the noisy versions of the test signals, where WGN $\mathbf{e} \sim \mathcal{N}(0,1)$ was added to each signal.

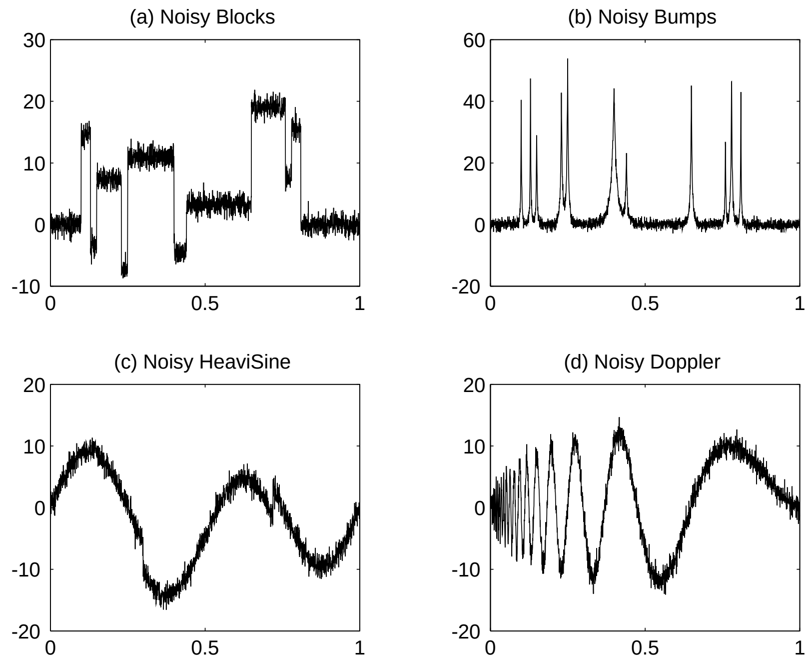


Figure 4.2: Noisy signals.

Figure 4.3 shows signals, denoised by simple wavelet shrinkage. Here the noisy signals were transformed into the wavelet domain with WT based on Daubechies nearly symmetric mother wavelet with 8 vanishing moments, the wavelet coefficients where modified with the VisuShrink estimator, and the modified coefficients were transformed into time domain. As one can see in Fig. 4.3, the enhanced signals contain discontinuities and other rapid time changes.

Figure 4.4 represents results of fully "translation invariant" (TI) denoising. Here the VisuShrink estimator was applied to all N circular shifts of each noisy signal, and the enhanced signals were obtained averaging all the unshifted denoised versions. As one can see, the pseudo-Gibbs oscillations are considerably reduced, at the clear cost of oversmoothing.

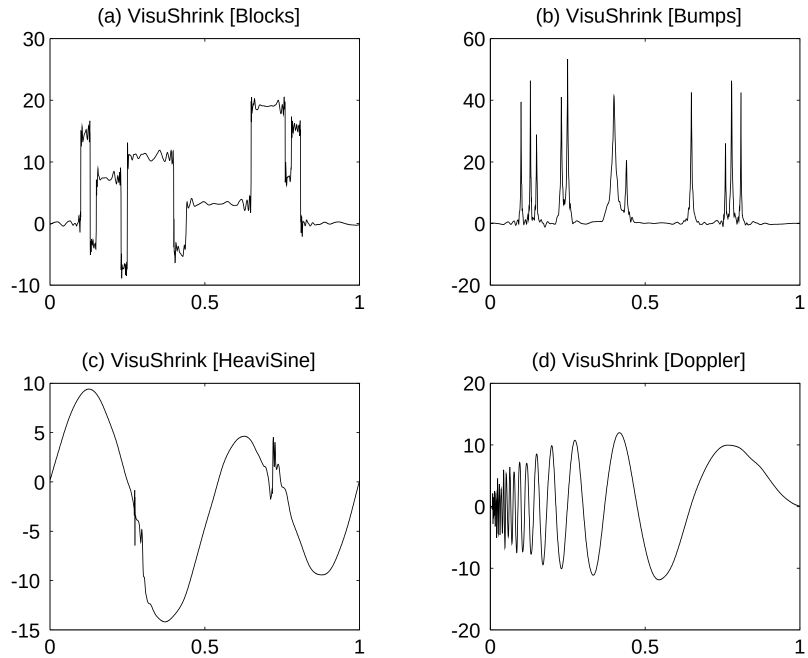


Figure 4.3: Signals, enhanced by VisuShrink estimator.

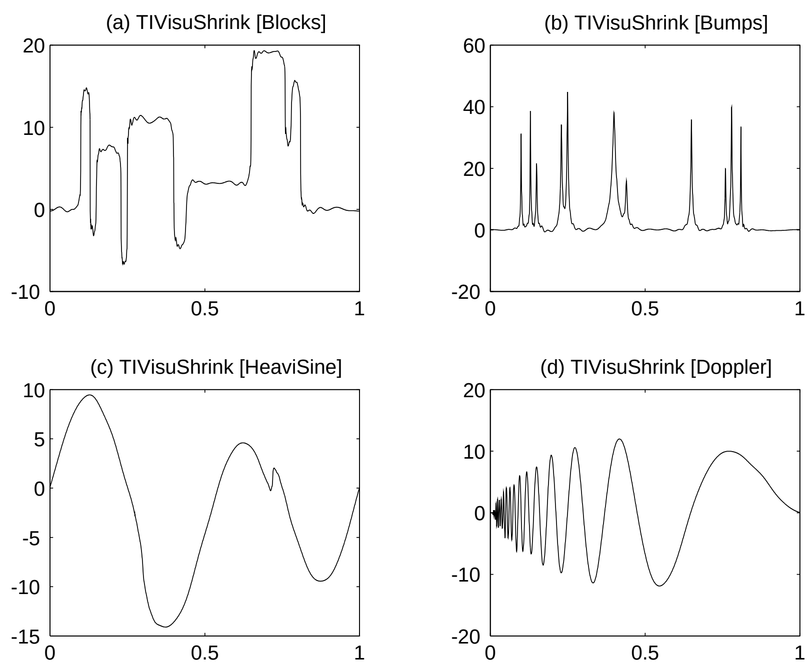


Figure 4.4: Signals, enhanced by Cycle-Spinning for all N circular shifts

4.3 Saito Adaptive Estimator

4.3.1 Problem Formulation

Again we consider the additive white Gaussian noise model of Section 4.1.1. For estimating $\mathbf{f}$ we'll use now a library of orthonormal bases (rather than a specific predetermined basis). Let's denote this library by $\mathcal{B} = \{B_1, B_2, ...B_m, ..., B_M\}$, where each element B_m $(1 \leq m \leq M)$ represents an orthonormal basis in the library and M is the number of bases in $\mathcal{B}$. The hope is that the best basis from such a library will represent the unknown signal $\mathbf{f}$ by a small number k $(< N)$ of elements, associated with B_m, i.e.,

$$\mathbf{f} = W_m^T \Theta_m^{(k)}, \tag{4.22}$$

where $W_m \in \mathrm{I\!R}^{N \times N}$ is the orthogonal transform matrix, $W_m^T \in \mathrm{I\!R}^{N \times N}$ is an orthogonal matrix whose column vectors are the basis elements of B_m, and $\Theta_m^{(k)} \in \mathrm{I\!R}^N$ is the expansion coefficients vector of $\mathbf{f}$ containing only k non-zero coefficients.

Now the problem of simultaneous noise suppression and signal compression can be stated as follows. Given the noisy observations $\mathbf{y}$ and a library of orthonormal bases $\mathcal{B}$ find the "best" k and B_m. In other words, the estimation problem is formulated as a model selection problem, where model is the basis B_m and the number of terms is k.

4.3.2 The Minimum Description Length (MDL) Principle

One of the most suitable criteria for defined problem is the so-called *Minimum Description Length* (MDL) information-theoretic criterion, that was proposed by Rissanen [28]. The MDL principle suggests that the "best" model among the given collection of models is the one giving the shortest description of the data and of the model itself. For each model in the collection, the length of description of the data is counted as the codelength of encoding the data using that model in bits. The length of description of a model is the codelength of specifying that model, e.g., the number of parameters and their values when dealing with a parametric model.

Let $\Omega = \{\Upsilon_i : i = 1, 2, ...\}$ be the collection of models at hand and let Υ denote the true model generating the data $\mathbf{f}$. Since the true model Υ is not known we use one of the models in the set Ω as its estimate: $\widehat{\Upsilon} = \Upsilon_m$. Given the index m, we can write the codelength for the whole process as

$$\mathcal{L}(\mathbf{f}, \Upsilon_m, m) = \mathcal{L}(m) + \mathcal{L}(\Upsilon_m|m) + \mathcal{L}(\mathbf{f}|\Upsilon_m, m), \tag{4.23}$$

where $\mathcal{L}(m)$ is the codelength for encoding the model number (e.g., for WPD the model number m uniquely determines the binary tree), $\mathcal{L}(\Upsilon_m|m)$ is the codelength for encoding parameters of the model and their values, and $\mathcal{L}(\mathbf{f}|\Upsilon_m, m)$ is the codelength for encoding the signal given the model Υ_m. The MDL criterion suggests picking the model Υ_m which results in the minimum of the total description length (4.23). Thus, we should minimize each of the terms in (4.23), and ignoring the integer constraint for the codelength, it leads to the so-called Shannon codelength

$$\mathcal{L}(x) = -\log_2 p(x), \tag{4.24}$$

where x is a symbol from a finite alphabet $\aleph$ and $p(x)$ is a probability mass function of x.

Instead of minimizing the ideal codelength (4.23), Rissanen proposed to minimize

$$MDL(\mathbf{f}, \widehat{\Upsilon}, m) = \mathcal{L}(m) + \sum_{j=1}^{k_m} \mathcal{L}^*([\widehat{v}_{m,j}]) + \frac{k_m}{2}\log_2 N + \mathcal{L}(\mathbf{f}|\widehat{\Upsilon}, m) \tag{4.25}$$

(see Appendix III.3.1). Here, $\{\widehat{v}_{m,j}\}_{j=1}^{k_m}$ are the k_m real-valued parameters that describe the maximum likelihood (ML) estimate $\widehat{\Upsilon}$ of the true model Υ, $[x]$ denotes an integer part of x, and $\mathcal{L}^*(\cdot)$ is the codelength defined by Rissanen [28] (Eq. (III.21)). The minimization of (4.25) provides the best compromise between low complexity in the model and high likelihood of the data.

4.3.3 Simultaneous Noise Suppression and Signal Compression

To invoke the MDL formalism, given (k, m) in (4.22), Saito [33] prepares a conceptual encoder which expands the data $\mathbf{y}$ on the basis B_m. He then transmits the number (k) of non-zero terms, the specification of the basis B_m, the k expansion coefficients $\Theta_m^{(k)}$, the variance (σ^2) of the WGN model, and the estimation error. The total codelength to be minimized may be expressed as the sum of the codelength of: (1) two natural numbers (k, m), (2) $(k+1)$ real-valued parameters $(\Theta_m^{(k)}, \sigma^2)$ given (k, m), and (3) the deviations of the observed data $\mathbf{y}$ from the (estimated) signal $\widehat{\mathbf{f}} = W_m^T \widehat{\Theta}_m^{(k)}$ given $(k, m, \Theta_m^{(k)}, \sigma^2)$. The appropriate total description length (4.25) now becomes

$$MDL(\mathbf{y}, \widehat{\Theta}_m^{(k)}, \widehat{\sigma}^2, k, m) = \mathcal{L}(k, m) + \mathcal{L}(\widehat{\Theta}_m^{(k)}, \widehat{\sigma}^2 | k, m) + \mathcal{L}(\mathbf{y} | \widehat{\Theta}_m^{(k)}, \widehat{\sigma}^2, k, m), \qquad (4.26)$$

where $\widehat{\Theta}_m^{(k)}$ and $\widehat{\sigma}^2$ are the ML estimates of $\Theta_m^{(k)}$ and σ^2, respectively. In order to suppress noise and simultaneously compress the signal, Saito defines the approximate MDL:

$$AMDL = \frac{3}{2} k \log_2 N + \frac{N}{2} \log_2 ||W_m \mathbf{y} - \eta^{(k)}(W_m \mathbf{y})||_{2,N}^2, \qquad (4.27)$$

(see Appendix III.3.2). Here, $\eta^{(k)}$ is a thresholding operation which keeps the k largest (in absolute value) elements intact and sets all other elements to zero. Minimizing AMDL we find the optimal value of k.

Unfortunately, the AMDL principle, as proposed by Saito, is not additive. However, the search for the index m is equivalent to the search for best basis, and it can be done by the minimization of some additive cost function. Thus, Saito employed the Shannon entropy as the primary cost function for the determination of the best basis B_m, and the AMDL principle as a secondary criterion.

Using (4.27) we can summarize the procedure that deals with the simultaneous noise suppression and signal compression. Saito's algorithm comprises the following steps:

Step 1 *Expand the data* $\mathbf{y}$ *into the library* $\mathcal{B} = \{B_m\}_{m=1}^{M}$*, i.e., obtain the expansion coefficients* $\mathbf{w}_m = \{w_{m,j}\}_{j=0}^{N-1}$ *for* $1 \leq m \leq M$*.*

Step 2 *Determine the optimal basis* $A \equiv B_m = \{\phi_{m,j}\}_{j=0}^{N-1}$*, i.e., the basis with the minimal value of some cost function.*

Step 3 *Pick*

$$\hat{k} = \underset{0 \leq k \leq N-1}{\arg\min}\,(AMDL). \tag{4.28}$$

Step 4 *Threshold the expansion coefficients:*

$$\widehat{\Theta}_m^{(\hat{k})} = \left\{\widehat{\theta}_{m,j}^{(\hat{k})}\right\}_{j=0}^{N-1} = \eta^{(\hat{k})}\mathbf{w}_m. \tag{4.29}$$

Step 5 *Reconstruct the signal estimate according to the optimal basis* A*:*

$$\widehat{\mathbf{f}} = \sum_{j=0}^{N-1} \widehat{\theta}_{m,j}^{(\hat{k})}\phi_{m,j}. \tag{4.30}$$

The library $\mathcal{B}$ of orthonormal bases may include any collection of orthonormal bases that can be organized as a binary tree. The Wavelet Packet and Local Trigonometric libraries are well known examples.

4.4 Cohen-Raz-Malah Shift-Invariant Denoising

4.4.1 MDL-Based Additive Information Cost Function

The MDL principle, as proposed by Saito, non-additive. Therefore, in order to use it for selecting the best basis, Cohen, Raz and Malah [5] proposed a modified additive MDL cost function.

Let's consider the model of an additive white Gaussian noise presented in Section 4.1.1. Also, denote by E_m the tree-set that corresponds to the terminal nodes of a SIWPD tree and describes the orthonormal basis $B_m \in \mathcal{B}$. k, as before, is the number of non-zero coefficients in the set Θ_m of expansion coefficients of the unknown signal $\mathbf{f}$ presented on the basis B_m. $\{j_{m,n}\}_{n=0}^{k-1}$ denote the set of position indexes of the non-zero coefficients in Θ_m. The encoding of the noisy observations $\mathbf{y}$, and hence the computation of the codelength, is carried out in three steps:

1) encoding the observed data assuming E_m, k and $\{j_{m,n}\}_{n=0}^{k-1}$ are given,

2) encoding the number of non-zero signal terms k and their locations $\{j_{m,n}\}_{n=0}^{k-1}$ assuming that E_m is given,

3) encoding the tree-set E_m.

Accordingly, the total description length of the data is given by

$$\mathcal{L}(\mathbf{y}) = \mathcal{L}\left(\mathbf{y} \mid E_m, k, \{j_{m,n}\}_{n=0}^{k-1}\right) + \mathcal{L}\left(k, \{j_{m,n}\}_{n=0}^{k-1} \mid E_m\right) + \mathcal{L}(E_m)\,. \tag{4.31}$$

In their work Cohen et. al. showed that the optimal number of signal terms $\hat{k}$ and their optimal locations $\{\hat{j}_{m,n}\}_{n=0}^{\hat{k}-1}$ are given by

$$\hat{k} = \#\left\{w_{m,n}^2 > 3\sigma^2 \ln N \mid 0 \le n \le N-1\right\} \tag{4.32}$$

and

$$\left\{\hat{\jmath}_{m,n}\right\}_{n=0}^{\hat{k}-1} = \left\{n \mid w_{m,n}^2 > 3\sigma^2 \ln N, 0 \le n \le N-1\right\} \tag{4.33}$$

(further details are given in Appendix III.4.1). Specifically, given E_m we compute the expansion coefficients of the observed data, and then subsequently identify $\hat{k}$ as the number of coefficients exceeding (in absolute value) the threshold $\sigma\sqrt{3 \ln N}$, and $\{\hat{\jmath}_{m,n}\}_{n=0}^{\hat{k}-1}$ as their locations. Thus, the sum of the first two terms in Eq. (4.31) is given by

$$\mathcal{L}(\mathbf{y} \mid E_m) = \frac{1}{2\sigma^2 \ln 2} \sum_{n=0}^{N-1} \min\left(w_{m,n}^2, 3\sigma^2 \ln N\right). \tag{4.34}$$

For $|E_m| \gg 1$, the codelength $\mathcal{L}(E_m)$ can be approximated by

$$\mathcal{L}(E_m) \approx 3|E_m|, \tag{4.35}$$

where the constant terms are ignored. Adding the codelength $\mathcal{L}(\mathbf{y} \mid E_m)$ (Eq. (4.34)) to Eq. (4.35), the total description length of the observed data is given by

$$\mathcal{L}(\mathbf{y}) = \mathcal{L}(E_m) + \mathcal{L}(\mathbf{y} \mid E_m) = 3\,|E_m| + \frac{1}{2\sigma^2 \ln 2} \sum_{n=0}^{N-1} \min\left(w_{m,n}^2\,,\ 3\sigma^2 \ln N\right). \tag{4.36}$$

The dependence of $\mathcal{L}(\mathbf{y})$ on the tree-set E_m is introduced here through the number of terminal nodes and the values of the expansion coefficients $\{w_{m,n}\}_{n=0}^{N-1}$. Since the total energy of the coefficients $\sum_{n=0}^{N-1} w_{m,n}^2 = \|\mathbf{y}\|_{2,N}^2$ is independent of E_m, we want that the relative energy, contained in the coefficients exceeding $\sigma\sqrt{3 \ln N}$ in magnitude, will be as large as possible. At the same time, we want to minimize the complexity of the expansion tree (the number of terminal nodes). Thus, search algorithm for the best tree-set E_m that minimizes $\mathcal{L}(\mathbf{y})$ is needed.

4.4.2 The Optimal Tree Design and Signal Estimation

Let $\mathcal{B}$ represent the SWP library of orthonormal bases. Since each basis B_m in the library is associated with a tree-set E_m, the search for the optimal E_m is equivalent to the search for the optimal basis in $\mathcal{B}$.

Let's denote by $\mathbf{w}_m = \{w_{m,n,j}\}$ the set of the expansion coefficients of the observed data $\mathbf{y}$ on the basis B_m, $n \in E_m$ is the index of the appropriate terminal node in the tree-set E_m, $0 \leq j \leq N_n - 1$ and N_n is the number of expansion coefficients that belong to the n-th terminal node. Then by Eq. (4.36)

$$\mathcal{L}(\mathbf{w}_m) = \sum_{n \in E_m} \mathcal{L}(\{w_{m,n,j}\}_{j=0}^{N_n-1}), \tag{4.37}$$

where

$$\mathcal{L}(\{w_{m,n,j}\}_{j=0}^{N_n-1}) = 3 + \frac{1}{2\sigma^2 \ln 2} \sum_{j=0}^{N_n-1} \min\left\{w_{m,n,j}^2\,,\ 3\sigma^2 \ln N\right\} \tag{4.38}$$

is the codelength for the terminal node $n \in E_m$. Thus, the optimal basis for $\mathbf{y}$ in $\mathcal{B}$ with respect to the MDL principle is $B_m \in \mathcal{B}$ for which $\mathcal{L}(B_m\mathbf{y})$ is minimal.

The codelength in Eq. (4.37) constitutes an additive cost function, resulting directly from the approximations derived in the previous section. Accordingly, we can apply the SIWPD (Section 3.2.2) to the observed data $\mathbf{y}$ and to use the Eq. (4.36) to find the optimal basis $A \equiv B_m = \{\phi_{m,j}\}_{j=0}^{N-1}$, that minimizes the description length of the observed data, as described in Section 3.2.3.

From Eqs. (III.40), (4.32) and (4.33), the optimal estimate $\widehat{\mathbf{f}}$ of $\mathbf{f}$ is obtained by expanding the observed data $\mathbf{y}$ on the optimal basis A and *hard-thresholding* the expansion coefficients by $\tau \equiv \sigma\sqrt{3 \ln N}$. Specifically,

$$\widehat{\mathbf{f}} = \sum_{j=0}^{N-1} \eta_h^\tau(w_{m,j})\phi_{m,j}, \tag{4.39}$$

where $w_{m,j} = \langle \mathbf{y}, \phi_{m,j} \rangle$, and $\eta_h^\tau(x)$ is the hard-thresholding operator.

The following steps summarize the optimal shift-invariant signal estimation by the MDL principle:

Step 1 *Expand the data* $\mathbf{y}$ *into the library* $\mathcal{B} = \{B_m\}_{m=1}^{M}$*, i.e., obtain the expansion coefficients* $\mathbf{w}_m = \{w_{m,j}\}_{j=0}^{N-1}$ *for* $1 \leq m \leq M$*.*

Step 2 *Determine the optimal basis* $A \equiv B_m = \{\phi_{m,j}\}_{j=0}^{N-1}$*, i.e., the basis that minimizes the MDL cost function defined by Eq. (4.36).*

Step 3 *Hard-threshold the expansion coefficients* $\mathbf{w}_m$ *by* $\tau = \sigma\sqrt{3 \ln N}$.

