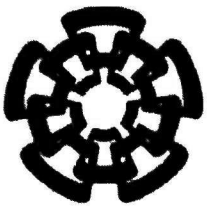
xx(97956.1)



CINVESTAV

Centro de Investigación y de Estudios Avanzados del IPN

Unidad Guadalajara de Ingeniería Avanzada

Excelencia en Investigación, Educación y Desarrollo Tecnológico

Codificación Perceptual de Audio Banda Extendida Usando Paquetes de Ondeletas

TESIS QUE PRESENTA

MIGUEL ÁNGEL ALONSO ARÉVALO

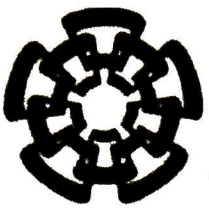
PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS

EN LA ESPECIALIDAD DE

INGENIERÍA ELÉCTRICA

Guadalajara, Jal., Junio de 2001



CINVESTAV

Centro de Investigación y de Estudios Avanzados del IPN
Unidad Guadalajara de Ingeniería Avanzada

Excelencia en Investigación, Educación y Desarrollo Tecnológico

Perceptual Wide-band Speech Audio Coding Using Wavelet Packets

THESIS PRESENTED BY

MIGUEL ÁNGEL ALONSO ARÉVALO

IN PARTIAL FULFILLMENT TO OBTAIN THE DEGREE OF

MASTER OF SCIENCE

IN THE FIELD OF

ELECTRICAL ENGINEERING

Guadalajara, Jal., June 2001

CODIFICACIÓN PERCEPTUAL DE AUDIO BANDA EXTENDIDA USANDO PAQUETES DE ONDELETAS

Tesis de Maestría en Ciencias
Ingeniería Eléctrica

Por:

Miguel Ángel Alonso Arévalo

Becario de CONACYT, expediente no.**129114**

Director de Tesis:

Dr. Manuel Edgardo Guzmán Rentería

CINVESTAV del IPN Unidad Guadalajara, Junio de 2001

PERCEPTUAL WIDE-BAND SPEECH AUDIO CODING USING WAVELET PACKETS

Master´s Thesis
in Electrical Engineering

by:

Miguel Ángel Alonso Arévalo

CONACYT´s scholarship holder, expedient number **129114**

Thesis Director:

Dr. Manuel Edgardo Guzmán Rentería

CINVESTAV del IPN Unidad Guadalajara, June 2001

Agradecimientos

A mis padres Ana Miriam y Miguel por su apoyo, por la formación que me han dado y por haberme brindado la oportunidad de recibir una buena educación.

A mis hermanos, Ani, Beatriz, Paul, a mis tías Alby, Bety, Elena y a mis abuelos, por su apoyo y motivación a seguir adelante.

A los Tavos, Eliús, Manueles, Albertos, Pillos, Puigs y Paulo. A todos ellos por su cariño, confianza y amistad incondicional desde hace tantos años.

A mis profesores el Dr. Arturo Veloz Guerrero y el Dr. Jean-Marc Boucher, por las enseñanzas transmitidas y su apoyo durante mi estancia en Francia.

A mis amigos Carlos, Lola, Andrés, Marta, Felipe, Txiki, Anabel, Esteban y el resto de los *troueros*, por su amistad y confianza.

A mis amigos Aida, Jacobo, Enrique, Raúl, Eduardo, Luis Alberto, Vladimir, Alfredo, Elizabeth, Vicente, Pablo, Héctor, Víctor e Iván, por todos esos momentos agradables durante la maestría en CINVESTAV.

Al CONACYT y a todo el pueblo de México, por apoyarme económicamente durante dos años y hacer posibles mis estudios de posgrado.

Al EGIDE, por apoyarme económicamente durante mi estancia en Francia.

A mis demás compañeros y maestros del CINVESTAV y de la ENST de Bretaña, por hacer de los momentos complicados y el trabajo, a veces difícil, algo más dinámico y agradable.

Resumen

En esta tesis se presenta un codificador de audio en el rango de voz banda extendida (50–7,000 Hz) basado en la transformada en paquetes de ondeletas. Dicha transformada permite aproximar eficientemente la descomposición en bandas críticas que realiza el sistema auditivo, tanto en el dominio del tiempo como de la frecuencia. Como resultado de este fenómeno, se hace uso de las propiedades de enmascaramiento espectral del sistema auditivo humano para poder disminuir la razón promedio de bits del codificador, ocultando perceptualmente el error de cuantización.