Step 4 *Reconstruct the signal estimate using Eq. (4.39).*

The computational complexity of executing an optimal SIWPD best-basis expansion is $O(N2^{L+1})$. Yet, as demonstrated in Section 3.2.3, one may resort to a *sub-optimal* SIWPD procedure entailing a reduced complexity at the expense of a longer description length, while still retaining the desirable shift-invariance property. The larger d and L, the larger the complexity, but the shorter the description length. This algorithm can be readily extended to searches over more than a single library.

Chapter 5 : Speech Denoising Algorithms

5.1 Introduction

In this section we develop WPD-based and LTD-based speech denoising algorithms, and study the consequences of shift-invariance on speech enhancement and the resulting artifacts.

5.2 Implementation and Quality Measures

All the software for this thesis was written for the $Matlab^{©}$ environment and is based on $Matlab^{©}$ and $WavBox^{©}$ software. All examinations were done for the 3 following sentences, each pronounced by a male and a female:

1) *A lathe is a big tool*

2) *An icy wind raked the beach*

3) *Joe brought a young girl.*

Each sentence is sampled at 8 KHz sampling frequency and has 16384 samples ($J = 14$).

Quality of the resulting speech signals was evaluated by listening and according to the following quantitative measures:

1) *Signal to Noise Ratio* (SNR):

$$SNR = 10\log\left(\frac{||\mathbf{f}||_2^2}{||\mathbf{f}-\widehat{\mathbf{f}}||_2^2}\right) \quad [\mathrm{dB}] \tag{5.1}$$

where $\widehat{\mathbf{f}}$ is an estimate of clean speech $\mathbf{f}$.

2) *Segmental SNR* (SEGSNR):

$$SEGSNR = \frac{1}{M}\sum_{i=1}^{M} SNR_i \quad [\mathrm{dB}] \tag{5.2}$$

where

$$SNR_i = 10\log\left(\frac{||\mathbf{f}_i||_2^2}{||\mathbf{f}_i-\widehat{\mathbf{f}}_i||_2^2}+1\right), \tag{5.3}$$

$\mathbf{f}_i$ and $\widehat{\mathbf{f}}_i$ are the i-th frames of clean speech and estimated speech signals respectively, M is the number of frames.

3) *Log-Spectral Distance* (LSD):

$$LSD = \frac{1}{M}\sum_{i=1}^{M} D_i \quad [\mathrm{dB}] \tag{5.4}$$

where

$$D_i = \left[\frac{1}{N}\sum_{k=1}^{N}\left(10\log|F_i(k)| - 10\log\left|\widehat{F}_i(k)\right|\right)^2\right]^{1/2}, \tag{5.5}$$

$$F_i(k) = DFT\left\{\mathbf{f}_i\right\}(k), \quad \widehat{F}_i(k) = DFT\left\{\widehat{\mathbf{f}}_i\right\}(k), \tag{5.6}$$

N is the number of samples in the frame.

5.3 WPD-Based Speech Denoising

5.3.1 Introduction

In this section we compare WPD-based speech denoising algorithms.

Obviously, a WP library allows choice of a mother wavelet with appropriate properties like linearity of phase, number of vanishing moments, time and frequency localization, and the WPD can be performed with different numbers of decomposition levels and different cost functions. Moreover, different gain functions can be applied to the WPD coefficients.

5.3.2 Wiener Filter

Let's consider the same model of additive white Gaussian noise (Section 4.1.1). When the speech and the noise signals are independent

$$E||\mathbf{w}||_{2,N}^2 = E||\Theta||_{2,N}^2 + E||\mathbf{z}||_{2,N}^2 = ||\Theta||_{2,N}^2 + N\sigma^2. \tag{5.7}$$

Among all linear estimators $\widehat{\Theta} = G(\mathbf{w}, \sigma) \cdot \mathbf{w}$, the *non-causal Wiener filter* obtains an optimal (in *mean squared error* sense) estimate of the clean signal:

$$G_w^*(\Theta, \sigma) = \frac{||\Theta||_{2,N}^2}{||\Theta||_{2,N}^2 + E||\mathbf{z}||_{2,N}^2}. \tag{5.8}$$

Since $||\Theta||_{2,N}^2$ is unknown, we use, on practice, it's estimate $\eta_s(||\mathbf{w}||_{2,N}^2, E||\mathbf{z}||_{2,N}^2)$, so

$$G_w(\mathbf{w}, \sigma) = \frac{\eta_s(||\mathbf{w}||_{2,N}^2, E||\mathbf{z}||_{2,N}^2)}{\eta_s(||\mathbf{w}||_{2,N}^2, E||\mathbf{z}||_{2,N}^2) + E||\mathbf{z}||_{2,N}^2}, \tag{5.9}$$

where η_s is soft-thresholding operator which was defined in Section 4.1.2. $G_w^*(\Theta, \sigma)$ is the *ideal gain*, and $G_w(\mathbf{w}, \sigma)$ is the gain that we practically use. Given $\mathbf{y}$, the variance σ^2 of the additive WGN still has to be estimated.

#	Speaker	Decomposition type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	WT	10	6.06	11.51	14.08	8.14	9.11
1	Female	WPD	10	6.06	11.51	**14.47**	8.37	8.86
1	Male	WT	10	5.96	11.53	12.67	7.06	10.39
1	Male	WPD	10	5.96	11.53	**13.3**	7.34	9.93
2	Female	WT	10	6.68	9.47	12.87	9.34	6.91
2	Female	WPD	10	6.68	9.47	**13.33**	8.61	7.35
2	Male	WT	10	6.73	9	11.97	7.73	7.8
2	Male	WPD	10	6.73	9	**12.69**	8.2	6.96
3	Female	WT	10	6.17	11.11	13.34	7.69	9.57
3	Female	WPD	10	6.17	11.11	**13.74**	7.89	9.43
3	Male	WT	10	5.92	11.51	12.83	7.06	10.28
3	Male	WPD	10	5.92	11.51	**13.5**	7.39	10.01

Table 5.1: Influence of decomposition type on WPD-based denoising performance. # is the number of the test sentence. SNRs and LSD are in [dB]. The input SNR, SEGSNR and LSD are the original SNRs and LSD, and output SNR, SEGSNR and LSD are the resulting SNRs and LSD.

5.3.3 Decomposition Type

Here we compare the performance of WT and WPD-based denoising. The comparison was performed for noisy speech at 10dB SNR. The noise added to the clean speech was WGN. The results of this comparison, using the Wiener gain function (Eq. (5.9)) and a decomposition of whole signal (i.e., without segmentation into frames), $L = 5$ and the Daubechies nearly symmetric mother wavelet (DNS) of 8 order (support width is 15 taps) are presented in Table 5.1. For WPD an entropy-based best-basis selection algorithm was implemented.

As we can see, for all of the above examples the WPD-based denoising performs better than WT-based one. Additional simulations have shown that this conclusion does not depend on the estimator type used for denoising, and for any of the described estimators WPD-based denoising sounds less noisy. The reason for the advantage of WPD is that this decomposition type allows better adaptation to the given signals properties.

#	Speaker	Estimator type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
2	Female	*VisuShrink*	10	6.68	9.47	10.08	6.76	6.88
2	Female	*RiskShrink*	10	6.68	9.47	12.69	8.13	6.47
2	Female	*SureShrink*	10	6.68	9.47	**14.73**	9.28	6.35
2	Female	*Wiener*	10	6.68	9.47	13.35	8.63	7.34
2	Female	*Saito*	10	6.68	9.47	9.55	6.52	7.44
2	Female	*Cohen*	10	6.68	9.47	11.28	7.62	6.85
2	Male	*VisuShrink*	10	6.73	9	9.02	6.29	8.68
2	Male	*RiskShrink*	10	6.73	9	11.74	7.68	6.42
2	Male	*SureShrink*	10	6.73	9	**14.19**	8.96	6.44
2	Male	*Wiener*	10	6.73	9	12.72	8.22	6.96
2	Male	*Saito*	10	6.73	9	7.37	5.45	8.82
2	Male	*Cohen*	10	6.73	9	10.25	6.89	7.29

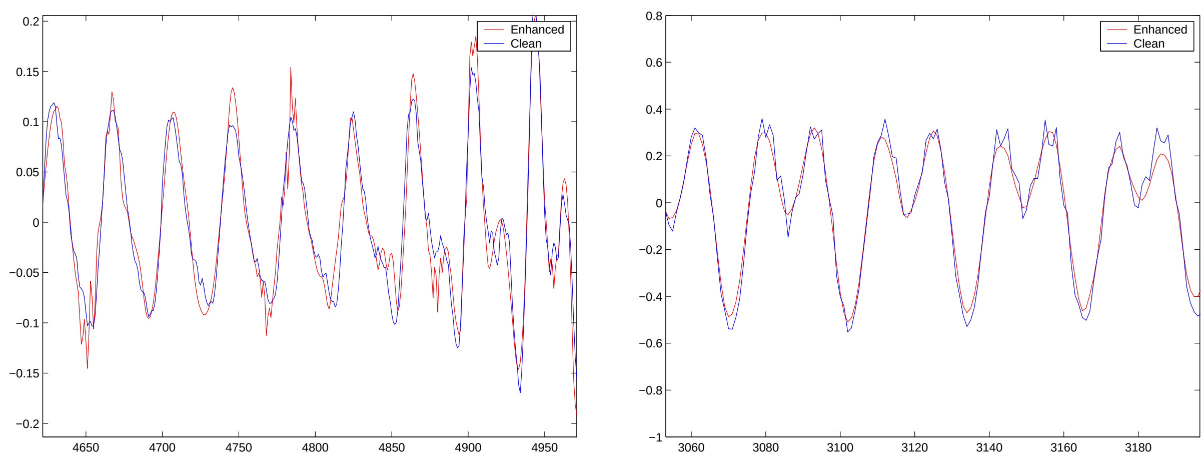
Table 5.2: Influence of estimator type on WPD-based denoising performance.

5.3.4 Estimator Type

As we saw in Section 3.3.4, different estimators can be used in wavelet-based denoising. In this section we compare different approaches for the estimation of a speech signal from it's noisy observation.

Tests (Table 5.2) were done under the same conditions as in Section 5.3.3, using WPD-based denoising with $L = 6$. The estimators *RiskShrink*, *VisuShrink* and *SureShrink* were used with soft thresholding.

SureShrink and Wiener estimators are clearly better than all the other examined estimators. But enhanced signals obtained by using the SureShrink estimator suffer from artifacts in the form of spike-like transitions. The reason is that the thresholding operation leads to discontinuities: there can be situations where in the neighborhood of high-amplitude coefficients only one expansion coefficient will be set to zero, or vice versa (only few high-amplitude coefficients will not be set to zero). For example, use of the entropy for the best-basis selection usually leads to a decomposition tree in which the high frequency bands are not decomposed (Figure 5.3), thus basis functions that correspond to these bands have short time support (good time localization properties) and are very oscillatory. Leaving only few coefficients in such a band will lead to strong artifacts (Figure 5.1(a)). The same artifacts characterize all soft and hard thresholding-based estimators. Moreover, in the high frequency regions speech is characterized by relatively low energy, thus speech, enhanced by the thresholding-based algorithms, sounds oversmoothed - the estimate of high frequency speech coefficients is zero for the most of the coefficients, and it seems like the speech was low-pass filtered (Figures 5.1(b), 5.2).

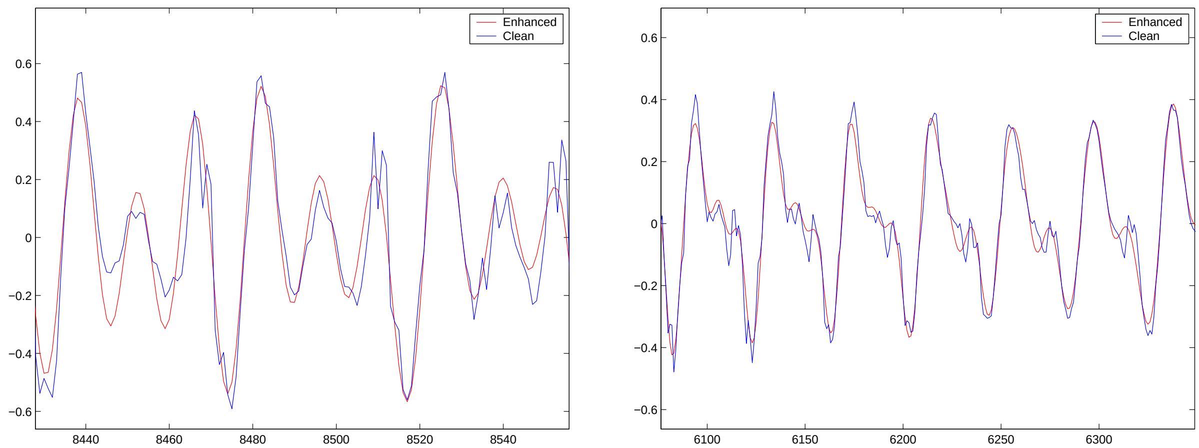


(a) (b)

Figure 5.1: Fragments of speech signals, enhanced by thresholding-based algorithms: (a) Artifacts in speech enhanced by using the *SureShrink* estimator, (b) Oversmoothing in speech enhanced by using the *RiskShrink* estimator.

Despite the fact that the speech enhanced by the Wiener estimator sounds noisier, it's quality is much better than for all other tested estimators. In the sequel we'll show that the use of the *decision directed* approach for *a priori* SNR estimation can suppress the background noise without introducing artifacts.

In order to improve the speech quality obtained by thresholding-based estimators, we should use an appropriate cost function and choose a mother wavelet with a wider time support (worse time localization). However, in the following subsections we'll show that



(a) (b)

Figure 5.2: Oversmoothing in speech signals, enhanced by thresholding-based algorithms: fragments of speech signals, enhanced by (a) *Saito* estimator, (b) *Cohen* et. al. estimator.

these measures do not suppress sufficiently the artifacts, which characterize thresholding-based estimators.

5.3.5 Cost Function and Lowest Decomposition Level

In order to check if any of the additive cost functions that were mentioned in Section 3.1.3 is better suited for speech denoising, simulations were carried out as presented in Table 5.3. The conditions are the same as in Section 5.3.3, the Wiener estimator and a WPD with $L = 6$ were used.

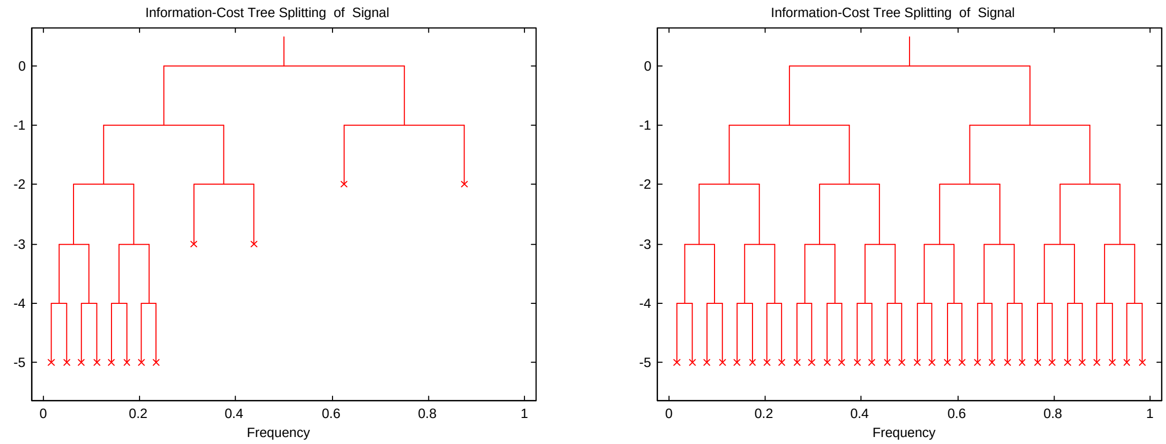
We see that the differences are not significant, but full subband decomposition results it the best denoising results. Further decomposition ($L > 6$) doesn't lead to serious improvement for any of above cost functions. In conclusion, none of the mentioned above additive cost functions is globally optimal.

The full subband decomposition is an attractive choice: it can be represented by a fixed tree structure and can be easily implemented. Moreover, it allows simple utilization of the decision directed a priori SNR estimation: In order to utilize the latter, we have

#	Speaker	Cost function	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	H	10	6.06	11.51	14.54	8.41	8.84
1	Female	$\mathcal{E}$	10	6.06	11.51	14.63	8.43	8.82
1	Female	ℓ^1	10	6.06	11.51	14.63	8.43	8.77
1	Female	Full Subband	10	6.06	11.51	**14.68**	8.45	8.83
1	Male	H	10	5.96	11.53	13.39	7.37	9.9
1	Male	$\mathcal{E}$	10	5.96	11.53	13.53	7.46	9.86
1	Male	ℓ^1	10	5.96	11.53	13.4	7.37	9.93
1	Male	Full Subband	10	5.96	11.53	**13.55**	7.47	9.88
2	Female	H	10	6.68	9.47	13.35	8.63	7.34
2	Female	$\mathcal{E}$	10	6.68	9.47	13.53	8.69	7.2
2	Female	ℓ^1	10	6.68	9.47	13.49	8.69	7.21
2	Female	Full Subband	10	6.68	9.47	**13.55**	8.69	7.2
2	Male	H	10	6.73	9	12.72	8.22	6.96
2	Male	$\mathcal{E}$	10	6.73	9	12.77	8.25	6.88
2	Male	ℓ^1	10	6.73	9	12.72	8.24	6.91
2	Male	Full Subband	10	6.73	9	**12.79**	8.26	6.89
3	Female	H	10	6.17	11.11	13.77	7.9	9.42
3	Female	$\mathcal{E}$	10	6.17	11.11	13.86	7.94	9.37
3	Female	ℓ^1	10	6.17	11.11	13.86	7.94	9.38
3	Female	Full Subband	10	6.17	11.11	**13.86**	7.93	9.39
3	Male	H	10	5.92	11.51	13.54	7.38	10.02
3	Male	$\mathcal{E}$	10	5.92	11.51	13.5	7.37	10.03
3	Male	ℓ^1	10	5.92	11.51	13.56	7.39	10
3	Male	Full Subband	10	5.92	11.51	**13.56**	7.39	10.02

Table 5.3: Influence of cost function on WPD-based denoising performance. H corresponds to the Shannon entropy, $\mathcal{E}$ to the log energy, and ℓ^1 to the concentration in ℓ^1 norm.

to track the a priori SNR for terminal tree nodes. For a fixed tree structure (like the full subband decomposition) the indices of terminal nodes do not change from frame to frame. Where as to utilize the decision directed a priori SNR estimation for a WPD (with an additive cost function) we have to track the a priori SNR at <u>all tree nodes</u> because the indices of terminal nodes may change from frame to frame.



(a) (b)

Figure 5.3: Examples of WPD trees: (a) Result of entropy-based best-basis selection algorithm, (b) Full subband decomposition tree.

5.3.6 Mother Wavelet: Phase Linearity and Design

It's well known that the linearity of phase is of high importance in some fields of Signal Processing. Thus, it's important to know should we use wavelet bases that possess phase linearity (biorthogonal wavelet bases), nearly symmetric wavelets or minimum phase wavelets. Moreover, as was mentioned in Section 5.3.4, in order to reduce the artifacts for thresholding-based denoising algorithms, we have to verify the influence of time-resolution properties of mother wavelet on denoising performance.

Tests (Table. 5.4) were done under the same conditions as in Section 5.3.3 using WPD with entropy-based best-basis selection, for Daubechies minimum phase (DMP) mother wavelet (8'th order), Daubechies nearly symmetric (DNS) mother wavelet (8'th order) and biorthogonal (BIOR) mother wavelets (5'th order for both decomposition and reconstruction mother wavelets). In order to use biorthogonal mother wavelets, in all the tests noise variance was estimated for each decomposition band.

According to the listening and the results presented in Table 5.4 we can see that it's not important if the mother wavelet possess linearity of phase or not when we decompose

#	Speaker	Mother wavelet	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	DMP	10	6.06	11.51	14.43	8.38	9.05
1	Female	DNS	10	6.06	11.51	**14.54**	8.41	8.84
1	Female	BIOR	10	6.06	11.51	14.16	8.19	9.15
1	Male	DMP	10	5.96	11.53	**13.4**	7.38	9.92
1	Male	DNS	10	5.96	11.53	13.39	7.37	9.9
1	Male	BIOR	10	5.96	11.53	12.58	6.96	10.47
2	Female	DMP	10	6.68	9.47	**13.48**	8.69	7.25
2	Female	DNS	10	6.68	9.47	13.35	8.63	7.34
2	Female	BIOR	10	6.68	9.47	13.04	8.44	7.67
2	Male	DMP	10	6.73	9	12.71	8.22	6.94
2	Male	DNS	10	6.73	9	**12.72**	8.22	6.96
2	Male	BIOR	10	6.73	9	11.69	7.57	7.88
3	Female	DMP	10	6.17	11.11	**13.83**	7.92	9.35
3	Female	DNS	10	6.17	11.11	13.77	7.9	9.42
3	Female	BIOR	10	6.17	11.11	13.19	7.59	9.54
3	Male	DMP	10	5.92	11.51	13.51	7.35	10.04
3	Male	DNS	10	5.92	11.51	**13.54**	7.38	10.02
3	Male	BIOR	10	5.92	11.51	12.78	6.98	10.3

Table 5.4: Influence of mother wavelet type on WPD-based denoising performance. $L = 6$.

the whole signal. Moreover, there isn't any advantage in using biorthogonal wavelets even when the analysis is done with framing.

In order to improve the performance of the thresholding-based denoising algorithms, we used a *generalized Meyer* mother wavelet. The Meyer mother wavelet [11] is defined by its QMF $m_0(\omega)$:

$$m_0(\omega) = \begin{cases} 1, & |\omega| \leq \frac{\pi}{3}, \\ \cos\left[\frac{\pi}{2} \cdot \nu(\frac{3}{\pi}|\omega| - 1)\right], & \frac{\pi}{3} \leq |\omega| \leq \frac{2\pi}{3}, \\ 0, & |\omega| \geq \frac{2\pi}{3}, \end{cases} \tag{5.10}$$

where $\nu(x)$ is the so-called *auxiliary function*, $x \in [0,1]$. The default choice of $Matlab^{©}$ is $\nu(x) = 35x^4 - 84x^5 + 70x^6 - 20x^7$.

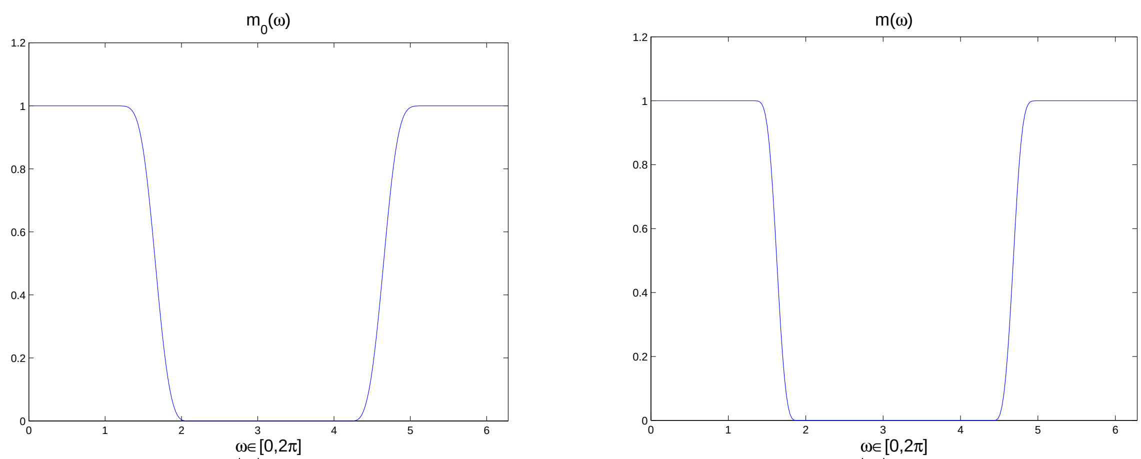
The *generalized Meyer* mother wavelet is given by a *modified* QMF $m(\omega)$:

$$m(\omega) = \begin{cases} 1, & |\omega| \le \frac{\pi}{2}(1-r), \\ \cos\left[\frac{\pi}{2} \cdot \nu\left(\frac{|\omega| - \frac{\pi}{2}(1-r)}{\pi r}\right)\right], & \frac{\pi}{2}(1-r) \le |\omega| \le \frac{\pi}{2}(1+r), \\ 0, & |\omega| \ge \frac{\pi}{2}(1+r), \end{cases} \tag{5.11}$$

where r is the *roll-off*. The choice of $r = 1/3$ corresponds to the standard Meyer mother wavelet:

$$m_0(\omega) = m(\omega)|_{r=1/3}. \tag{5.12}$$

The graphs of $m_0(\omega)$ and $m(\omega)|_{r=1/5}$ are given in Figure (5.4).



(a) (b)

Figure 5.4: Design of Meyer mother wavelet: (a) Roll-off $r = 1/3$ yields standard Meyer mother wavelet, (b) Modified quadrature mirror filter $m(\omega)$ ($r = 1/5$).

Given $m(\omega)$ we can compute its discrete version (approximation):

$$H(k) = m(w)|_{w=\frac{2\pi}{N}k}, \quad k = 0, 1, ...N-1. \tag{5.13}$$

Taking the real part of the IDFT of the sequence $\{H(k)\}_{k=0}^{N-1}$ gives an approximation of modified Meyer QMF $h(n)$:

$$h(n) = Re\{IDFT(\{H(k)\}_{k=0}^{N-1})(n)\}, \quad n = 0, 1, ...N-1. \tag{5.14}$$

#	Speaker	Estimator type,N,r	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
2	Female	$VisuShrink$,32,$\frac{1}{3}$	10	6.68	9.47	10.62	7.04	6.8
2	Female	$RiskShrink$,32,$\frac{1}{3}$	10	6.68	9.47	13.22	8.44	6.54
2	Female	$SureShrink$,32,$\frac{1}{3}$	10	6.68	9.47	**14.81**	9.45	6.29
2	Female	$Wiener$,32,$\frac{1}{3}$	10	6.68	9.47	13.53	8.71	7.26
2	Female	$Saito$,32,$\frac{1}{3}$	10	6.68	9.47	10.01	6.88	6.6
2	Female	$Cohen$,32,$\frac{1}{3}$	10	6.68	9.47	11.87	7.82	6.5
2	Female	$VisuShrink$,64,$\frac{1}{3}$	10	6.68	9.47	10.69	7.08	6.74
2	Female	$RiskShrink$,64,$\frac{1}{3}$	10	6.68	9.47	13.28	8.57	6.47
2	Female	$SureShrink$,64,$\frac{1}{3}$	10	6.68	9.47	**14.83**	9.46	6.19
2	Female	$Wiener$,64,$\frac{1}{3}$	10	6.68	9.47	13.6	8.76	7.24
2	Female	$Saito$,64,$\frac{1}{3}$	10	6.68	9.47	10.19	6.99	6.54
2	Female	$Cohen$,64,$\frac{1}{3}$	10	6.68	9.47	11.88	7.94	6.45
2	Female	$VisuShrink$,64,$\frac{1}{5}$	10	6.68	9.47	10.79	7.15	6.7
2	Female	$RiskShrink$,64,$\frac{1}{5}$	10	6.68	9.47	13.33	8.6	6.43
2	Female	$SureShrink$,64,$\frac{1}{5}$	10	6.68	9.47	**14.92**	9.47	6.12
2	Female	$Wiener$,64,$\frac{1}{5}$	10	6.68	9.47	13.63	8.77	7.2
2	Female	$Saito$,64,$\frac{1}{5}$	10	6.68	9.47	10.22	7.09	6.5
2	Female	$Cohen$,64,$\frac{1}{5}$	10	6.68	9.47	11.9	7.97	6.44

Table 5.5: Use of generalized Meyer mother wavelet: influence of time and frequency localization on WPD-based denoising performance.

The filter $g(n)$ can be easily computed via Eq. (3.8). According to uncertainty principle, the smaller the roll-off, the bigger the order N, that is needed for good approximation (small reconstruction error).

In Table 5.5 results of WPD-based denoising for different roll-offs r and orders N are presented.

Speech, enhanced by thresholding-based algorithms, sounds better for Meyer mother wavelet ($N = 32, 64$) then for other mother wavelets tested before. Reducing time localization we increase the time-support of bases functions thus suppressing the artifacts. Moreover, when using mother wavelets with better frequency localization (smaller roll-off), the quality of enhanced speech improves, resulting in better SNRs and LSD.

In spite of improvement of speech quality for thresholding-based algorithms, the qua-

#	Speaker	Library type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	WPD	10	6.06	11.51	16.01	9.12	8.17
1	Female	SIWPD	10	6.06	11.51	**16.16**	9.16	7.88
1	Male	WPD	10	5.96	11.53	**15.03**	8	9.65
1	Male	SIWPD	10	5.96	11.53	14.68	7.86	9.52
2	Female	WPD	10	6.68	9.47	**14.92**	9.47	6.3
2	Female	SIWPD	10	6.68	9.47	14.44	9.33	6.44
2	Male	WPD	10	6.73	9	14.4	9.11	6.45
2	Male	SIWPD	10	6.73	9	**14.41**	9.11	6.42
3	Female	WPD	10	6.17	11.11	15.06	8.31	9.35
3	Female	SIWPD	10	6.17	11.11	**15.36**	8.43	9.25
3	Male	WPD	10	5.92	11.51	14.89	7.64	9.87
3	Male	SIWPD	10	5.92	11.51	**14.89**	7.65	9.97

Table 5.6: Influence of shift invariance on WPD-based denoising performance. Use of SureShrink estimator.

lity is worse comparing to enhanced by Wiener filter speech.

5.3.7 Shift Invariance

As it was mentioned in Section 4.2, some of the artifacts that are produced by denoising were related by scientists to the lack of shift invariance. Thus it was of high importance to check if the SIWPD-based denoising leads to improvement when compared to WPD-based denoising.

Results of WPD and SIWPD-based denoising for SureShrink estimator are presented in Table 5.6, and for Wiener estimator - in Table 5.7. There the entropy-based best basis selection algorithm with $L = 6$ and generalized Meyer mother wavelet $(N = 64, r = 1/5)$ was used.

According to the results that are presented in Tables 5.6, 5.7 we can see that the use of SIWPD does not necessarily lead to improvement in the resulting SNR of the enhanced speech, the speech quality does not improve too. Moreover, if we return to

#	Speaker	Library type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	WPD	10	6.06	11.51	14.79	8.55	8.75
1	Female	SIWPD	10	6.06	11.51	**14.91**	8.57	8.73
1	Male	WPD	10	5.96	11.53	13.49	7.43	9.91
1	Male	SIWPD	10	5.96	11.53	**13.64**	7.52	9.83
2	Female	WPD	10	6.68	9.47	13.63	8.77	7.25
2	Female	SIWPD	10	6.68	9.47	**13.77**	8.8	7.09
2	Male	WPD	10	6.73	9	**12.79**	8.27	6.89
2	Male	SIWPD	10	6.73	9	12.78	8.26	6.85
3	Female	WPD	10	6.17	11.11	14	8.03	9.35
3	Female	SIWPD	10	6.17	11.11	**14.09**	8.05	9.37
3	Male	WPD	10	5.92	11.51	**13.77**	7.48	9.95
3	Male	SIWPD	10	5.92	11.51	13.75	7.47	9.98

Table 5.7: Influence of shift invariance on WPD-based denoising performance. Use of Wiener estimator.

the test signals of Donoho and Johnstone [10] (Section 4.2) and perform SIWPD-based denoising, the conclusion is the same. The reason for improvement obtained by the Cycle Spinning denoising approach is that the averaging over time-shifts performs smoothing, thus suppressing the Gibbs-like artifacts and blurring sharp transitions in the signals.

Test signal	Library type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
Blocks	WPD	17	18.68	9.94	18.23	19.75	9.73
Blocks	SIWPD	17	18.68	9.94	**18.24**	19.85	9.64
Bumps	WPD	17	14.19	9.69	21.34	16.25	2.91
Bumps	SIWPD	17	14.19	9.69	**21.39**	16.25	4
HeaviSine	WPD	17	16.1	4.22	**21.65**	20.68	0.84
HeaviSine	SIWPD	17	16.1	4.22	21.61	20.67	0.87
Doppler	WPD	17	16.32	7.7	20.85	21.01	1.11
Doppler	SIWPD	17	16.32	7.7	**21.33**	21.09	1.79

Table 5.8: Influence of shift invariance on WPD-based denoising performance. Use of Visu-Shrink estimator for Donoho-Johnstone test signals. DNS mother wavelet (8'th order).

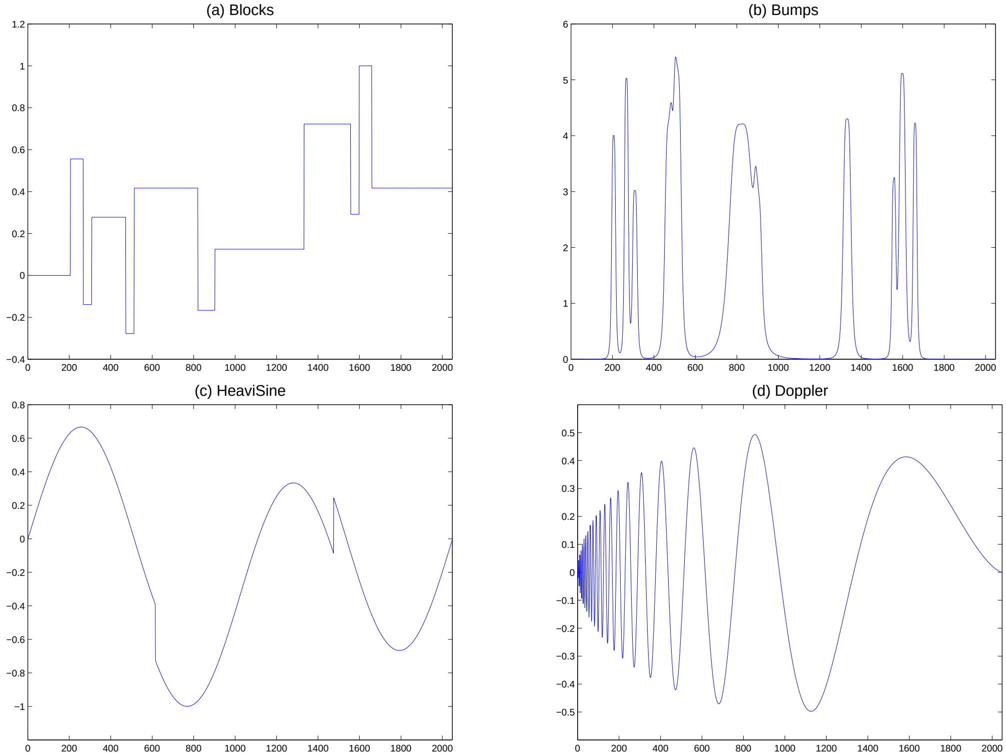


Figure 5.5: Test signals.

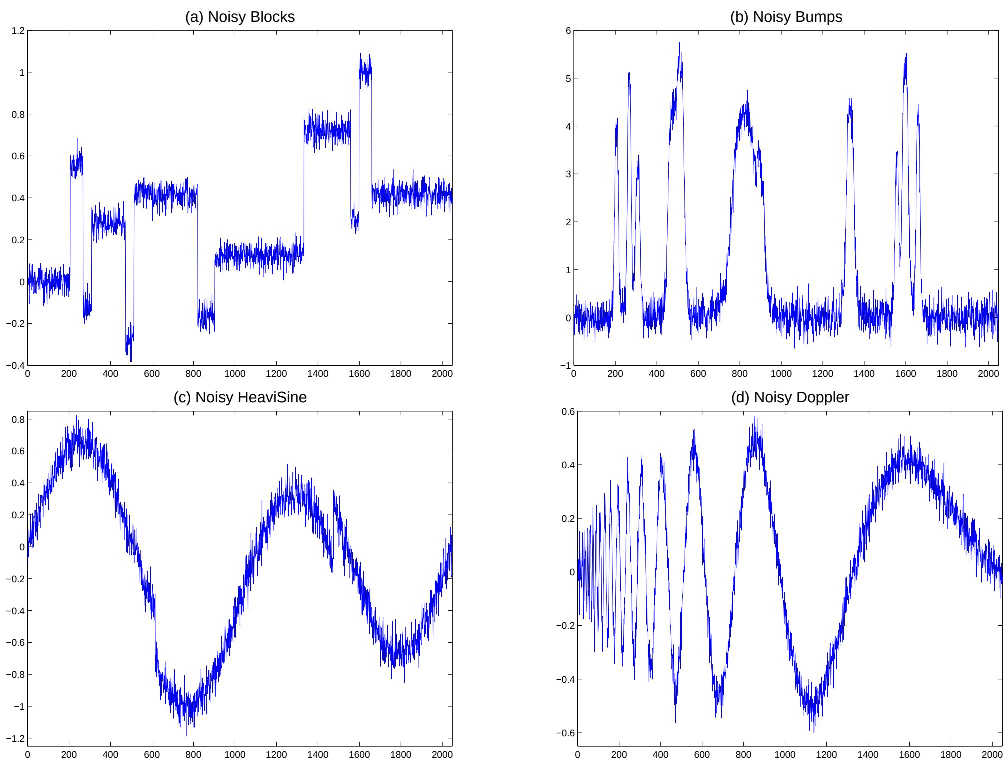


Figure 5.6: Noisy signals.

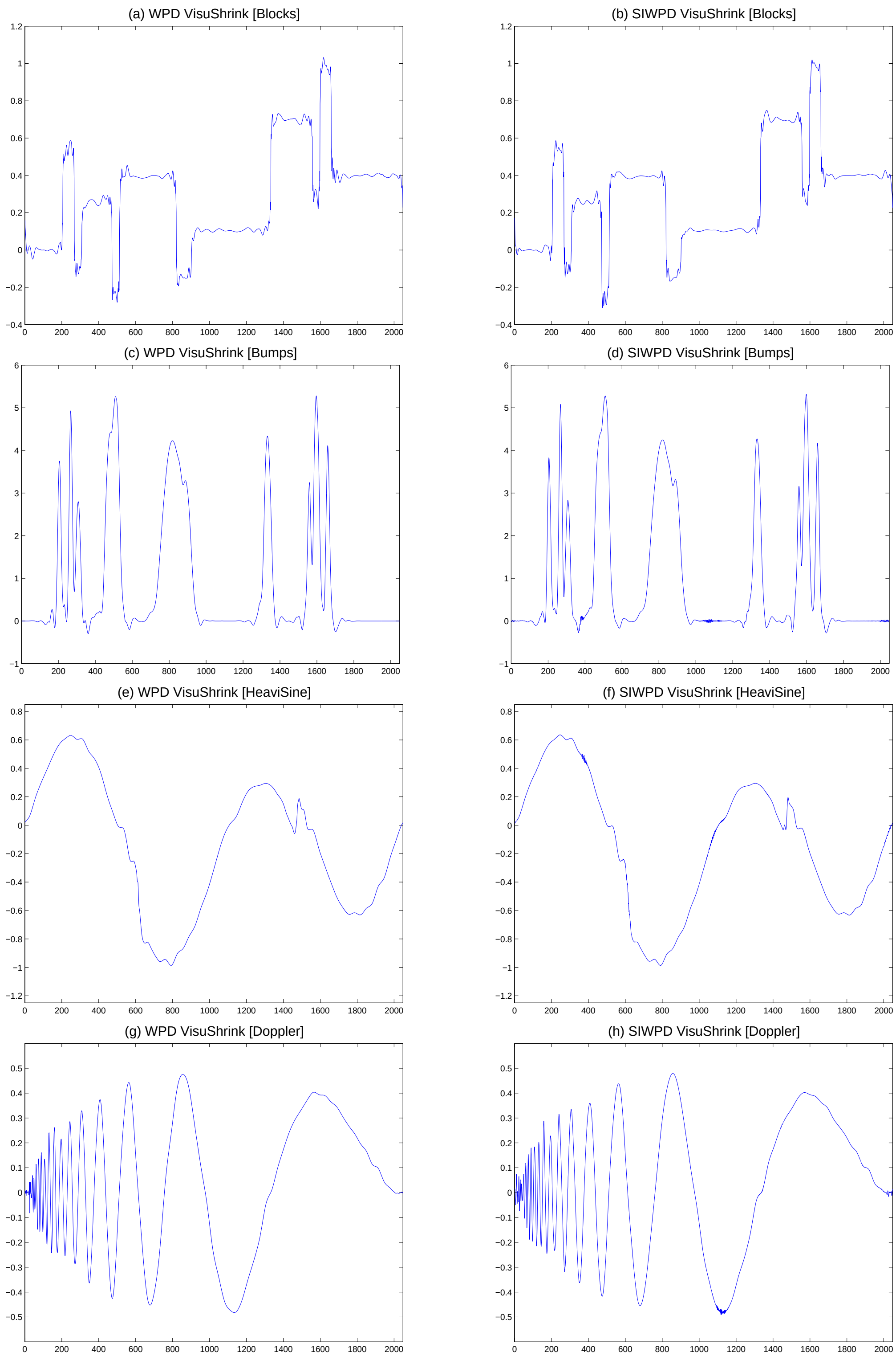


Figure 5.7: Signals, enhanced by VisuShrink estimator.

5.3.8 Framing and Utilization of Decision Directed a Priori SNR Estimation

It's obvious that when speech is corrupted by stationary colored noise, the variance of the noise process has to be estimated for each decomposition band. Moreover, estimation of the noise process variance for each of the decomposition bands makes it possible to utilize the decision directed a priori SNR estimation and to implement a Voice Activity Detector [21].

In order to utilize the decision directed a priori SNR estimation we have to segment the speech signal into frames. In general, segmentation of a speech signal can have many advantages, like obtaining better adaptation to transitions in time and decreasing the time delay in the denoising unit.

The Wiener gain function, defined by Eq. (5.8), can be rewritten in terms of the *a priori* SNR introduced in Section 2.2.2:

$$G_w^*(\Theta_{\ell,n}(j), \sigma_{\ell,n}(j)) = \frac{\xi_{\ell,n}(j)}{\xi_{\ell,n}(j) + 1}, \tag{5.15}$$

where j is the index of the analysis frame ($j = 1, 2, ...M$), $\sigma^2_{\ell,n}(j) = E||\mathbf{z}_{\ell,n}(j)||^2_{2,d}/d$ is the noise variance in the band indexed by the pair (ℓ, n), $\xi_{\ell,n}(j) = \frac{||\Theta_{\ell,n}(j)||^2_{2,d}}{\sigma^2_{\ell,n}(j)d}$ is the *a priori* SNR of the wavelet coefficients in that band, $\Theta_{\ell,n}(j)$ are the clean speech wavelet coefficients in the band, d is the number of the coefficients in the band ($d = 2^{(\ell+J)}$).

It's clear that the value of $\xi_{\ell,n}(j)$ has to be estimated and the estimated value $\widehat{\xi}_{\ell,n}(j)$ can then be used for gain calculation:

$$G_w(\mathbf{w}_{\ell,n}(j), \sigma_{\ell,n}(j)) = \frac{\widehat{\xi}_{\ell,n}(j)}{\widehat{\xi}_{\ell,n}(j) + 1}. \tag{5.16}$$

Like in Ephraim-Malah speech denoising algorithm, the estimation of the a priori SNR

#	Speaker	N, L, α	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
3	Female	128,5,0	10	6.17	11.11	15.96	8.55	9.17
3	Female	128,5,0.9	10	6.17	11.11	**16.5**	8.85	8.98
3	Female	128,J,0	10	6.17	11.11	13.49	7.58	10.1
3	Female	128,J,0.9	10	6.17	11.11	16.19	8.71	9.21
3	Female	256,5,0	10	6.17	11.11	15.89	8.5	8.91
3	Female	256,5,0.9	10	6.17	11.11	**16.4**	8.85	8.98
3	Female	256,J,0	10	6.17	11.11	13.77	7.78	10.02
3	Female	256,J,0.9	10	6.17	11.11	16.26	8.61	9.12
3	Female	512,5,0	10	6.17	11.11	15.85	8.61	9.34
3	Female	512,5,0.9	10	6.17	11.11	**16.37**	8.82	9.14
3	Female	512,J,0	10	6.17	11.11	13.8	7.78	10.01
3	Female	512,J,0.9	10	6.17	11.11	16.34	8.88	9.19
3	Female	1024,5,0	10	6.17	11.11	15.23	8.11	9.03
3	Female	1024,5,0.9	10	6.17	11.11	**16.28**	8.73	9.23
3	Female	1024,J,0	10	6.17	11.11	13.9	7.81	10.02
3	Female	1024,J,0.9	10	6.17	11.11	15.98	8.59	9.22

Table 5.9: Influence of frame size N, lowest decomposition level L and smoothing parameter α on WPD-based denoising performance. $L = J$ corresponds to the lowest allowed decomposition level ($\log_2 N$). The approximation of generalized Meyer mother wavelet with 64 taps and roll-off of 10% was used.

is performed using the decision directed approach:

$$\widehat{\xi}_{\ell,n}(j) = \alpha \frac{||\widehat{\Theta}_{\ell,n}(j-1)||_{2,d}^2}{E||\mathbf{z}_{\ell,n}(j-1)||_{2,d}^2} + (1-\alpha)\eta_s(\gamma_{\ell,n}(j),1), \quad j = 2,3,...M, \tag{5.17}$$

where $\gamma_{\ell,n}(j) = \frac{||\mathbf{w}_{\ell,n}(j)||_{2,d}^2}{E||\mathbf{z}_{\ell,n}(j)||_{2,d}^2}$ is the so-called *a posteriori* SNR and α is a smoothing parameter. The initial condition is:

$$\widehat{\xi}_{\ell,n}(1) = \alpha + (1-\alpha)\eta_s(\gamma_{\ell,n}(1),1). \tag{5.18}$$

If $\alpha = 0$, we return to the gain function, defined in (5.16):

$$G_w(\mathbf{w}_{\ell,n}(j), \sigma_{\ell,n}(j)) = \frac{\eta_s(\gamma_{\ell,n}(j),1)}{\eta_s(\gamma_{\ell,n}(j),1)+1} = \frac{\eta_s(||\mathbf{w}_{\ell,n}(j)||_{2,d}^2, E||\mathbf{z}_{\ell,n}(j)||_{2,d}^2)}{||\mathbf{w}_{\ell,n}(j)||_{2,d}^2}. \tag{5.19}$$

In order to choose the optimal frame size N, lowest decomposition level L and smoothing parameter α, examinations of full-subband WPD-based denoising with Hanning window of different lengths (50% overlapping) were made (Table 5.9).

#	Speaker	Library type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	WPD	10	6.06	11.51	17.37	9.58	8.52
1	Female	SIWPD	10	6.06	11.51	**17.38**	9.56	8.55
1	Male	WPD	10	5.96	11.53	**15.95**	8.48	9.62
1	Male	SIWPD	10	5.96	11.53	15.92	8.44	9.63
2	Female	WPD	10	6.68	9.47	16.01	9.58	6.62
2	Female	SIWPD	10	6.68	9.47	**16.03**	9.59	6.63
2	Male	WPD	10	6.73	9	**15.05**	9.54	6.31
2	Male	SIWPD	10	6.73	9	15	9.51	6.29
3	Female	WPD	10	6.17	11.11	**16.56**	9.01	9.12
3	Female	SIWPD	10	6.17	11.11	16.54	9	9.1
3	Male	WPD	10	5.92	11.51	**15.7**	7.94	9.96
3	Male	SIWPD	10	5.92	11.51	15.65	7.9	10

Table 5.10: Influence of shift invariance on WPD-based denoising performance. $N = 256$, $\alpha = 0.9$, $L = 8$ and the approximation of generalized Meyer mother wavelet with 64 taps and roll-off of 10% was used.

According to the results in Table 5.9 and the resulting speech quality, the preferable frame size is 256 samples. When using the framing with $\alpha = 0$, the optimal lowest decomposition level was found to be $L = 5$. The perception of the enhanced speech is better than for denoising without framing, but the background noise becomes colored. The reason is that the gains in a given band fluctuate from frame to frame. The smoothing operation ($\alpha \neq 0$) alleviates this problem: it arises from the physical model of speech generation and is based on the fact that spectral envelope of speech changes relatively slow.

The best results, in terms of SNR and the quality of the enhanced speech were obtained for the full subband decomposition ($\alpha = 0.9$, $L = \log_2 N$ - the lowest allowed decomposition level, where N is the length of the analysis frame). Usage of the *"decision directed"* approach to *a priori* SNR estimation with $\alpha = 0.9$ smoothes the undesirable fluctuations of the gain from frame to frame. The resulting speech sounds better than for denoising with $\alpha = 0$, and the residual noise sounds more like white noise. Therefore it's important to use the decision directed approach for a priori SNR estimation and full-subband WPD in the speech denoising process. The full-subband WPD results in a low computational complexity and allows simple implementation of the decision directed approach to a priori SNR estimation.

In Table 5.10 we present comparative results of full-subband WPD-based and full-subband SIWPD-based denoising. For full-subband SIWPD the best-basis search algorithm still have to be used. We used entropy as the cost function for the best-basis search algorithm. For both decompositions, the noise variance was estimated for each of subbands and a Wiener filter combined with the decision directed a priori SNR estimation was used. As it was previously mentioned, one can see that the shift invariance property of WPD does not improve performance of the wavelet-based speech denoising algorithm.

5.4 LTD-Based Speech Denoising

As it was mentioned in Section 3.3, LTD is a joint time-frequency representation that performs adaptive time axis segmentation. In this section we present briefly a LTD-based speech denoising algorithm. An extended discussion can be found in Appendix IV.

In our investigations we have found that the conclusions, made in the previous section with regard to WPD-based speech denoising, hold also for LTD-based denoising of speech: the optimal type of decomposition is "full-subsegment" decomposition, the optimal length of time segments is 256 samples, improving frequency localization improves speech quality and corresponding quantitative measures, and the Wiener estimator, combined with the decision directed a priori SNR estimation, is the best choice.

The Wiener gain function, defined by Eq. (5.9), can be adapted to LTD-based denoising:

$$G_w(w_{\ell,n,k}, \sigma_{\ell,n,k}) = \frac{\widehat{\xi}_{\ell,n,k}}{\widehat{\xi}_{\ell,n,k} + 1}. \tag{5.20}$$

where $\xi_{\ell,n,k} = \frac{|\theta_{\ell,n,k}|^2}{\sigma^2_{\ell,n,k}}$ is the *a priori* SNR of the clean speech LTD coefficient $\theta_{\ell,n,k}$, $\sigma^2_{\ell,n,k} = E|z_{\ell,n,k}|^2$ is the noise variance of the k-th "frequency" LTD coefficient in the time segment indexed by the pair (ℓ, n). The estimation of the a priori SNR is done using the decision directed approach:

$$\widehat{\xi}_{\ell,n,k} = \alpha \frac{|\widehat{\theta}_{\ell,n-1,k}|^2}{E|z_{\ell,n,k}|^2} + (1-\alpha)\eta_s(\gamma_{\ell,n,k}, 1), \tag{5.21}$$

where the *a posteriori* SNR, $\gamma_{\ell,n,k}$, is defined by

$$\gamma_{\ell,n,k} = \frac{|w_{\ell,n,k}|^2}{E|z_{\ell,n,k}|^2}. \tag{5.22}$$

The initial condition is:

$$\widehat{\xi}_{\ell,0,k} = \alpha + (1-\alpha)\eta_s(\gamma_{\ell,0,k}, 1). \tag{5.23}$$

Utilization of the decision directed a priori SNR estimation for full “subsegment” LTD requires tracking of the a priori SNR for $k \in \{0, 1, ..., 2^{(J+\ell)} - 1\}$, $n \in \{0, 1, ..., 2^{-\ell} - 1\}$ with $\ell = -L$ (constant).

According to the results shown in Table IV.4 and the resulting speech quality, the preferable lowest decomposition level is $L = 6$, the best value of α is $\alpha = 0.9$, and full “subsegment” decomposition is preferred.

5.5 WPD Applied to DCT Coefficients

When WPD applied to DCT coefficients it can be viewed as joint time-frequency representation that performs adaptive segmentation of time axis like LTD. In this section we present briefly speech denoising algorithm, which is based on applying WPD to DCT coefficients of the noisy speech signal. An extended discussion can be found in Appendix IV, were we show that the conclusions, made with regard to WPD-based speech denoising, hold also for speech denoising based on this type of representation.

In the case of this type of representation utilization of the decision directed a priori SNR estimation can be done exactly as for LTD. According to results shown in Table IV.8 and the resulting speech quality, the preferable lowest decomposition level is $L = 6$, the best value of α is $\alpha = 0.9$, and the best decomposition type is full “subsegment” decomposition. Since improved frequency localization leads to improved SNR and speech quality, the best choice of mother wavelet is DNS mother wavelet of 4’th order.

Chapter 6 : Alternative Speech Denoising Algorithms : A Comparative Performance Analysis

6.1 Ideal Denoising

6.1.1 Introduction

In this chapter we compare the speech denoising algorithms proposed herein to existing speech denoising algorithms.

Soon [36] et al have emphasized the advantages of DCT-based speech enhancement techniques compared to similar DFT-based. In this chapter we carry out a similar study comparing our algorithms with previous state of the art denoising schemes.

6.1.2 Real-valued Transforms vs. DFT

Let's summarize the definitions of the Wiener estimator (Eq. (5.9)) for three transform types, as follows:

1) WPD:

$$G_w = \frac{\eta_s(||\mathbf{w}_{\ell,n}(j)||^2_{2,d}, E||\mathbf{z}_{\ell,n}(j)||^2_{2,d})}{||\mathbf{w}_{\ell,n}(j)||^2_{2,d}}, \tag{6.1}$$

where $E||\mathbf{z}_{\ell,n}(j)||^2_{2,d} = \sigma^2_{\ell,n}d$ is an estimate of the noise variance in the band, indexed by the pair (ℓ, n), in the j-th time frame, and $d = 2^{(\ell+J)}$ is the number of wavelet coefficients in each subband at the ℓ-th resolution level.

2) LTD/WPD applied to DCT-I coefficients:

$$G_w = \frac{\eta_s(|w_{\ell,n,k}|^2, E|z_{\ell,n,k}|^2)}{|w_{\ell,n,k}|^2}, \tag{6.2}$$

where $E|z_{\ell,n,k}|^2 = \sigma^2_{\ell,n,k}$ is variance estimate of the noise k-th spectral component within the segment indexed by the pair (ℓ, n).

3) DFT/DCT:

$$G_w = \begin{cases} \left(1 - \frac{E|Z_k(j)|^2}{|Y_k(j)|^2}\right), & \frac{E|Z_k(j)|^2}{|Y_k(j)|^2} < 1 \\ 0, & otherwise \end{cases} = \frac{\eta_s(|Y_k(j)|^2, E|Z_k(j)|^2)}{|Y_k(j)|^2}, \tag{6.3}$$

where again $E|Z_k(j)|^2 = \sigma^2_k$ is variance estimate of noise k-th spectral component in the j-th time frame. Estimation of noise spectral components can be accomplished by averaging the spectrum over speech-free segments.

The decision directed a priori SNR estimation can be applied to any of these gain functions, and its purpose is to smooth out frame to frame gain fluctuations caused by fluctuations of the *squared spectral amplitude* of the noise process ($|Z_k|^2$ for DFT-based Wiener filter). For all of the mentioned above gain functions, the decision directed a priori SNR estimation improves the quality of enhanced speech and results in higher SNR. Obviously, usage of the (in practice, unknown) exact value of the squared spectral component for each k, ($|Z_k|^2$), of the noise process, instead of its estimate ($E|Z_k|^2$), must, necessarily, improve the performance of the denoising algorithms. We refer to a denoising

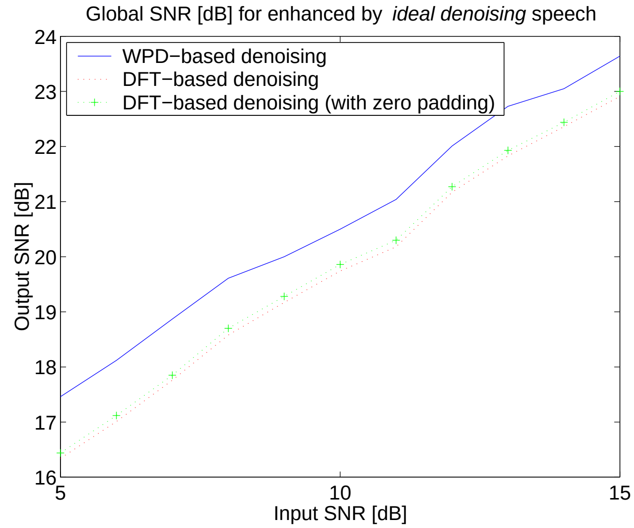


Figure 6.1: WPD-based vs. DFT-based ideal speech denoising.

process based on the exact values of the squared spectral components of the noise signal as "*ideal denoising*".

Results of ideal denoising process applied to test signal #1, when in each of gain functions (6.1), (6.2) and (6.3) the estimate of the squared spectral amplitude of the noise is replaced by its exact value, are presented in Tables 6.1, 6.2 and shown in Figure 6.1. The WPD-based denoising executes a full-subband decomposition with a generalized Meyer mother wavelet (64 taps, 10% roll-off) was used.

As one can see, the WPD-based denoising attains a consistently higher SNR. The reasons for that are:

1) While performing denoising with N samples per time frame, we obtain N complex-valued spectral (DFT) coefficients. WPD (or any real-valued transform) yields N inde-

Transform	Input SNR	5	6	7	8	9	10
WPD	Output SNR	**17.46**	**18.12**	**18.87**	**19.61**	**20**	**20.48**
DFT	Output SNR	16.34	17.01	17.76	18.58	19.17	19.74
DFT_{2N}	Output SNR	16.44	17.12	17.85	18.7	19.28	19.86

Table 6.1: WPD-based vs. DFT-based ideal speech denoising. DFT_{2N} corresponds to DFT of zero padded segments (double length).

Transform	Input SNR	11	12	13	14	15
WPD	Output SNR	**21.04**	**22.01**	**22.73**	**23.05**	**23.64**
DFT	Output SNR	20.18	21.17	21.83	22.36	22.91
DFT_{2N}	Output SNR	20.3	21.27	21.93	22.44	23

Table 6.2: WPD-based vs. DFT-based ideal speech denoising. DFT_{2N} corresponds to DFT of zero padded segments (double length).

pendent real-valued coefficients. Consequently, DFT-based denoising has at its disposal only half ($\sim N/2$) the number of amplitude coefficients.

2) While performing a DFT-based denoising, if $|Y_k|^2 > |Z_k|^2$ we do not modify the phase of noisy signal, taking it as the optimal estimate of clean speech phase [18]. Whenever $|Y_k|^2 \leq |Z_k|^2$ we take the zero phase as an estimate of clean speech phase. On the other hand, speech denoising, based on some real-valued transform, achieve an exact phase reconstruction of the clean speech signal (expressed via the sign of the real-valued transform coefficients) (see Appendix II). This is an intrinsic advantage of real-valued transforms compared to DFT-based denoising.

A fair comparison between a WPD and a DFT-based ideal denoising is achieved by zero padding the time segments in DFT-based denoising, thus improving its frequency resolution, and subsequently, improving the global SNR of enhanced speech by $0.08 \div 0.12$ [dB] (Fig. 6.1). The results show that the exact phase reconstruction associated with real-valued transforms leads to global SNR improvement by $0.69 \div 1.12$ [dB].

6.1.3 Simulation Results

In this subsection we present the results of *ideal denoising* for all the proposed algorithms, the DFT and DCT-based Wiener estimator, and the DFT-based Ephraim-Malah (E-M) log-spectral amplitude estimator. The results, presented in Table 6.3, show that DCT and

WPD-based Wiener estimator overperform the DFT_{2N}-based one. Moreover, they show that the Wiener estimator, based on CPD and WPD applied to DCT-I coefficients, attains lower SNR than the Wiener estimator, based on WPD. The reason is that WPD, DCT and DFT-based denoising algorithms divide the time axis into overlapping segments, while introducing some kind of averaging and improving SNR. A fair comparison is achieved by applying denoising algorithms, which are based on CPD and WPD applied to DCT-I coefficients, using time-segments of 512 samples per segment with 25% overlap and Hanning window; the number of decomposition levels is $L = 1$. The results are presented in the Table 6.4.

#	Speaker	Estimator, transform	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	Wiener,WPD	10	6.06	11.51	**20.48**	11.87	2.43
1	Female	Wiener,DCT	10	6.06	11.51	20.45	11.86	2.38
1	Female	Wiener,CPD	10	6.06	11.51	19.71	11.19	3.31
1	Female	Wiener,WPD (DCT-I)	10	6.06	11.51	19.78	11.2	3.25
1	Female	Wiener,DFT_{2N}	10	6.06	11.51	19.86	11.46	2.36
1	Female	E-M,DFT_{2N}	10	6.06	11.51	19.51	10.96	2.86

Table 6.3: Comparative performance of different speech denoising algorithms.

#	Speaker	Estimator, transform	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	Wiener,CPD	10	6.06	11.51	20.36	11.8	2.42
1	Female	Wiener,WPD (DCT-I)	10	6.06	11.51	**20.5**	11.9	2.43

Table 6.4: Comparative performance of two different speech denoising algorithms.

In Table 6.5 we show the influence of frequency resolution and localization on WPD-based ideal denoising performance. The full-subband decomposition ($L = 8$) with prior segmentation into frames (256 samples per frame) was used. As previously mentioned (Sections 5.3.6, IV.1.3 and IV.2.3), improvement in frequency resolution and localization

#	Speaker	L, r	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	8,0.1	10	6.06	11.51	**20.48**	11.87	2.43
1	Female	7,0.1	10	6.06	11.51	19.72	11.44	2.53
1	Female	6,0.1	10	6.06	11.51	19.14	11.02	2.63
1	Female	8,0.2	10	6.06	11.51	**20.12**	11.69	2.55
1	Female	8,1/3	10	6.06	11.51	19.93	11.58	2.6
1	Female	8,DNS(8)	10	6.06	11.51	19.3	11.19	2.58

Table 6.5: Influence of frequency resolution and localization on WPD-based ideal denoising performance. r denotes roll-off of modified Meyer QMF $m(\omega)$.

#	Speaker	η	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	6	10	6.06	11.51	18.41	10.5	3.68
1	Female	128	10	6.06	11.51	**19.71**	11.19	3.31
1	Male	6	10	5.96	11.53	17.52	9.32	3.36
1	Male	128	10	5.96	11.53	**18.21**	9.86	2.95
2	Female	6	10	6.68	9.47	17.36	11.37	3.93
2	Female	128	10	6.68	9.47	**18.21**	11.81	3.95
2	Male	6	10	6.73	9	16.49	11.02	3.1
2	Male	128	10	6.73	9	**17.02**	11.23	2.5
3	Female	6	10	6.17	11.11	17.72	9.74	2.35
3	Female	128	10	6.17	11.11	**18.63**	10.17	2.16
3	Male	6	10	5.92	11.51	17.28	8.72	3.55
3	Male	128	10	5.92	11.51	**17.96**	9.03	3.04

Table 6.6: Influence of CP basis functions frequency localization on LTD-based ideal denoising performance.

leads to improved speech quality and higher global SNR.

In the Table 6.6 we present results of CPD-based ideal denoising for different values of η (Section 3.3.2): 6 and 128. Clearly, an improved frequency localization results in improved denoising performance.

6.2 Practical Denoising

Herein, we compare the proposed speech denoising algorithms to the Ephraim-Malah (E-M) MMSE Log-SA estimator [19] (Section 2.2.2) and to DFT-based Wiener estimator. In all cases the decision directed a priori SNR estimation was utilized. The smoothing parameter $\alpha = 0.92$ was empirically found to be the best value for E-M algorithm, while for all other algorithms $\alpha = 0.9$ was used. The results are summarized in Table 6.7.

These results indicate that, for each of the tested speech signals, the DFT-based Wiener estimator attains the highest global SNR, segmental SNR and LSD. The quality of the enhanced speech is similar for all the algorithms. Examples of the results of *Linear Prediction Coding* (LPC) analysis for two frames of enhanced speech signals are depicted in Figure 6.2. The notable difference is the level and type of a residual background noise. All of the algorithms, Ephraim-Malah being the exception, introduce a colored background noise, that was found to be disturbing the listener. The DFT-based Wiener estimator is characterized by the lowest level of the residual noise, and is superior to the proposed algorithms. The Ephraim-Malah algorithm is characterized by a higher level of background noise then DFT and WPD-based Wiener estimator, but, advantageously, the background noise is almost white. It's important to note that Wiener estimator, by definition, minimizes the mean squared error estimating an unknown signal $\mathbf{f}$, while E-M algorithm minimizes the mean squared error estimating the log-spectra of the unknown signal. Hence, it's expected that the DFT-based Wiener estimator achieves higher SNR than the E-M algorithm.

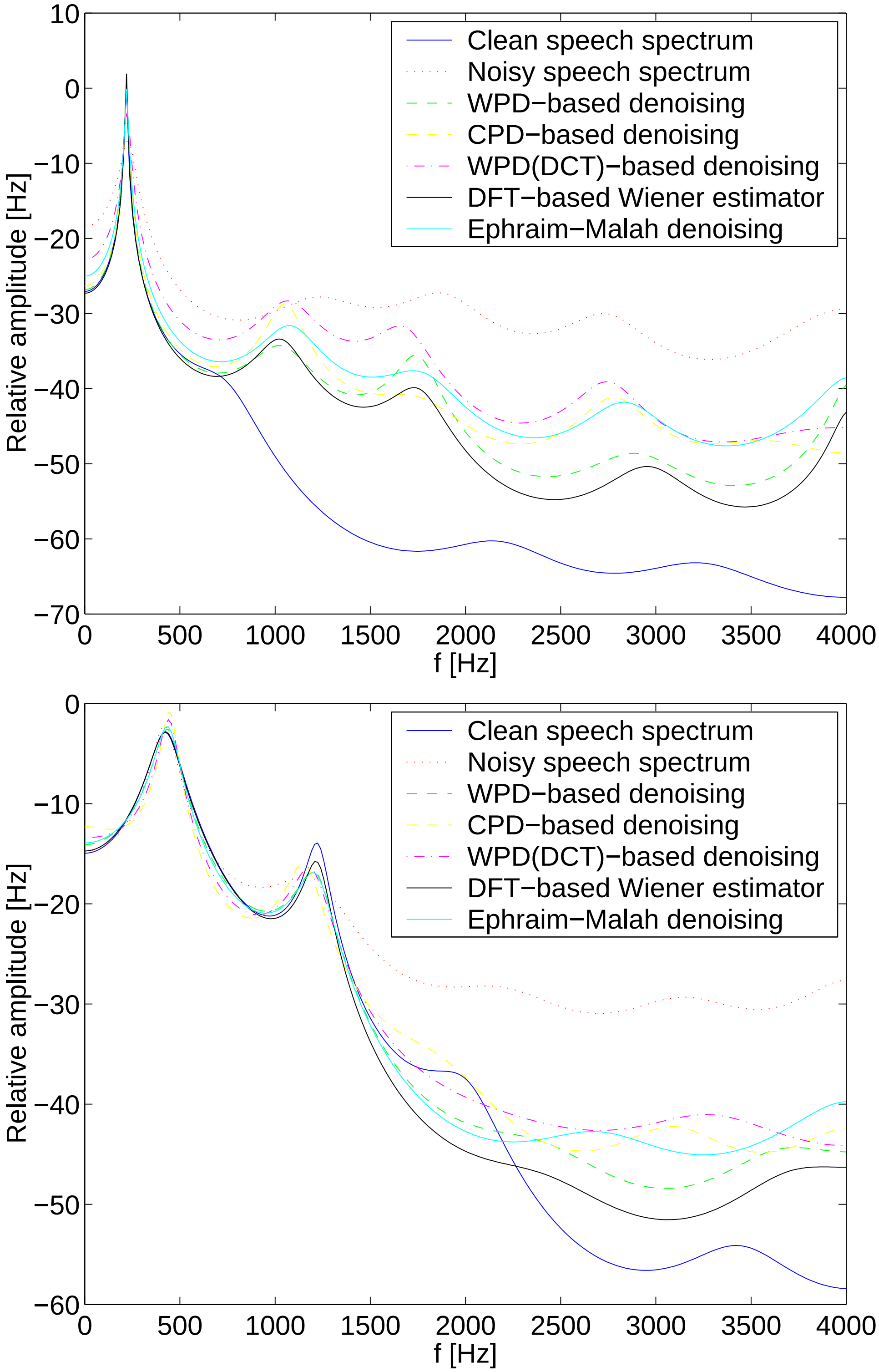


Figure 6.2: LPC analysis ($p = 10$) for clean, noisy and enhanced speech signals.

6.3 Discussion

The comparisons, presented in the previous subsections, show that despite the advantages of WPD and CPD-based algorithms under ideal denoising conditions, in practice (i.e., with an estimated noise variance) the DFT-based denoising algorithms are found to be better. The reasons are:

1) Given the noisy observations $\mathbf{y}$, we can't know the exact values of the noise squared spectral components. Hence, using only the estimated averages of the noise squared spectral components we can't exactly reconstruct the clean speech phase.

2) It is shown in Appendix I, that if the additive noise is white and Gaussian, the variance of its squared spectral components, obtained by real-valued transform, is twice (except for the DC coefficient) the variance of the noise squared spectral components, obtained by the DFT. This leads to higher deviations of noise squared spectral amplitude from its estimated value, and subsequently to higher frame to frame gains fluctuations (segment to segment gains fluctuations for CPD-based denoising) thus reducing the resulting global and segmental SNR. The frame to frame gains fluctuations cause the residual background noise to be colored.

Although, the proposed speech denoising algorithms do not outperform the DFT-based algorithms, they still possess several advantages:

1) The WPD-based denoising can be easily incorporated into a WPD-based speech coding system.

2) It's important to note that LTD can be used as a time-segmentation tool. Thus, the LTD-based denoising algorithm can be easily implemented in speech analysis systems, which require adaptive segmentation.

3) The LTD-based speech denoising algorithm can be used in conjunction with the Shift-Invariant Adaptive Polarity Local Trigonometric Decomposition (SIAP-LTD) [4], that possess the shift-invariance property, which is potentially critical for recognition applications.

#	Speaker	Denoising algorithm	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	WPD	10	6.06	11.51	17.37	9.58	8.52
1	Female	CPD	10	6.06	11.51	16.69	9.17	8.56
1	Female	WPD (DCT-I)	10	6.06	11.51	16.49	9.04	8.81
1	Female	DFT	10	6.06	11.51	**17.83**	9.91	8.07
1	Female	E-M	10	6.06	11.51	17.22	9.45	8.67
1	Male	WPD	10	5.96	11.53	15.95	8.48	9.62
1	Male	CPD	10	5.96	11.53	15.44	8.1	9.69
1	Male	WPD (DCT-I)	10	5.96	11.53	15.31	8.07	9.73
1	Male	DFT	10	5.96	11.53	**16.41**	8.75	9.5
1	Male	E-M	10	5.96	11.53	16.00	8.44	9.69
2	Female	WPD	10	6.68	9.47	16.01	9.58	6.62
2	Female	CPD	10	6.68	9.47	15.13	9.43	6.76
2	Female	WPD (DCT-I)	10	6.68	9.47	15.02	9.35	6.95
2	Female	DFT	10	6.68	9.47	**16.22**	10.26	6.29
2	Female	E-M	10	6.68	9.47	15.7	9.79	6.81
2	Male	WPD	10	6.73	9	15.05	9.54	6.31
2	Male	CPD	10	6.73	9	14.37	9.06	6.39
2	Male	WPD (DCT-I)	10	6.73	9	14.22	9	6.49
2	Male	DFT	10	6.73	9	**15.4**	9.85	6.2
2	Male	E-M	10	6.73	9	15.06	9.5	6.48
3	Female	WPD	10	6.17	11.11	16.56	9.01	9.12
3	Female	CPD	10	6.17	11.11	15.94	8.59	9.2
3	Female	WPD (DCT-I)	10	6.17	11.11	15.83	8.42	9.35
3	Female	DFT	10	6.17	11.11	**17.01**	9.24	8.99
3	Female	E-M	10	6.17	11.11	16.46	8.93	9.23
3	Male	WPD	10	5.92	11.51	15.7	7.94	9.96
3	Male	CPD	10	5.92	11.51	15.24	7.68	9.99
3	Male	WPD (DCT-I)	10	5.92	11.51	15.16	7.53	10.09
3	Male	DFT	10	5.92	11.51	**16.11**	8.12	9.83
3	Male	E-M	10	5.92	11.51	15.74	7.97	10.01

Table 6.7: Comparison of the proposed speech denoising algorithms to the state of the art speech denoising algorithms.

Chapter 7 : Summary and Conclusions

7.1 Summary

We have developed speech denoising algorithms based on WPD and LTD. The proposed speech denoising algorithms utilize the *decision directed* approach to *a priori* SNR estimation. It takes into account slow changes of the speech spectral envelope and overcomes the undesirable fluctuations in the estimate of the noise squared spectral amplitude, thus improving the denoising performance. We have shown that artifacts, introduced by wavelet-based denoising algorithms [14, 16, 10, 33, 5], applied to speech enhancement, can be particularly suppressed by increasing temporal support of the basis functions. Moreover, improvement in frequency localization of the basis functions improves the speech denoising performance. It also has been shown that shift-invariance achieved by Shift-Invariant Wavelet Packet Decomposition (SIWPD) does not contribute to artifacts suppression and does not guarantee an improved denoising performance.

We have compared the proposed speech denoising algorithms to the state of the art speech denoising algorithms [18, 19]. Denoising based on the presumption of prior knowledge of the squared spectral amplitude of the noise is referred to as *ideal* denoising.

Quite expectedly, simulations confirm that such ideal denoising attains higher SNR than the practical one. Moreover, we have proved that for WGN, the of squared spectral amplitude of the coefficients, obtained by a real-valued orthonormal transform, is twice the variance obtained by using the DFT. This explains the result that the state of the art speech denoising algorithms perform better (although close) than the proposed speech denoising algorithms.

Despite the fact that the speech denoising algorithms proposed herein do not possess clear advantage over the DFT-based algorithms, they may have merit in a wider sense. For example, WPD-based denoising can be easily incorporated into a WPD-based speech coding system. Also, LTD can be used as a time-segmentation tool. Thus, the LTD-based denoising algorithm can be conveniently implemented in speech analysis systems, which require adaptive time-segmentation.

7.2 Future Research

There are a number of potentially promising topics for future study:

1) The LTD-based speech denoising algorithm can be used in conjunction with the Shift-Invariant Adaptive Polarity Local Trigonometric Decomposition (SIAP-LTD) [4], that possesses the shift-invariance property, which is potentially critical for recognition applications.

2) Saito and Coifman [32] have described a best-basis method for signal classification problems, which is based on the conventional WPD and LTD. They have used cross entropy as a basis selection criterion. Thus, it picks out the most significant basis functions to serve as feature extractors, that are subsequently utilized in an ordinary classifier.

The sensitivity of the expansion coefficients to signal shifts is a serious drawback of such classification methods. Saito [31] proposed to reduce the sensitivity to shift-variance by creating from each training signal a few circularly-shifted versions. This not only increases the computational complexity and required memory resources, but generates feature extractors that remain shift-variant. The proposed in [5] Shift-Invariant Wavelet Packet Decompositions (as well as SIAP-LTD [4]) are expected to overcome this difficulty.

Appendix I : Fluctuations of Squared Spectral Amplitude

Real-valued Transforms

Let $\underline{\mathbf{x}} = \{x_i\}_{i=0}^{N-1}$ be samples of WGN, where $x_i \sim \mathcal{N}(0, \sigma_x^2)$. When we use a real-valued orthonormal transformation

$$\underline{\mathbf{X}} = \mathbf{T} \cdot \underline{\mathbf{x}}, \tag{I.1}$$

where $\underline{\mathbf{X}} = \{X_k\}_{k=0}^{N-1}$ and $\mathbf{T}$ represents an orthonormal matrix. It is well known that $X_k \sim \mathcal{N}(0, \sigma_x^2)$. Let us turn now to the squared amplitude $|X_k|^2$.

It's known [25], that if $X_k \sim \mathcal{N}(0, \sigma_x^2)$, then

$$E\{|X_k|^n\} = \begin{cases} 1 \cdot 3 \cdot ... \cdot (n-1) \cdot \sigma_x^n \ , \ n = 2m \\ 2^m \cdot m! \cdot \sigma_x^n \cdot \sqrt{\frac{2}{\pi}} \ , \qquad n = 2m+1. \end{cases} \tag{I.2}$$

Thus,

$$\mu_{|X_k|^2} = E\{|X_k|^2\} = \sigma_x^2 \tag{I.3}$$

and

$$\begin{aligned} \sigma^2_{|X_k|^2} &= Var\{|X_k|^2\} = E\{(|X_k|^2 - E\{|X_k|^2\})^2\} = \\ &= E\{|X_k|^4\} - (E\{|X_k|^2\})^2 = 3\sigma_x^4 - \sigma_x^4 = 2\sigma_x^4. \end{aligned} \tag{I.4}$$

Here, the squared spectral amplitude $|X_k|^2$ has the so-called chi-square distribution with one degree of freedom:

$$f_y(y) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})\sigma_x} y^{\frac{n}{2}-1} e^{-\frac{y}{2\sigma_x^2}} U(y), \tag{I.5}$$
$$f_{|X_k|^2}(|X_k|^2) = f_y(y)|_{y=|X_k|^2,\ n=1},$$

where $U(y)$ is a step function.

Discrete Fourier Transform

Let $\underline{\mathbf{X}} = \{X_k\}_{k=0}^{N-1}$ denote the DCT coefficients of $\underline{\mathbf{x}} = \{x_i\}_{i=0}^{N-1}$ which are samples of a WGN process with zero mean and variance σ_x^2. It is well known [40] that the DCT coefficients $\underline{\mathbf{X}}$ can be obtained by applying a $2N$-point DFT to the sequence $\underline{\mathbf{u}} = \{u_i\}_{i=0}^{2N-1}$, defined by

$$u_i = \begin{cases} x_i, & i = 0, 1, ...N-1, \\ 0, & i = N, ...2N-1. \end{cases} \tag{I.6}$$

The DCT coefficients $\{X_k\}_{k=0}^{N-1}$ and the DFT coefficients $\{U_k\}_{k=0}^{N-1}$ are related then by

$$X_k = c_k \cdot |U_k| \cdot \cos{(\vartheta_k - \frac{\pi k}{2N})}, \quad k = 0, 1, ...N-1. \tag{I.7}$$

Here, ϑ_k represents the phase of U_k and has uniform distribution $\vartheta_k \sim \mathcal{U}[0; 2\pi]$, and

$$c_k = \begin{cases} 1, & k = 0, \\ \sqrt{2}, & k = 1, 2, ...N-1. \end{cases} \tag{I.8}$$

Forming

$$|X_k|^2 = c_k^2 \cdot |U_k|^2 \cdot \cos^2{(\vartheta_k - \frac{\pi k}{2N})}$$

with c_k, $|U_k|^2$ and ϑ_k independent and c_k a constant for a fixed k, leads to

$$E\{|X_k|^2\} = c_k^2 \cdot E\{|U_k|^2\} \cdot E\{\cos^2{(\vartheta_k - \frac{\pi k}{2N})}\}. \tag{I.9}$$

It's clear that

$$c_k^2 = \begin{cases} 1, & k = 0, \\ 2, & k = 1, 2, ...N-1 \end{cases}$$

and

$$E\{\cos^2(\vartheta_k - \frac{\pi k}{2N})\} = \begin{cases} \cos^2 0, & k = 0 \\ \int_0^{2\pi} \cos^2(\vartheta_k - \frac{\pi k}{2N}) \frac{1}{2\pi} d\vartheta_k, & k = 1, 2, ...N-1 \end{cases} =$$

$$= \begin{cases} 1, & k = 0, \\ \frac{1}{2}, & k = 1, 2, ...N-1. \end{cases}$$

Hence

$$E\{|X_k|^2\} = \begin{cases} 1 \cdot E\{|U_k|^2\} \cdot 1, & k = 0 \\ 2 \cdot E\{|U_k|^2\} \cdot \frac{1}{2}, & k = 1, 2, ...N-1 \end{cases} = E\{|U_k|^2\} \tag{I.10}$$

and according to (I.3)

$$\mu_{|U_k|^2} = E\{|U_k|^2\} = \sigma_x^2. \tag{I.11}$$

For $\sigma^2_{|U_k|^2}$ we get:

$$\sigma^2_{|U_k|^2} = Var\{|U_k|^2\} = E\{(|U_k|^2 - E\{|U_k|^2\})^2\} =$$

$$E\{|U_k|^4\} - (E\{|U_k|^2\})^2 = E\{|U_k|^4\} - \sigma_x^4. \tag{I.12}$$

Here

$$|X_k|^4 = c_k^4 \cdot |U_k|^4 \cdot \cos^4(\vartheta_k - \frac{\pi k}{2N}),$$

c_k, $|U_k|^2$ and ϑ_k are independent, c_k is a constant for a fixed k. Thus

$$E\{|X_k|^4\} = c_k^4 \cdot E\{|U_k|^4\} \cdot E\{\cos^4(\vartheta_k - \frac{\pi k}{2N})\},$$

and

$$E\{|U_k|^4\} = \frac{E\{|X_k|^4\}}{c_k^4 \cdot E\{\cos^4(\vartheta_k - \frac{\pi k}{2N})\}}. \tag{I.13}$$

It's clear that

$$c_k^4 = \begin{cases} 1, & k=0, \\ 4, & k=1,2,...N-1, \end{cases}$$

and

$$E\{\cos^4(\vartheta_k - \frac{\pi k}{2N})\} = \begin{cases} 1 \quad (\vartheta_0 = 0), & k=0 \\ \int_0^{2\pi} \cos^4(\vartheta_k - \frac{\pi k}{2N})\frac{1}{2\pi}d\vartheta_k, & k=1,2,...N-1 \end{cases} =$$

$$= \begin{cases} 1, & k=0 \\ \frac{1}{2\pi}\left(\frac{3}{8}\vartheta_k + \frac{1}{4}\mathrm{sin}2\vartheta_k + \frac{1}{32}\mathrm{sin}4\vartheta_k\right)|_0^{2\pi}, & k=1,2,...N-1 \end{cases} = \begin{cases} 1, & k=0, \\ \frac{3}{8}, & k=1,2,...N-1. \end{cases}$$

Finally,

$$E\{|U_k|^4\} = \begin{cases} E\{|X_k|^4\}, & k=0, \\ \frac{E\{|X_k|^4\}}{4\cdot\frac{3}{8}}, & k=1,2,...N-1, \end{cases}$$

and using (I.2), we obtain

$$E\{|U_k|^4\} = \begin{cases} 3\sigma_x^4, & k=0, \\ 2\sigma_x^4, & k=1,2,...N-1, \end{cases}$$

and substitution of the last equation into (I.12) gives:

$$\sigma^2_{|U_k|^2} = \begin{cases} 2\sigma_x^4, & k=0, \\ \sigma_x^4, & k=1,2,...N-1, \end{cases} \tag{I.14}$$

Appendix II : Clean Speech Phase Reconstruction

As it was claimed in Section 6.1.2, it is shown here that real-valued transform-based speech denoising, that assumes knowledge of the squared spectral amplitude of the noise, allows perfect reconstruction of clean speech phase.

Let X_k denote the expansion coefficients of the clean speech signal, Z_k the expansion coefficients of noise process, and $Y_k = X_k + Z_k$ the expansion coefficients of noisy speech. Given the exact value of the noise squared-spectral-amplitude, $|Z_k|^2$, an estimate $\widehat{X}_k$ of X_k, obtained by the use of Wiener filter, is defined by

$$\widehat{X}_k = Y_k \cdot G_k = Y_k \cdot \begin{cases} \left(\frac{|Y_k|^2 - |Z_k|^2}{|Y_k|^2}\right), & |Y_k|^2 > |Z_k|^2, \\ 0, & otherwise. \end{cases} \tag{II.1}$$

There are three possible cases:

1) $\text{sign}(X_k) = \text{sign}(Z_k)$. Then

$\text{sign}(\widehat{X}_k) = \text{sign}(Y_k) = \text{sign}(X_k + Z_k) = \text{sign}(X_k)$.

2) $\text{sign}(X_k) = -\text{sign}(Z_k)$. We distinguish between the following three cases:

a) $|Z_k| \geq |X_k| \;\Rightarrow\; |Y_k|^2 = |X_k + Z_k|^2 \leq |Z_k|^2$, and the reconstructed phase will be 0.

b) $\frac{1}{2}|X_k| \leq |Z_k| \leq |X_k| \;\Rightarrow\; |Y_k|^2 = |X_k + Z_k|^2 \leq \frac{1}{4}|X_k|^2 < |Z_k|^2$, and the reconstructed phase will be 0.

c) $|Z_k| < \frac{1}{2}|X_k| \Rightarrow |Y_k|^2 = |X_k + Z_k|^2 > \frac{1}{4}|X_k|^2 > |Z_k|^2$, $\text{sign}(\widehat{X}_k) = \text{sign}(Y_k) = \text{sign}(X_k + Z_k) = \text{sign}(X_k)$.

Thus, real-valued transform-based speech denoising, that assumes knowledge of the squared spectral amplitude of the noise, allows perfect reconstruction of clean speech phase.

Appendix III : Derivation of Ephraim-Malah and State of the Art Wavelet-Based Estimators

III.1 Ephraim-Malah Log-Spectral Amplitude Estimator

The estimation problem of the STSA is formulated as that of estimation of the amplitude of each Fourier expansion coefficient of the speech signal $\mathbf{f} = \{f(t), 0 \leq t \leq T\}$, given the noisy process $\mathbf{y} = \{y(t), 0 \leq t \leq T\}$. Both speech and noise process are assumed to be Gaussian. The Fourier expansion coefficients of the speech process, as well as the noise process, are modeled as statistically independent Gaussian random variables. The Gaussian model is motivated by the central limit theorem, as each Fourier expansion coefficient is a weighted sum of random variables. The statistical independence assumption is motivated by the fact that the correlation between the spectral components reduces as the analysis interval length increases.

Let the $F_k = A_k e^{j\alpha_k}$, Z_k and $Y_k = R_k e^{j\vartheta_k}$ denote the k-th Fourier expansion coefficients of the speech signal, the noise process and the noisy observations, respectively, in the analysis interval $[0, T]$. According to the formulation of the estimation problem given above, we are looking for the estimator $\widehat{A}_k$, which minimize the following distortion measure:

$$E\{(\log A_k - \log \widehat{A}_k)^2\}. \tag{III.1}$$

This estimator is easily shown to be

$$\widehat{A}_k = \exp\{E[\ln A_k|\mathbf{y}]\} \tag{III.2}$$

and it is independent of the basis chosen for the log in (III.2). As it was noted in [18], the estimator (III.2) equals

$$\widehat{A}_k = \exp\{E[\ln A_k|Y_k]\}. \tag{III.3}$$

The evaluation of $E[\ln A_k|\mathbf{y}]$ for the Gaussian model assumed here is conveniently done by utilizing the moment generating function of $\ln A_k$ given Y_k. Let $D_k = \ln A_k$. Then the moment generating function $\Phi_{D_k|Y_k}$ of D_k given Y_k equals

$$\Phi_{D_k|Y_k} = E\{\exp(\mu D_k)|Y_k\} = E\{A_k^\mu|Y_k\}. \tag{III.4}$$

$E[\ln A_k|Y_k]$ is obtained from $\Phi_{D_k|Y_k}$ by

$$E[\ln A_k|Y_k] = \frac{d}{d\mu}\Phi_{D_k|Y_k}(\mu)|_{\mu=0}. \tag{III.5}$$

From (III.4)

$$\Phi_{D_k|Y_k}(\mu) = E\{A_k^\mu|Y_k\} = \frac{\int_0^\infty \int_0^{2\pi} a_k^\mu p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k da_k}{\int_0^\infty \int_0^{2\pi} p(Y_k|a_k,\alpha_k)p(a_k,\alpha_k)d\alpha_k da_k}. \tag{III.6}$$

On the basis of the Gaussian model assumed here, $p(Y_k|a_k,\alpha_k)$ and $p(a_k,\alpha_k)$ are given by [18]:

$$p(Y_k|a_k,\alpha_k) = \frac{1}{\pi\lambda_z(k)}\exp\left\{-\frac{1}{\lambda_z(k)}|Y_k - a_k e^{j\alpha_k}|^2\right\}, \tag{III.7}$$

$$p(a_k, \alpha_k) = \frac{a_k}{\pi \lambda_f(k)} \exp\left\{-\frac{a_k^2}{\lambda_f(k)}\right\}, \tag{III.8}$$

where $\lambda_z(k) \equiv E\{|Z_k|^2\}$ and $\lambda_f(k) \equiv E\{|F_k|^2\}$ are the variances of the noise and the signal k-th spectral components. Substituting (III.7) and (III.8) into (III.6), and using the integral representation of the modified Bessel function of zero order $I_0(\cdot)$, we obtain:

$$\Phi_{D_k|Y_k}(\mu) = \frac{\int_0^\infty a_k^{\mu+1} \exp(-a_k^2/\lambda_k) I_0(2a_k\sqrt{v_k/\lambda_k}) da_k}{\int_0^\infty a_k \exp(-a_k^2/\lambda_k) I_0(2a_k\sqrt{v_k/\lambda_k}) da_k}, \tag{III.9}$$

where λ_k satisfies $1/\lambda_k = 1/\lambda_f + 1/\lambda_z$, and v_k is defined by

$$v_k \equiv \frac{\xi_k}{1+\xi_k}\gamma_k \ , \quad \xi_k \equiv \frac{\lambda_f(k)}{\lambda_z(k)} \ , \quad \gamma_k \equiv \frac{R_k^2}{\lambda_z(k)}. \tag{III.10}$$

ξ_k and γ_k are the *a priori* and *a posteriori* SNRs, respectively.

Further mathematical transformations lead to

$$\frac{d}{d\mu}\Phi_{D_k|Y_k}(\mu) = \frac{1}{2}\ln\lambda_k - \frac{1}{2}\left(c + \sum_{r=1}^{\infty}\frac{(-v_k)^r}{r!}\frac{1}{r}\right) = \frac{1}{2}\ln\lambda_k + \frac{1}{2}\left(\ln v_k + \int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right). \tag{III.11}$$

The integral in (III.11) is known as the exponential integral of v_k, and can be efficiently calculated. Substitution of (III.11) into (III.5), using (III.10) and (III.3) gives the desired amplitude estimator:

$$\widehat{A}_k = \frac{\xi_k}{1+\xi_k}\exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right\} R_k. \tag{III.12}$$

III.2 Donoho-Johnstone SureShrink Estimator

Consider the noise model that was described previously (Section 4.1.1) with $\sigma = 1$. Let $\Theta = \{\theta_i\}_{i=0}^{d-1}$ be a d-dimensional vector of speech wavelet coefficients, and $\mathbf{w} = \{w_i\}_{i=0}^{d-1}$ be it's noisy observation with $w_i \sim N(\theta_i, 1)$. Let $\widehat{\Theta} = \widehat{\Theta}\{\mathbf{w}\}$ be a particular fixed estimator of Θ. Charles Stein [37] introduced a method for estimation of the loss $||\widehat{\Theta} - \Theta||^2$ in an unbiased fashion.

Let

$$\widehat{\Theta}(\mathbf{w}) = \mathbf{w} + \mathbf{g}(\mathbf{w}), \tag{III.13}$$

where $\mathbf{g} = \{g_i\}_{i=0}^{d-1}$ is a function from R^d into R^d. Stein showed that when $\mathbf{g}(\mathbf{w})$ is weakly differentiable, then

$$E_\Theta||\widehat{\Theta} - \Theta||^2 = d + 2\,\nabla\cdot\mathbf{g}(\mathbf{w})\} + E_\Theta\{||\mathbf{g}(\mathbf{w})||^2, \tag{III.14}$$

where $\nabla\cdot\mathbf{g} \equiv \sum_i \frac{\partial}{\partial w_i} g_i$.

Now consider the soft thresholding estimator $\widehat{\theta}_i^{(t)} = \eta_s(w_i, t)$, and apply Stein's result. According to (4.10)

$$\widehat{\theta}_i^{(t)} = \begin{cases} w_i - t \cdot \mathrm{sign}(w_i), & |w_i| > t \\ 0, & |w_i| < t \end{cases}$$

and from (III.13) we get:

$$g_i(\mathbf{w}) = \begin{cases} -t \cdot \mathrm{sign}(w_i), & |w_i| > t \\ -w_i, & |w_i| < t \end{cases} \quad \Rightarrow \quad \frac{\partial}{\partial w_i} g_i(\mathbf{w}) = \begin{cases} 0, & |w_i| > t \\ \text{-1}, & |w_i| < t \end{cases},$$

$$g_i^2(\mathbf{w}) = \begin{cases} t^2, & |w_i| > t \\ |w_i|^2, & |w_i| < t \end{cases} = \{\min(|w_i|, t)\}^2.$$

Substitution of these equations into (III.14) gives an unbiased estimate of risk:

$$E_\Theta||\widehat{\Theta}^{(t)} - \Theta||^2 = E_\Theta\{SURE(\mathbf{w}, t)\},$$

where

$$SURE(\mathbf{w}, t) = d - 2 \cdot \#\{i : |w_i| \le t\} + \sum_{i=0}^{d-1} \{\min(|w_i|, t)\}^2. \tag{III.15}$$

The basic idea of *SureShrink* estimator is usage of this estimator of risk for threshold selection:

$$t_{Sure}(\mathbf{w}) = \arg \min_{0 \le t \le \lambda_d} SURE(\mathbf{w}, t) \tag{III.16}$$

Solution of this optimization problem is very simple. Suppose we've ordered the vector $\mathbf{w}$ in the rising of $|w_i|$ manner. Between two neighboring points w_i the function $SURE(\mathbf{w}, t)$ is the monotonic rising function of t. Thus, the threshold t_{Sure} is absolute value of one of the coefficients $\{w_i\}_{i=0}^{d-1}$.

It was found that for σ close to 0 use of optimal threshold (III.16) results in higher risk than use of universal threshold λ_d (section 4.1.4). For $\sigma \gg 0$ use of t_{Sure} results in smaller risk than use of λ_d. Consequently, *SureShrink* estimator employs a hybrid scheme: if the energy of $\mathbf{w}$ is non-negligible, the estimator employs the optimal threshold t_{Sure}, and for negligible $||\mathbf{w}||_{2,d}^2$ it employs the universal threshold λ_d. The value

$$s_d^2 \equiv \frac{1}{d} \sum_{i=0}^{d-1} (w_i^2 - 1) \tag{III.17}$$

was chosen as a measure of energy of $\mathbf{w}$, and the test $s_d^2 > \eta_d/\sqrt{d}$ was chosen as an indicator for threshold selection. Here η_d is given by

$$\eta_d = (\log_2 d)^{3/2}. \tag{III.18}$$

For $\sigma \ne 1$ fine modifications are needed.

III.3 Saito Adaptive Estimator

III.3.1 The Minimum Description Length (MDL) Principle

Let $\mathbf{x} = \{x_i\}_{i=1}^{N}$ be a string of symbols drawn from a finite alphabet $\aleph$, which are independently and identically distributed with probability mass function $p(x),\ x \in \aleph$. The Shannon code has the shortest codelength on the average, and satisfies the so-called Kraft inequality:

$$\sum_{x_i \in \aleph} 2^{\mathcal{L}(x_i)} \leq 1, \tag{III.19}$$

which is the necessary and sufficient for the existence of an instantaneously decodable code, i.e., a code such that there is no codeword which is the prefix of any other codeword in the coding system. The shortest codelength on average for the whole sequence **f** becomes

$$\mathcal{L}(\mathbf{x}) = \sum_{i=0}^{N-1} \mathcal{L}(x_i) = -\sum_{i=0}^{N-1} \log_2 p(x_i). \tag{III.20}$$

Let's turn now to incoding the integer m. Suppose we don't know how large it can be. Rissanen [28] proposed that the code of such a natural number should be the binary representation of m, preceded by the code describing its length $\log m$, preceded by the code describing the length of the code for $\log m$, and so forth. This recursive strategy leads to

$$\mathcal{L}^*(m) = \log_2{}^*m + \log_2 c_0 = \log_2 m + \log_2 \log_2 m + \cdots + \log_2 c_0, \tag{III.21}$$

where the sum involves only the non-negative terms and the constant $c_0 \approx 2.865064$ (it was computed to satisfy the Kraft inequality (III.19) with equality). This can be generalized for an integer m by defining

$$\mathcal{L}^*(m) = \begin{cases} 1, & m = 0, \\ \log_2{}^*|m| + \log_2 4c_0, & \text{otherwise.} \end{cases} \tag{III.22}$$

Since we don't know the true model Υ_m generating the data $\mathbf{f}$, we'll have to use some estimate $\widehat{\Upsilon}_m$. The maximum likelihood (ML) estimate $\widehat{\Upsilon}_m$ minimize

$$\mathcal{L}(\mathbf{f}|\widehat{\Upsilon}_m, m) = -\log_2 p(\mathbf{f}|\widehat{\Upsilon}_m, m) \tag{III.23}$$

by definition (it maximizes $p(\mathbf{f}|\widehat{\Upsilon}_m, m)$ with $p(\mathbf{f}|\widehat{\Upsilon}_m, m) \leq 1$). Thus, the ML estimate is the natural choice of the model.

The ML estimate $\widehat{\Upsilon}_m$ can be described by deterministic real-valued parameters. However, it's known that for deterministic real-valued parameters $v_i \in \mathbb{R}$ the exact code generally requires infinite length of bits. Thus, in practice, some truncation must be done for transmission. Let δ be the precision and v_δ be the truncated value, i.e., $|v - v_\delta| < \delta$. Then, the number of bits required for v_δ is the sum of the codelength of its integer part $[v]$ and the number of fractional binary digits of the truncation precision δ, i.e.,

$$\mathcal{L}(v_\delta) = \mathcal{L}^*([v]) + \log_2(1/\delta). \tag{III.24}$$

The finer truncation precision we use, the smaller the term (III.23), but the larger the term $\mathcal{L}(\widehat{\Upsilon}_m|m)$ becomes. Suppose that the model Υ_m has k_m real-valued parameters, i.e., $\Upsilon_m = \{v_{m,1}, v_{m,2}, ..., v_{m,k_m}\}$. Rissanen showed that the optimal truncation precision δ^* is of order $1/\sqrt{N}$ and

$$\begin{aligned} \min_\delta \mathcal{L}(\mathbf{f}, \Upsilon_{m,\delta}, m, \delta) &= \mathcal{L}(m) + \mathcal{L}(\widehat{\Upsilon}_{m,\delta^*}|m) + \mathcal{L}(\mathbf{f}|\widehat{\Upsilon}_{m,\delta^*}, m) + O(k_m) \\ &\approx \mathcal{L}(m) + \sum_{j=1}^{k_m} \mathcal{L}^*([\widehat{v}_{m,j}]) + \frac{k_m}{2}\log_2 N + \mathcal{L}(\mathbf{f}|\widehat{\Upsilon}_m, m) + O(k_m), \end{aligned} \tag{III.25}$$

where $\widehat{\Upsilon}_m$ is the optimal non-truncated model and $\widehat{\Upsilon}_{m,\delta^*}$ is the optimally truncated version. For sufficiently large N, the last term may be omitted, and instead of minimizing the ideal codelength (4.23), Rissanen proposed to minimize

$$MDL(\mathbf{f}, \widehat{\Upsilon}_m, m) = \mathcal{L}(m) + \sum_{j=1}^{k_m} \mathcal{L}^*([\widehat{v}_{m,j}]) + \frac{k_m}{2}\log_2 N + \mathcal{L}(\mathbf{f}|\widehat{\Upsilon}_m, m). \tag{III.26}$$

The minimum of (III.26) gives the best compromise between the low complexity in the model and high likelihood of the data.

Even though the list of models Ω doesn't include the true model, the MDL method achieves the best result among the available models. It's also important to note that the MDL principle doesn't attempt to find the absolutely minimum description of the data. It always requires an available collection of the models and simply suggests picking the best model from the collection.

III.3.2 Simultaneous Noise Suppression and Signal Compression

Since we assumed the noise component is additive WGN, the probability of observing the data given all model parameters is

$$P(\mathbf{y}|\Theta_m^{(k)}, \sigma^2, k, m) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{||\mathbf{y} - W_m^T\Theta_m^{(k)}||_{2,N}^2}{2\sigma^2}\right). \tag{III.27}$$

For the ML estimate of σ^2, first consider the log likelihood of (III.27):

$$\ln p(\mathbf{y}|\Theta_m^{(k)}, \sigma^2, k, m) = -\frac{N}{2}\ln 2\pi\sigma^2 - \frac{||\mathbf{y} - W_m^T\Theta_m^{(k)}||_{2,N}^2}{2\sigma^2}. \tag{III.28}$$

Taking the derivative with respect to σ^2 and setting it to zero, we obtain

$$\widehat{\sigma}^2 = \frac{1}{N}||\mathbf{y} - W_m^T\Theta_m^{(k)}||_{2,N}^2. \tag{III.29}$$

Substitution of this into (III.28) leads to

$$\ln p(\mathbf{y}|\Theta_m^{(k)}, \widehat{\sigma}^2, k, m) = -\frac{N}{2} \ln \left(\frac{2\pi}{N} ||\mathbf{y} - W_m^T \Theta_m^{(k)}||_{2,N}^2 \right) - \frac{N}{2}. \tag{III.30}$$

Let $\mathbf{w}_m = W_m \mathbf{y}$ denote the vector of expansion coefficients of $\mathbf{y}$ in the basis B_m. Since this basis is orthonormal, we have:

$$||\mathbf{y} - W_m^T \Theta_m^{(k)}||_{2,N}^2 = ||\mathbf{w}_m - \Theta_m^{(k)}||_{2,N}^2.$$

Thus, it's easy to see that maximization of (III.30) is equivalent to minimization of $||\mathbf{w}_m - \Theta_m^{(k)}||_{2,N}^2$. Since $\Theta_m^{(k)}$ contains only k nonzero elements, the minimum of $||\mathbf{w}_m - \Theta_m^{(k)}||_{2,N}^2$ can be achieved by taking the largest k coefficients in magnitudes of $\mathbf{w}_m$ as the ML estimate of $\Theta_m^{(k)}$, i.e.,

$$\widehat{\Theta}_m^{(k)} = \eta^{(k)} \mathbf{w}_m = \eta^{(k)}(W_m \mathbf{y}),$$

where $\eta^{(k)}$ is a thresholding operation which keeps the k largest (in absolute value) elements intact and sets all other elements to zero. Finally for $\widehat{\sigma}^2$ we get:

$$\widehat{\sigma}^2 = \frac{1}{N} ||W_m \mathbf{y} - \eta^{(k)}(W_m \mathbf{y})||_{2,N}^2. \tag{III.31}$$

Let's now assume that we don't have any prior information on (k,m) so that the cost $\mathcal{L}(k,m)$ is the same for all cases, i.e., we can drop the first term of (4.26) for minimization purpose.

As for the second term, by normalizing the sequence $\mathbf{w}_m$ by $||\mathbf{y}||_{2,N}$, we can assume that the magnitude of each coefficient in $\widehat{\Theta}_m^{(k)}$ is strictly less than 1. Thus, its integer part is zero and we do not need to encode the integer part if we transmit the real-valued parameter $||\mathbf{y}||_{2,N}^2$. Now the description length $\mathcal{L}(\widehat{\Theta}_m^{(k)}, \widehat{\sigma}^2|k,m)$ becomes approximately $(\mathcal{L}^*([\widehat{\sigma}^2]) + \mathcal{L}^*([||\mathbf{y}||_{2,N}^2]) + \frac{k+2}{2}\log_2 N)$ bits. Moreover, we need to specify the indices of the non-zero coefficients, i.e., where the k non-zero elements are in vector $\widehat{\Theta}_m^{(k)}$. This requires $k\log_2 N$ bits. Thus, as a result,

$$\mathcal{L}(\widehat{\Theta}_m^{(k)}, \widehat{\sigma}^2 | k, m) = \frac{3}{2} k\log_2 N + c, \tag{III.32}$$

where c is a constant independent of (k, m).

From (III.27) we get:

$$\mathcal{L}(\mathbf{y} | \widehat{\Theta}_m^{(k)}, \widehat{\sigma}^2, k, m) = \frac{N}{2} \log_2 ||W_m \mathbf{y} - \eta^{(k)}(W_m \mathbf{y})||_{2,N}^2 + c', \tag{III.33}$$

where c' is a constant independent of (k, m). Using (III.32) and (III.33), Saito defines the approximate MDL (AMDL):

$$AMDL = \frac{3}{2} k\log_2 N + \frac{N}{2} \log_2 ||W_m \mathbf{y} - \eta^{(k)}(W_m \mathbf{y})||_{2,N}^2. \tag{III.34}$$

III.4 Cohen-Raz-Malah Adaptive Estimator

III.4.1 MDL-Based Additive Information Cost Function

Let's denote by μ the parameter vector that describes the true model Υ_m generating the unknown signal $\mathbf{f}$:

$$\mu = \left(E_m, k, \{j_{m,n}\}_{n=0}^{k-1}, \{\theta_{j_{m,n}}\}_{n=0}^{k-1}\right). \tag{III.35}$$

It was established by Rissanen [30] that the shortest codelength for encoding the data set $\{w_{m,n}\}_{n=0}^{N-1}$ using the probabilistic model $P(\{w_{m,n}\}_{n=0}^{N-1} \mid \mu)$, where μ is an unknown parameter vector, is asymptotically given by

$$\mathcal{L}\left(\{w_{m,n}\}_{n=0}^{N-1}\right) = -\log_2 P\left(\{w_{m,n}\}_{n=0}^{N-1} | \widehat{\mu}\right) + \frac{q}{2}\log_2 N, \tag{III.36}$$

where $\widehat{\mu}$ is the ML estimator of μ:

$$\widehat{\mu} = \arg\max_{\mu} P\left(\{w_{m,n}\}_{n=0}^{N-1} | \mu\right), \tag{III.37}$$

and q is the number of free real-valued parameters in the vector μ. Recalling that the expansion coefficients of the noise $\{z_{m,n}\}_{n=0}^{N-1}$ are i.i.d. $\sim \mathcal{N}(0, \sigma^2)$, it follows from Section 4.1.1 that the probability of observing the data given all model parameters is

$$P(\mathbf{y}|\mu) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{n=0}^{k-1}(w_{m,j_{m,n}} - \theta_{m,j_{m,n}})^2 + \sum_{n=k}^{N-1} w_{m,j_{m,n}}^2\right)\right). \tag{III.38}$$

Thus, from (III.36), the codelength required to encode the observed data, assuming E_m, k and $\{j_{m,n}\}_{n=0}^{k-1}$ are given, is

$$\begin{aligned}\mathcal{L}\left(\mathbf{y} \mid E_m, k, \{j_{m,n}\}_{n=0}^{k-1}\right) &= -\log_2 P\left(\mathbf{y} \mid E_m, k, \{j_{m,n}\}_{n=0}^{k-1}, \{\widehat{\theta}_{m,j_{m,n}}\}_{n=0}^{k-1}\right) + \frac{k}{2}\log_2 N \\ &= \frac{1}{2\sigma^2 \ln 2}\sum_{n=k}^{N-1} w_{m,j_{m,n}}^2 + \frac{N}{2}\log_2(2\pi\sigma^2) + \frac{k}{2}\log_2 N,\end{aligned} \tag{III.39}$$

where

$$\widehat{\theta}_{m,j_{m,n}} = w_{m,j_{m,n}}, \quad 0 \le n \le k-1 \tag{III.40}$$

are the ML estimates of $\{\theta_{m,j_{m,n}}\}_{n=0}^{k-1}$.

It was shown in Section 4.3.3 that the codelength for encoding k and $\{j_{m,n}\}_{n=0}^{k-1}$ is

$$\mathcal{L}\left(k, \{j_{m,n}\}_{n=0}^{k-1} | E_m\right) \approx k \log_2 N. \tag{III.41}$$

Since our goal is to obtain the shortest codelength, the optimal number of signal terms $\hat{k}$ and their optimal locations $\{\hat{j}_n\}_{n=0}^{\hat{k}-1}$ are obtained by minimizing the sum of codelengths given by (III.39) and (III.41):

$$\begin{aligned} \mathcal{L}\left(\mathbf{y} \mid E_m\right) &= \frac{1}{2\sigma^2 \ln 2} \sum_{n=k}^{N-1} w_{m,j_{m,n}}^2 + \frac{3k}{2} \log_2 N \\ &= \frac{1}{2\sigma^2 \ln 2} \left[\sum_{n=k}^{N-1} w_{m,j_{m,n}}^2 + \sum_{n=0}^{k-1} (3\sigma^2 \ln N) \right], \end{aligned} \tag{III.42}$$

where the constant terms are discarded. Clearly,

$$\sum_{n=0}^{N-1} \min\left(w_{m,n}^2, 3\sigma^2 \ln N\right) \le \sum_{n=k}^{N-1} w_{m,j_{m,n}}^2 + \sum_{n=0}^{k-1} (3\sigma^2 \ln N) \tag{III.43}$$

for all $0 \le k \le N-1$ and $\{j_{m,n}\}_{n=0}^{k-1} \subset \{0, \ldots, N-1\}$. Equality in (III.43) holds for the optimal value given by

$$\hat{k} = \#\left\{w_{m,n}^2 > 3\sigma^2 \ln N \mid 0 \le n \le N-1\right\} \tag{III.44}$$

and

$$\left\{\hat{j}_{m,n}\right\}_{n=0}^{\hat{k}-1} = \left\{n \mid w_{m,n}^2 > 3\sigma^2 \ln N, 0 \le n \le N-1\right\}. \tag{III.45}$$

Specifically, given E_m we compute the expansion coefficients of the observed data, and then $\hat{k}$ is the number of coefficients exceeding the threshold $\sigma\sqrt{3 \ln N}$ in absolute value, and $\{\hat{j}_{m,n}\}_{n=0}^{\hat{k}-1}$ are their locations (notice that $\hat{k} = 0$ implies $\hat{\mathbf{f}} \equiv 0$). Thus the codelength in (III.42) reduces to

$$\mathcal{L}\left(\mathbf{y} \mid E_m\right) = \frac{1}{2\sigma^2 \ln 2} \sum_{n=0}^{N-1} \min\left(w_{m,n}^2, 3\sigma^2 \ln N\right). \tag{III.46}$$

To encode the tree-set E_m, with the SIWPD tree we associate a 3-ary string as follows: for each node of the SIWPD tree we use 0 if its shift-index is identical to the shift-index of its child-nodes, we use 1 if its child-nodes have a different shift-index, and we use 2 if it is a terminal-node. Now, we traverse the tree from node to node, top-down from left to right, starting at the root at the top. The string for the example shown in Fig. III.1 is 0210222.

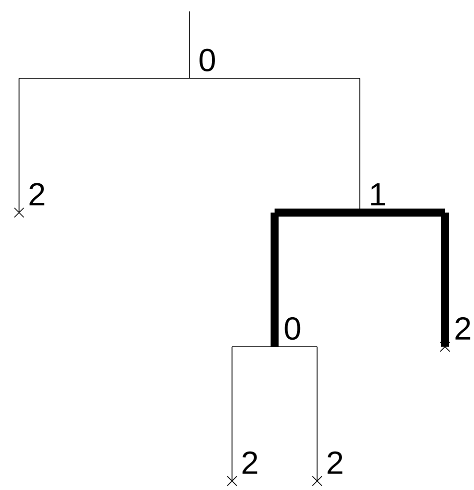


Figure III.1: Exemplifying the description of SIWPD trees by 3-ary strings. Terminal nodes are represented by 2s, and internal nodes by either 0s or 1s, depending on their expansion mode. In the present example, the string is 0210222.

A SIWPD tree, that corresponds to the basis B_m, includes $|E_m|$ terminal nodes and $|E_m| - 1$ internal nodes, where $|E_m|$ is the cardinality of E_m. Since the tree always ends with a terminal node, the last 2 in the string can be discarded, and thus we need to encode a sequence containing $(|E_m| - 1)$ 2s and $(|E_m| - 1)$ symbols from $\{0, 1\}$. The description length of such a sequence is

$$\mathcal{L}(E_m) = \log_2 \begin{pmatrix} 2|E_m| - 2 \\ |E_m| - 1 \end{pmatrix} + (|E_m| - 1) + \log_2 |E_m| , \tag{III.47}$$

where the first term is required to specify the locations of 2s in the sequence, the second term to discriminate between 0s and 1s, and the third term to encode the number of terminal terms. Applying Stirling's formula to the factorials, the description length of the tree is given by

$$\mathcal{L}\left(E_m\right) = 3|E_m| + \log_2 \frac{|E_m|}{\sqrt{|E_m| - 1}} + \frac{\alpha_1 - 4\alpha_2}{24(|E_m| - 1)\ln 2} + c', \tag{III.48}$$

where α_1, α_2 and c' are constants independent of E_m $(0 < \alpha_1, \alpha_2 < 1)$. For $|E_m| \gg 1$, the codelength can be approximated by

$$\mathcal{L}\left(E_m\right) \approx 3|E_m|, \tag{III.49}$$

where the constant terms are ignored. Adding the codelength $\mathcal{L}\left(\mathbf{y} \mid E_m\right)$ (Eq. (III.46)) to Eq. (III.49), the total description length of the observed data is given by

$$\mathcal{L}\left(\mathbf{y}\right) = \mathcal{L}\left(E_m\right) + \mathcal{L}\left(\mathbf{y} \mid E_m\right) = 3\left|E_m\right| + \frac{1}{2\sigma^2 \ln 2} \sum_{n=0}^{N-1} \min\left(w_{m,n}^2\,,\; 3\sigma^2 \ln N\right). \tag{III.50}$$

Appendix IV : Proposed Speech Denoising Algorithms

IV.1 LTD-Based Speech Denoising

IV.1.1 Estimator Type

As we saw in Section 3.3.4, different estimators can be used in wavelet-based denoising. Like WPD, LTD is an orthonormal transformation which preserve the risk (4.7) and can be organized in binary tree. Thus all the estimators that were described previously (except Cohen-Raz-Malah estimator which was developed for SIWPD), can be used in LTD-based speech denoising.

Let's denote by $\Theta = \{\Theta_{\ell,n}\}_{(\ell,n)\in E}$ the clean speech LTD coefficients, $\widehat{\Theta} = \{\widehat{\Theta}_{\ell,n}\}_{(\ell,n)\in E}$ - the estimated coefficients, $\mathbf{w} = \{\mathbf{w}_{\ell,n}\}_{(\ell,n)\in E}$ - the noisy speech LTD coefficients, and $\mathbf{z} = \{\mathbf{z}_{\ell,n}\}_{(\ell,n)\in E}$ - the noise process LTD coefficients. Here $\mathbf{w}_{\ell,n} = \{w_{\ell,n,k}\}_{k=0}^{d-1}$ are the LTD coefficients in the segment (terminal tree node), which is indexed by the pair (ℓ, n), k is the frequency-domain position index (contrary to WPD, where k is the time-domain position index), and d is the number of the coefficients in this segment ($d = 2^{(\ell+J)}$).

Tests (Table IV.1) were done under the same conditions as in Section 5.3.3, using

#	Speaker	Estimator type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
2	Female	*VisuShrink*	10	6.68	9.47	10.34	6.91	6.81
2	Female	*RiskShrink*	10	6.68	9.47	12.72	8.03	6.93
2	Female	*SureShrink*	10	6.68	9.47	**14.48**	9.06	5.9
2	Female	*Wiener*	10	6.68	9.47	12.92	8.05	8.08
2	Female	*Saito*	10	6.68	9.47	9.67	6.7	7.49
2	Male	*VisuShrink*	10	6.73	9	9.05	6.28	7.31
2	Male	*RiskShrink*	10	6.73	9	11.7	7.59	6.99
2	Male	*SureShrink*	10	6.73	9	**14.02**	8.81	6.03
2	Male	*Wiener*	10	6.73	9	12.65	7.97	7.61
2	Male	*Saito*	10	6.73	9	7.35	5.45	8.68

Table IV.1: Influence of estimator type on LTD-based denoising performance. $L = 6$.

Cosine (DCT-IV) *Packet Decomposition*-based (CPD-based) denoising with $L = 6$ and *Wickerhauser Symmetric Bell* with $\eta = 6$ (default choice of $WavBox^{©}$ software). The estimators *RiskShrink*, *VisuShrink* and *SureShrink* were used with soft thresholding.

SureShrink and Wiener estimators perform better than all the other examined estimators. Speech, enhanced by thresholding the CPD coefficients of the noisy speech, sounds better than when it is enhanced by thresholding the WPD coefficients of the noisy speech (DNS mother wavelet of the 8'th order). However, CPD-based denoising still introduces artifacts similar to WPD-based denoising: setting some CPD coefficients to zero corresponds to subtraction of basis functions from noisy signal. The reason for the better quality obtained by CPD-based denoising is that LTD is the dual of WPD with respect to time-frequency tiling: decomposing a signal from resolution level ℓ into resolution level $\ell - 1$ increases the time support of WPD bases functions, and reduce the time support of Cosine Packet (CP) basis functions. As mentioned above (Section 3.3), the $\Psi_{j,k}$ (CP basis function) is supported in time on $[a_j - \eta, a_{j+1} + \eta]$ (a_j and a_{j+1} are the segmentation points). For $J = 14$ and $L = 6$ the minimal time support of CP basis functions is $2^{(J-L)} + 2\eta = 256 + 12 = 268$ taps, which corresponds to $\ell = -L$ resolution level. For the same J and L the minimal time support of WP basis functions is $(2M-1)2^1 = 16 \cdot 2 = 32$ taps (here $M = 8$ is the order of DNS mother wavelet function and its time support is $(2M-1)$), which corresponds to $\ell = -1$ resolution level. Thus, subtraction of WP basis function with such a time support will lead to an artifact which is stronger and disturbs a listener more than an artifact produced by subtraction of CP basis function.

Despite the fact that speech enhanced by a Wiener estimator sounds noisier, its quality is much better than for all other tested estimators. We'll show in the sequel that the use of the *decision directed* approach for *a priori* SNR estimation suppresses background noise without introducing artifacts.

In order to improve the quality of speech enhanced by LTD-based estimators we should look for an appropriate cost function and verify the effect of better frequency localization on the enhanced speech quality.

IV.1.2 Cost Function and Lowest Decomposition Level

In order to check if any of the additive cost functions that were listed in Section 3.1.3 is better suited for speech denoising, they were examined in simulations. The results are presented in Table IV.2. The conditions are the same as in Section 5.3.3, the Wiener estimator and CPD with $L = 6$ were used.

We see that the differences are not significant and that the full "subsegment" decomposition gives the best denoising results. Further decomposition ($L > 6$) doesn't lead to improvement for any of above cost functions. The reason is that the smaller the segments, the more annoying are fluctuations of gains from segment to segment. Thus, $L = 6$ was chosen as an optimal value (256 taps in minimal time segment).

As mentioned in Section 5.3.5, the full "subsegment" decomposition is an attractive

#	Speaker	Cost function	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	H	10	6.06	11.51	13.33	7.5	10.19
1	Female	$\mathcal{E}$	10	6.06	11.51	13.35	7.5	10.09
1	Female	ℓ^1	10	6.06	11.51	13.34	7.5	10.09
1	Female	Full subsegment	10	6.06	11.51	**13.35**	7.5	9.95
1	Male	H	10	5.96	11.53	13.03	7.08	10.52
1	Male	$\mathcal{E}$	10	5.96	11.53	13.04	7.07	10.52
1	Male	ℓ^1	10	5.96	11.53	13.03	7.07	10.52
1	Male	Full subsegment	10	5.96	11.53	**13.11**	7.11	10.42
2	Female	H	10	6.68	9.47	12.92	8.05	8.08
2	Female	$\mathcal{E}$	10	6.68	9.47	12.97	8.08	8
2	Female	ℓ^1	10	6.68	9.47	12.98	8.07	7.96
2	Female	Full subsegment	10	6.68	9.47	**12.99**	8.05	7.85
2	Male	H	10	6.73	9	12.65	7.97	7.61
2	Male	$\mathcal{E}$	10	6.73	9	12.68	7.99	7.63
2	Male	ℓ^1	10	6.73	9	12.63	7.96	7.63
2	Male	Full subsegment	10	6.73	9	**12.73**	8.01	7.53
3	Female	H	10	6.17	11.11	13.08	7.39	10.2
3	Female	$\mathcal{E}$	10	6.17	11.11	13.15	7.43	10.09
3	Female	ℓ^1	10	6.17	11.11	13.14	7.43	10.09
3	Female	Full subsegment	10	6.17	11.11	**13.17**	7.41	10
3	Male	H	10	5.92	11.51	13.02	6.93	10.73
3	Male	$\mathcal{E}$	10	5.92	11.51	13.02	6.94	10.68
3	Male	ℓ^1	10	5.92	11.51	13.01	6.92	10.67
3	Male	Full subsegment	10	5.92	11.51	**13.02**	6.94	10.65

Table IV.2: Influence of cost function on LTD-based denoising performance. $L = 6$. H corresponds to the Shannon entropy, $\mathcal{E}$ to the log energy, and ℓ^1 to the concentration in ℓ^1 norm (Section 3.1.3).

choice: it can be represented by a fixed tree structure and can be easily implemented. Moreover, it allows simple utilization of the decision directed a priori SNR estimation: In order to utilize the latter, we have to track the a priori SNR for time segments of the same length. For the full subsegment decomposition these (time segments) are the terminal nodes and this avoids the need to keep the values of the a priori SNR for all nodes in the tree, but the terminal nodes.

IV.1.3 Window Function and Frequency Localization

As we saw in the Section 5.3.6, the better the frequency localization of basis functions, the better the performance of speech denoising algorithms. Thus, we made some simulations to verify if this conclusion also holds for LTD-based speech denoising.

As previously mentioned, the frequency uncertainty of CP basis functions $\Psi_{j,k}$ equals to that of the Fourier transform of the window function. The last one has its support on the interval $[a_j - \eta, a_{j+1} + \eta]$. Thus, the bigger the value of η, the smaller is the frequency uncertainty of CP basis functions. That is, the better their frequency localization.

In order to improve the frequency localization, we have to use the maximal allowed η. As stated in Section 3.3.2, η has to satisfy:

$$a_{j+1} - a_j \geq 2\eta > 0. \tag{IV.1}$$

Given J and L, the maximal allowed η is given by

$$\eta_{max} = 2^{J-L}/2, \tag{IV.2}$$

where 2^{J-L} is the length of the minimal segment. For our examinations $J = 14$, $L = 6$, and hence $\eta_{max} = 128$.

Tests were done (Table. IV.3) under the same conditions as in Section 5.3.3, using CP full “subsegment” decomposition and two values of η: 6 and 128.

Based on listening and on the results presented in Table IV.3, we can indeed state that the better frequency localization (bigger time support of window function), the better the quality of enhanced speech and the resulting values of SNR and LSD.

#	Speaker	η	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	6	10	6.06	11.51	13.35	7.5	9.95
1	Female	128	10	6.06	11.51	**13.67**	7.73	10.03
1	Male	6	10	5.96	11.53	13.11	7.11	10.42
1	Male	128	10	5.96	11.53	**13.3**	7.26	10.4
2	Female	6	10	6.68	9.47	12.99	8.05	7.85
2	Female	128	10	6.68	9.47	**13.24**	8.28	7.8
2	Male	6	10	6.73	9	12.73	8.01	7.53
2	Male	128	10	6.73	9	**12.89**	8.14	7.5
3	Female	6	10	6.17	11.11	13.17	7.41	10
3	Female	128	10	6.17	11.11	**13.44**	7.62	9.96
3	Male	6	10	5.92	11.51	13.02	6.94	10.65
3	Male	128	10	5.92	11.51	**13.14**	7.03	10.62

Table IV.3: Influence of CP basis functions frequency localization on LTD-based denoising performance.

IV.1.4 Utilization of Decision Directed a Priori SNR Estimation

In order to utilize the decision directed a priori SNR estimation we have to segment the speech signal into frames. As discussed earlier, time axis segmentation is inherent to LTD, thus the decision directed a priori SNR estimation can be easily applied.

The Wiener gain function, defined by Eq. (5.9), can be adapted to LTD-based denoising:

$$G_w(w_{\ell,n,k}, \sigma_{\ell,n,k}) = \frac{\widehat{\xi}_{\ell,n,k}}{\widehat{\xi}_{\ell,n,k}+1}. \tag{IV.3}$$

where $\xi_{\ell,n,k} = \frac{|\theta_{\ell,n,k}|^2}{\sigma^2_{\ell,n,k}}$ is the *a priori* SNR of the clean speech LTD coefficient $\theta_{\ell,n,k}$, $\sigma^2_{\ell,n,k} = E|z_{\ell,n,k}|^2$ is the noise variance of the k-th "frequency" LTD coefficient in the time segment indexed by the pair (ℓ, n). The estimation of the a priori SNR is done using the decision directed approach:

$$\widehat{\xi}_{\ell,n,k} = \alpha \frac{|\widehat{\theta}_{\ell,n-1,k}|^2}{E|z_{\ell,n,k}|^2} + (1-\alpha)\eta_s(\gamma_{\ell,n,k}, 1), \tag{IV.4}$$

where the *a posteriori* SNR, $\gamma_{\ell,n,k}$, is defined by

#	Speaker	Cost Function, L, α	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
3	Female	H,5,0	10	6.17	11.11	13.23	7.52	10.21
3	Female	H,5,0.9	10	6.17	11.11	14.75	8.22	9.4
3	Female	H,6,0	10	6.17	11.11	13.33	7.58	10.14
3	Female	H,6,0.9	10	6.17	11.11	**14.91**	8.31	9.4
3	Female	H,7,0	10	6.17	11.11	13.23	7.5	10.17
3	Female	H,7,0.9	10	6.17	11.11	14.72	8.14	9.39
3	Female	$\mathcal{E}$,5,0	10	6.17	11.11	13.25	7.53	10.17
3	Female	$\mathcal{E}$,5,0.9	10	6.17	11.11	15.24	8.26	9.37
3	Female	$\mathcal{E}$,6,0	10	6.17	11.11	13.38	7.59	10.06
3	Female	$\mathcal{E}$,6,0.9	10	6.17	11.11	1**5.71**	8.53	9.22
3	Female	$\mathcal{E}$,7,0	10	6.17	11.11	13.34	7.57	10.05
3	Female	$\mathcal{E}$,7,0.9	10	6.17	11.11	15.61	8.45	9.19
3	Female	ℓ^1,5,0	10	6.17	11.11	13.25	7.54	10.18
3	Female	ℓ^1,5,0.9	10	6.17	11.11	15.27	8.27	9.35
3	Female	ℓ^1,6,0	10	6.17	11.11	13.33	7.58	10.09
3	Female	ℓ^1,6,0.9	10	6.17	11.11	**15.59**	8.41	9.3
3	Female	ℓ^1,7,0	10	6.17	11.11	13.29	7.55	10.09
3	Female	ℓ^1,7,0.9	10	6.17	11.11	15.47	8.37	9.3
3	Female	FS,5,0	10	6.17	11.11	13.28	7.55	10.17
3	Female	FS,5,0.9	10	6.17	11.11	15.29	8.27	9.35
3	Female	FS,6,0	10	6.17	11.11	13.42	7.61	10.02
3	Female	FS,6,0.9	10	6.17	11.11	**15.94**	8.59	9.2
3	Female	FS,7,0	10	6.17	11.11	13.41	7.59	10.08
3	Female	FS,7,0.9	10	6.17	11.11	15.82	8.59	9.21

Table IV.4: Influence of cost function, lowest decomposition level L and smoothing parameter α on LTD-based denoising performance. FS corresponds to full "subsegment" LTD. $\eta = \eta_{max}$.

$$\gamma_{\ell,n,k} = \frac{|w_{\ell,n,k}|^2}{E|z_{\ell,n,k}|^2}. \tag{IV.5}$$

The initial condition is:

$$\widehat{\xi}_{\ell,0,k} = \alpha + (1-\alpha)\eta_s(\gamma_{\ell,0,k}, 1). \tag{IV.6}$$

If $\alpha = 0$, we return to the gain function, defined in (IV.3):

$$G_w(w_{\ell,n,k}, \sigma_{\ell,n,k}) = \frac{\eta_s(\gamma_{\ell,n,k}, 1)}{\eta_s(\gamma_{\ell,n,k}, 1) + 1} = \frac{\eta_s(|w_{\ell,n,k}|^2, \sigma^2_{\ell,n,k})}{|w_{\ell,n,k}|^2}. \tag{IV.7}$$

Utilization of the decision directed a priori SNR estimation for full "subsegment" LTD requires tracking of the a priori SNR for $k \in \{0, 1, ..., 2^{(J+\ell)} - 1\}$, $n \in \{0, 1, ..., 2^{-\ell} - 1\}$ with

$\ell = -L$ (constant). Using an adaptive best-basis selection requires the tracking of the a priori SNR for $k \in \{0, 1, ..., 2^{(J+\ell)} - 1\}$, $n \in \{0, 1, ..., 2^{-\ell} - 1\}$ and all $\ell \in \{-1, -2, ..., -L\}$.

To choose the best cost function, determine the best values of the lowest decomposition level L and the smoothing parameter α, several tests were made (Table IV.4). According to the results shown in Table IV.4 and the resulting speech quality, the preferable lowest decomposition level is $L = 6$, the best value of α is $\alpha = 0.9$, and full "subsegment" decomposition is preferred.

Increasing L leads to worsening of the frequency localization and decreases the time support of basis function. But, on the other hand, the time segments become shorter and the use of decision directed a priori SNR estimation improves the denoising performance. A further increase of L causes the gains fluctuations to be more annoying and reduces the quality of the enhanced speech.

IV.2 WPD Applied to DCT Coefficients

IV.2.1 Estimator Type

Similar to LTD-based speech denoising, all the estimators that were described previously (except Cohen-Raz-Malah estimator) can be used for speech denoising based on WPD applied to DCT coefficients.

Let's use the same notations as in Section IV.1.1 for clean, noisy and enhanced speech expansion coefficients. Tests (Table IV.5) were done under the same conditions as in Section 5.3.3, using denoising, based on WPD of DCT-I coefficient, with $L = 6$ and DNS mother wavelet of 8'th order. The estimators *RiskShrink*, *VisuShrink* and *SureShrink* were used with soft thresholding.

#	Speaker	Estimator type	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
2	Female	*VisuShrink*	10	6.68	9.47	10.73	6.77	7.16
2	Female	*RiskShrink*	10	6.68	9.47	13.17	8.16	6.67
2	Female	*SureShrink*	10	6.68	9.47	**14.36**	8.72	7.06
2	Female	*Wiener*	10	6.68	9.47	13	8.16	8.21
2	Female	*Saito*	10	6.68	9.47	9.16	6.36	7.9
2	Male	*VisuShrink*	10	6.73	9	9.01	6.09	7.92
2	Male	*RiskShrink*	10	6.73	9	11.77	7.61	6.56
2	Male	*SureShrink*	10	6.73	9	**13.68**	8.54	7.29
2	Male	*Wiener*	10	6.73	9	12.65	8.05	7.8
2	Male	*Saito*	10	6.73	9	7.55	5.51	8.01

Table IV.5: Influence of estimator type on performance of speech denoising, based on WPD applied to DCT-I coefficients.

SureShrink and Wiener estimators perform better than all the other examined estimators. Speech enhanced by this type of denoising sounds similar to the speech enhanced by LTD-based denoising. However, it still introduces artifacts similar to those obtained by WPD-based denoising.

Despite the fact that the speech enhanced by Wiener estimator sounds noisier its quality is much better than for all other tested estimators.

IV.2.2 Cost Function and Lowest Decomposition Level

The conditions in the simulations, the results of which are presented in Table IV.6, are the same as in Section 5.3.3. In the simulation the Wiener estimator and a WPD with $L = 6$ were used.

We see that the differences are not significant and full "subsegment" decomposition

#	Speaker	Cost function	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	H	10	6.06	11.51	13.45	7.59	10.29
1	Female	$\mathcal{E}$	10	6.06	11.51	13.44	7.58	10.28
1	Female	ℓ^1	10	6.06	11.51	13.45	7.59	10.24
1	Female	Full subsegment	10	6.06	11.51	**13.52**	7.61	10.21
1	Male	H	10	5.96	11.53	12.92	7.05	10.66
1	Male	$\mathcal{E}$	10	5.96	11.53	12.95	7.05	10.65
1	Male	ℓ^1	10	5.96	11.53	12.93	7.05	10.65
1	Male	Full subsegment	10	5.96	11.53	**12.98**	7.06	10.62
2	Female	H	10	6.68	9.47	13	8.16	8.21
2	Female	$\mathcal{E}$	10	6.68	9.47	13.03	8.16	8.17
2	Female	ℓ^1	10	6.68	9.47	13.04	8.16	8.19
2	Female	Full subsegment	10	6.68	9.47	**13.06**	8.18	8.14
2	Male	H	10	6.73	9	12.65	8.05	7.8
2	Male	$\mathcal{E}$	10	6.73	9	12.62	8.03	7.77
2	Male	ℓ^1	10	6.73	9	12.67	8.07	7.77
2	Male	Full subsegment	10	6.73	9	**12.7**	8.08	7.72
3	Female	H	10	6.17	11.11	13.09	7.43	10.25
3	Female	$\mathcal{E}$	10	6.17	11.11	13.08	7.43	10.26
3	Female	ℓ^1	10	6.17	11.11	13.08	7.43	10.23
3	Female	Full subsegment	10	6.17	11.11	**13.12**	7.45	10.2
3	Male	H	10	5.92	11.51	12.93	6.94	10.77
3	Male	$\mathcal{E}$	10	5.92	11.51	12.9	6.92	10.79
3	Male	ℓ^1	10	5.92	11.51	12.93	6.95	10.77
3	Male	Full subsegment	10	5.92	11.51	**12.98**	6.96	10.76

Table IV.6: Influence of cost function on performance of speech denoising, based on WPD applied to DCT-I coefficients.

gives the best denoising results. Further decomposition ($L > 6$) doesn't lead to improvement for any of above cost functions. The reason is that the smaller the segments, the more annoying are the segment to segment gains fluctuations. Thus, $L = 6$ was chosen as optimal value (256 taps in minimal time segment).

IV.2.3 Mother Wavelet and Frequency Localization

As we saw in Section 5.3.6, the better the frequency localization of basis functions, the better the performance of speech denoising algorithms. Thus, we made simulations to verify if this conclusion also holds for speech denoising based on WPD applied to DCT coefficients.

In order to improve the frequency localization for this type of joint time-frequency representation, we have to decrease the time support of the mother wavelet function. Thus, tests (Table. IV.7) were done under the same conditions as in Section 5.3.3, using full "subsegment" WPD with DNS mother wavelet of two orders (8 and 4) and the Wiener estimator. According to the results presented in Table IV.7, we can see that the better the frequency localization (smaller time support of mother wavelet function), the better are the obtained values of SNR and LSD.

IV.2.4 Utilization of Decision Directed a Priori SNR Estimation

Utilization of the decision directed a priori SNR estimation for this type of representation can be done exactly as for LTD. In order to choose an optimal cost function, lowest decomposition level L and smoothing parameter α, a number of examinations (Table IV.8) were made, in which WPD with DNS mother wavelet of 4'th order were used. According to results shown in Table IV.8 and the resulting speech quality, the preferable lowest

#	Speaker	Order	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
1	Female	8	10	6.06	11.51	13.45	7.59	10.29
1	Female	4	10	6.06	11.51	**13.5**	7.63	10.13
1	Male	8	10	5.96	11.53	12.92	7.05	10.66
1	Male	4	10	5.96	11.53	**13**	7.06	10.58
2	Female	8	10	6.68	9.47	13	8.16	8.21
2	Female	4	10	6.68	9.47	**13.04**	8.2	8.11
2	Male	8	10	6.73	9	12.65	8.05	7.8
2	Male	4	10	6.73	9	**12.7**	8.08	7.74
3	Female	8	10	6.17	11.11	13.09	7.43	10.25
3	Female	4	10	6.17	11.11	**13.13**	7.49	9.21
3	Male	8	10	5.92	11.51	12.93	6.94	10.77
3	Male	4	10	5.92	11.51	**12.95**	6.98	10.7

Table IV.7: Influence of frequency localization on speech denoising performance.

decomposition level is $L = 6$, the best value of α is $\alpha = 0.9$, and the best decomposition type is full "subsegment" decomposition.

#	Speaker	Cost Function,L, α	Input SNR	Input SEGSNR	Input LSD	Output SNR	Output SEGSNR	Output LSD
3	Female	H,5,0	10	6.17	11.11	12.91	7.38	10.27
3	Female	H,5,0.9	10	6.17	11.11	14.26	7.87	9.53
3	Female	H,6,0	10	6.17	11.11	12.96	7.41	10.27
3	Female	H,6,0.9	10	6.17	11.11	**14.41**	7.98	9.48
3	Female	H,7,0	10	6.17	11.11	12.97	7.41	10.26
3	Female	H,7,0.9	10	6.17	11.11	14.39	7.97	9.48
3	Female	$\mathcal{E}$,5,0	10	6.17	11.11	12.91	7.36	10.25
3	Female	$\mathcal{E}$,5,0.9	10	6.17	11.11	14.49	7.93	9.54
3	Female	$\mathcal{E}$,6,0	10	6.17	11.11	13.01	7.42	10.23
3	Female	$\mathcal{E}$,6,0.9	10	6.17	11.11	**14.72**	8	9.47
3	Female	$\mathcal{E}$,7,0	10	6.17	11.11	12.97	7.4	10.26
3	Female	$\mathcal{E}$,7,0.9	10	6.17	11.11	14.69	7.99	9.48
3	Female	ℓ^1,5,0	10	6.17	11.11	12.96	7.41	10.25
3	Female	ℓ^1,5,0.9	10	6.17	11.11	14.62	7.99	9.48
3	Female	ℓ^1,6,0	10	6.17	11.11	12.99	7.42	10.24
3	Female	ℓ^1,6,0.9	10	6.17	11.11	**14.64**	7.95	9.49
3	Female	ℓ^1,7,0	10	6.17	11.11	12.95	7.39	10.24
3	Female	ℓ^1,7,0.9	10	6.17	11.11	13.64	7.95	9.49
3	Female	FS,5,0	10	6.17	11.11	12.99	7.41	10.25
3	Female	FS,5,0.9	10	6.17	11.11	14.7	7.98	9.48
3	Female	FS,6,0	10	6.17	11.11	13.03	7.42	10.23
3	Female	FS,6,0.9	10	6.17	11.11	**14.83**	8.02	9.45
3	Female	FS,7,0	10	6.17	11.11	12.99	7.4	10.24
3	Female	FS,7,0.9	10	6.17	11.11	14.67	7.9	9.45

Table IV.8: Influence of cost function, lowest decomposition level L and smoothing parameter α on denoising performance. FS corresponds to full "subsegment" decomposition.

References

[1] J. Berger, R. R. Coifman and M. Goldberg, "Removing noise from music using local trigonometric bases and wavelet packets", J. Audio Eng. Soc., Vol. 42, Dec. 1994, pp. 808–818.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. on Acoust., Speech and Signal Processing, Vol. ASSP-27, Apr. 1979, pp. 113-120.

[3] I. Cohen, S. Raz and D. Malah, "Shift invariant wavelet packet bases", Proc. of the 20th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-95, Detroit, Michigan, 8–12 May 1995, pp. 1081–1084.

[4] I. Cohen, S. Raz and D. Malah, "Orthonormal shift-invariant adaptive local trigonometric decomposition", Signal Processing, Vol. 57, No. 1, Feb. 1997, pp. 43–64.

[5] I. Cohen, S. Raz and D. Malah, "Orthonormal shift-invariant wavelet packet decomposition and representation", Signal Processing, Vol. 57, No. 3, Mar. 1997, pp. 251–270. (also EE PUB No. 953, Technion - Israel Institute of Technology, Haifa, Israel, Jan. 1995).

[6] I. Cohen, S. Raz and D. Malah, "Translation-invariant denoising using the minimum description length criterion", Signal Processing, Vol. 75, 1999, pp. 201–223.

[7] I. Cohen, S. Raz and D. Malah, "Time-frequency analysis and noise suppression with shift-invariant wavelet packets", Proc. of the 11th Int. Conf. on High-Power Electromagnetics, EUROEM'98, Tel-Aviv, Israel, 14–19 June 1998.

[8] I. Cohen, S. Raz and D. Malah, "MDL-based translation-invariant denoising and robust time-frequency representations", Proc. of the 4th IEEE-SP Int. Symposium on Time-Frequency and Time-Scale Analysis, Pittsburgh, Pennsylvania, 6–9 Oct. 1998.

[9] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection", IEEE Trans. Inform. Theory, Vol. 38, No. 2, Mar. 1992, pp. 713–718.

[10] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising", in: A. Antoniadis and G. Oppenheim, ed., *Wavelet and Statistics*, Lecture Notes in Statistics, Springer-Verlag, 1995, pp. 125–150.

[11] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM Press,Philadelphia, Pennsylvania, 1992.

[12] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands", Proc. of EUROSPEECH'95, Vol. 2, 1995, pp. 1513–1516.

[13] D. L. Donoho and I. M. Johnstone, "Ideal denoising in an orthonormal basis chosen from a library of bases", Comptes Rendus Acad. Sci., Ser. I, Vol. 319, 1994, pp. 1317–1322.

[14] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage", Biometrica, Vol. 81, 1994, pp. 425–455.

[15] D. L. Donoho, "Unconditional bases are optimal bases for data compression and for statistical estimation", Applied and Computational Harmonic Analysis, Vol. 1, 1994, pp. 100–115.

[16] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage", Technical Report, Dept.of Statistics, Stanford Univ., July, 1994.

[17] D. L. Donoho, "De-noising by soft thresholding", IEEE Trans. Inform. Theory, Vol. 41, May 1995, pp. 613–627.

[18] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", IEEE Trans. on Acoust., Speech and Signal Processing, Vol. ASSP-32, No. 6, December 1984, pp. 1109–1121.

[19] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", IEEE Trans. on Acoust., Speech and Signal Processing, Vol. ASSP-33, No. 2, April 1985, pp. 443–445.

[20] B. Jawerth, W. Sweldens, "Biorthogonal smooth local trigonometric bases", the Journal of Fourier Analysis and Applications, Vol. 2, No. 2, 1995, pp. 109–133.

[21] D. Malah, R. V. Cox, A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments", IEEE Proc. of ICASSP99, Vol. 2, 1999, pp. 789–792.

[22] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet decomposition", IEEE Trans. PAMI, Vol. 11, No. 7, July 1989, pp. 674–693.

[23] R. Martin, "Spectral Subtraction Based on Minimum Statistics", in Proc. Seventh Euro. Signal Processing Conf. (EUSIPCO), 1994, pp. 1182–1185.

[24] D. B. Paul, "The spectral envelope estimation vocoder", IEEE Trans. on Acoust., Speech and Signal Processing, Vol. ASSP-29, Aug. 1981, pp. 786-794.

[25] J. G. Proakis, "Digital Communications", Third Edition, 1995, pp. 41–43.

[26] S. Raz, "Joint time-frequency representations and their applications ", Technion - Israel Institute of Technology, Haifa, Israel, lecture notes, Part 1, 1998.

[27] J. Rissanen, "Modeling by shortest data description", Automatica, Vol. 14, 1978, pp. 465–471.

[28] J. Rissanen, "A universal prior for integers and estimation by minimum description length", Ann. Statist., Vol. 11, No. 2, 1983, pp. 416–431.

[29] J. Rissanen, "Universal coding, information, prediction, and estimation", IEEE Trans. Inform. Theory, Vol. 30, No. 4, July 1984, pp. 629–636.

[30] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[31] N. Saito, *Local Feature Extraction and Its Applications Using a Library of Bases*, Ph.D. Dissertation, Yale Univ., New Haven, Dec. 1994.

[32] N. Saito and R. R. Coifman, "Local discriminant bases", in: A. F. Laine and A. M. Unser, ed., *Mathematical Imaging: Wavelet Applications in Signal and Image Processing*, Proc. SPIE, Jul. 1994, Vol. 2303.

[33] N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion", Proc. SPIE, Vol. 2242, 1994,pp. 224–235.

[34] N. Saito and R. R. Coifman, "On local orthonormal bases for classification and regression", Proc. of the 20th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-95, Detroit, Michigan, 8–12 May 1995, pp. 1529–1532.

[35] S. Saito, "Speech science and technology", 1992, pp. 5–9.

[36] I. Y. Soon, S. N. Koh, C. K. Yeo, "Noisy speech enhancement using discrete cosine transform", Speech Communication 24, 1998, pp. 249–257.

[37] C. Stein, "Estimation of the mean of a multivariate normal distribution", Annals of Statistics 9, No. 6, 1981, pp. 1135–1151.

[38] C. Taswell, "Near-best basis selection algorithms with non-additive information cost functions", Proc. of the 2nd IEEE-SP Int. Symposium on Time-Frequency and Time-Scale Analysis, Philadelphia, PA, 25–28 Oct. 1994, pp. 13–16.

[39] C. Taswell, "WavBox 4: A software toolbox for wavelet transforms and adaptive wavelet packet decompositions", in: A. Antoniadis and G. Oppenheim, ed., *Wavelet and Statistics*, Lecture Notes in Statistics, Springer-Verlag, 1995, pp. 361–376.

[40] J. M. Tribolet, R. E. Crochiere "Frequency domain coding of speech", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP–27, No. 5, Oct. 1979, pp. 512–530.

[41] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", IEEE Trans. Speech Audio Processing, Vol. 7, No. 2, March 1999, pp. 126–137.

[42] M. V. Wickerhauser, "Inria lectures on wavelet packet algorithms", tech. rep., INRIA, Roquencourt, France, 1991, minicourse lecture notes.

[43] M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, AK Peters, Ltd, Wellesley, Massachusetts, 1994.

Hebrew Abstract

The problem of enhancing speech degraded by uncorrelated additive noise, when the noisy speech alone is available, has received much attention. This is due to a variety of potential applications speech enhancement possesses. Furthermore, technologies enabling the implementation of such intricate algorithms are now available. The main purpose of denoising techniques is to improve the quality and comprehension of speech. It's also useful to enhance the speech prior to the implementation of techniques such as coding and recognition. Unfortunately, while existing speech denoising algorithms appear to improve the quality of speech, they typically do not improve its comprehension.

Wavelet bases are widely used for estimating signals embedded in noise. While traditional methods often remove noise by low-pass filtering, thus blurring the sharp features in the signal, wavelet-based methods show good performance for a wide diversity of signals. The *wavelet shrinkage* method, developed by Donoho and Johnstone [17], uses a fixed transform of the noisy data into the wavelet-domain, applies soft or hard thresholding to the resulting coefficients, and subsequently transforms the modified wavelet-domain coefficients back into the original space. It was recognized that the success of such a denoising scheme is determined by the extent to which the transform compresses the unknown signal into few significant coefficients [15]. Given a library of bases and a noisy measurement, researchers proposed several different approaches to select a "best" basis and a threshold

value, leading to the best signal estimate [13].

Saito [33] proposed to use an information-theoretic criterion, called the *Minimum Description Length* (MDL) principle [30], for noise removal. He claimed that the MDL criterion gives the best compromise between the estimation fidelity (noise suppression) and the efficiency of representation (signal compression).

It has been observed [10, 1, 33] that denoising with the conventional wavelet transform and wavelet packet decomposition (WPD) may exhibit visual artifacts, such as pseudo-Gibbs phenomena in the neighborhood of discontinuities. These artifacts were related to the lack of *shift-invariance*, and proposed to reduce them by averaging over different translations: Applying a range of shifts to the noisy data, denoising the shifted versions with the wavelet transform, then unshifting and averaging the denoised data. This procedure, termed *Cycle-Spinning* [10], generally yields better visual performance on smooth parts of the signal.

Cohen, Raz and Malah [5] presented an extension of WPD into a Shift-Invariant WPD (SIWPD). Moreover, they reformulated the MDL principle as an additive information cost function [8] and presented an adaptive translation-invariant denoising algorithm.

The main purpose of the thesis was to modify and improve existing denoising algorithms and to study the consequences of shift-invariance on speech enhancement and the resulting artifacts. First, we implemented the state of the art speech denoising algorithms and wavelet-based denoising algorithms. These algorithms served as benchmarks. We then developed the speech denoising algorithms, based on WPD and Local Trigonometric Decomposition (LTD), which utilize the decision directed approach to a priori SNR estimation.

We have shown that artifacts, introduced by wavelet-based denoising algorithms [14, 16, 10, 33, 5], applied to speech enhancement, can be particularly suppressed by increasing temporal support of the basis functions. Moreover, improvement in frequency localization of the basis functions improves the speech denoising performance. It also has been shown that shift-invariance achieved by Shift-Invariant Wavelet Packet Decomposition (SIWPD) does not contribute to artifacts suppression and does not guarantee an improved denoising performance.

Denoising based on the presumption of prior knowledge of the squared spectral amplitude of the noise is referred to as *ideal* denoising. Quite expectedly, simulations confirm that such ideal denoising attains higher SNR than the practical one. We have proven that ideal speech denoising, based on some real-valued transform, achieve an exact phase reconstruction of the clean speech signal (expressed via the sign of the real-valued transform coefficients). This is an intrinsic advantage of real-valued transforms compared to DFT-based denoising. The results show that the exact phase reconstruction associated with real-valued transforms leads to global SNR improvement by $0.69 \div 1.12$ [dB] while comparing WPD-based vs. DFT-based ideal denoising.

We have compared the proposed speech denoising algorithms to the state of the art speech denoising algorithms [18, 19]. Simulation results indicate that, for each of the tested speech signals, the DFT-based Wiener estimator attains the highest global SNR, segmental SNR and LSD. The quality of the enhanced speech is similar for all the algorithms. The notable difference is the level and type of a residual background noise. All of the algorithms, Ephraim-Malah being the exception, introduce a colored background noise, that was found to be disturbing the listener. The DFT-based Wiener estimator is characterized by the lowest level of the residual noise, and is superior to the proposed algorithms. The Ephraim-Malah algorithm is characterized by a higher level of background noise then DFT and WPD-based Wiener estimator, but, advantageously, the background noise is almost white.

Despite the advantages of WPD and LTD-based algorithms under ideal denoising conditions, in practice (i.e., with an estimated noise variance) the DFT-based denoising algorithms are found to be better. The reasons are:

1) Given the noisy observations, we can't know the exact values of the noise squared spectral components. Hence, using only the estimated averages of the noise squared spectral components we can't exactly reconstruct the clean speech phase.

2) It is shown in Appendix I, that if the additive noise is white and Gaussian, the variance of its squared spectral components, obtained by real-valued transform, is twice (except for the DC coefficient) the variance of the noise squared spectral components, obtained by the DFT. This leads to higher deviations of noise squared spectral amplitude from its estimated value, and subsequently to higher frame to frame gains fluctuations (segment to segment gains fluctuations for LTD-based denoising) thus reducing the resulting global and segmental SNR. The frame to frame gains fluctuations cause the residual background noise to be colored.

Despite the fact that the speech denoising algorithms proposed herein do not possess clear advantage over the DFT-based algorithms, they may have merit in a wider sense. For example, WPD-based denoising can be easily incorporated into a WPD-based speech coding system. Also, LTD can be used as a time-segmentation tool. Thus, the LTD-based denoising algorithm can be conveniently implemented in speech analysis systems, which require adaptive time-segmentation.

The organization of this thesis is as follows. In the next chapter we review the state of the art speech denoising algorithms and the so-called "*decision directed*" approach to *a* *priori* SNR estimation, that was introduced by Ephraim and Malah in [18]. In Chapter 3 we review the basics of joint time frequency representations: Wavelet packet analysis and best-basis expansion, the extension of wavelet packet bases for obtaining shift-invariance, and local trigonometric bases. In Chapter 4 we review different wavelet-based denoising algorithms, including the so-called "*translation-invariant*" denoising algorithm of Coifman and Donoho, and the Cohen-Raz-Malah shift-invariant denoising algorithm, based on shift-invariant WPD.

The main contribution of this thesis begins in Chapter 5, where we present several speech denoising algorithms, based on WPD, Cosine Packet Decomposition and WPD applied to DCT-I coefficients. We utilize the decision directed a priori SNR estimation for each of the mentioned joint time-frequency representations. Importance of shift-invariance, time support and frequency localization are discussed. In Chapter 6 we introduce a comparative performance analysis of different speech denoising algorithms, and present some interesting conclusions corresponding a comparison of DFT-based and real-valued transform-based denoising. Required proofs are given in the Appendices.

Finally, in Chapter 7 we conclude with a summary and discussion on future research directions.