El presente trabajo comienza con un breve repaso de las técnicas de análisis de señales por medio de transformada y descomposición en subbandas. Además, algunas técnicas de asignación de bits son revisadas. A continuación, los principios básicos del análisis por medio ondeletas son presentados. Esto lleva a la relación entre bancos de filtros y ondeletas, particularmente a los bancos de filtros paraunitarios de dos bandas y su generalización a estructuras de árbol.

A continuación, se presenta un análisis del sistema auditivo humano. Primeramente desde el punto de vista fisiológico, para después pasar a un estudio de las propiedades de enmascaramiento espectral del oído. Se elige el modelo psicoacústico de MPEG para su implementación en el presente trabajo.

Se presenta la integración del codificador y de sus componentes, además de su implementación. Las pruebas realizadas demuestran que el codificador propuesto logra una calidad casi transparente para señales de audio genéricas a una razón promedio de bits de 56.1 kbit/s. A continuación se muestran los resultados de las pruebas objetivas y subjetivas del codificador desarrollado. Finalmente se presentan las conclusiones de la tesis y posibles mejoras futuras para el codificador.

Résumé

Cette thése présente une méthode de codage des signaux audio générique dans la bande 50–7,000 Hz. Cette méthode est basée sur la transformée en paquets d'ondelettes. La technique proposée dans ce travail permet d'approcher efficacement la décomposition qui fait l'oreille en bandes critiques dans les domaines temps-fréquence. On profite des propriétés de masquage fréquentiel de l'oreille pour reduire le débit moyen du codeur, tant que l'erreur de quantification reste perceptuellement caché.

Ce travail commence par une étude des techniques d'analyse en sous-bandes et par transformées, en plus on déduit des régles d'allocation de bits nécessaire pour le codage. Ensuite, des fondements de l'analyse par ondelettes sont présentés. Après, on montre la relation entre les ondelettes et les bancs de filtres. Une attention particulière est accordée au banc de filtres para-unitaires à deux canaux et á la structure d'arbre qu'on peut déduire á partir d'eux.

Le system auditif humain est analysé. Au début, un point de vue physiologique du system auditif est donné. Après, un étude des propriétés du masquage fréquentiel est présenté. Le modèle psychoacoustique de MPEG est choisi pour être implémenté dans le codeur proposé.

La mise en œuvre du codeur et ses composantes est adressé. Les résultats des tests realisés montrent que le système de codage proposé atteint une qualité proche de la transparence avec un débit moyen de 56.1 kbit/s. Une évaluation objective et subjective du sytème proposé est présentée. Finalement, on présente les conclusions et des possibles améliorations dans l'avenir.

Abstract

This thesis presents a wide-band speech audio coding method based on the wavelet packet transform. These techniques permit us to efficiently approximate the auditory critical band decomposition in the time and the frequency domains. As a result, we make use of the spectral masking properties of the human hearing mechanism to decrease the average bit rate of the encoder, while perceptually hiding the quantization error.

This work begins with a brief review of the subband and transform signal analysis techniques. Bit allocation rules, necessary for coding, are also addressed. Next, the wavelets fundamentals are presented. Then, the relationship between wavelets and filter banks is discussed. Particular attention is paid to two-channel paraunitary filter banks and their generalization to tree structures.

The Human auditory system is analyzed. Firstable, a physiological point of view of the auditory system is provided. Then, a study of spectral masking and its properties is presented. The MPEG psychoacoustical model is chosen for implementation in the present work.

The integration of the coder components and its implementation are addressed. Experiments show that the proposed coding system achieves nearly transparent quality for generic audio signals in the range of 50–7,000 Hz, at an average bit rate of 56.1 kbits/s. Objective and subjective quality assessments for the proposed coder are provided. Finally, the conclusions, some future considerations and possible improvements are given.

Glossary of abbreviations

ADPCM	adaptive differential PCM
ASPEC	adaptive spectral perceptual entropy coding
DCC	digital compact cassette
CELP	code exited linear predictor
CWT	continuous wavelet transform
DAB	digital audio broadcasting
DCT	discrete cosine transform
DPWT	discrete parameters wavelet transform
DTWT	discrete time wavelet transform
DWT	discrete wavelet transform
FFT	fast Fourier transform
HAS	human auditory system
ISDN	integrated services digital network
ITU	international telecommunication union
JND	just noticeable distortion
LPC	linear prediction coding
MDCT	modified discrete cosine transform
MOS	mean opinion score
MPEG	moving pictures experts group
MUSICAM	masking-pattern universal sub-band integrated coding and multiplexing
PCM	pulse code modulation
PDF	probability density function
PSD	power spectrum density
QMF	quadrature mirror filter
SFM	spectral flatness measure
SPL	sound pressure level
STFT	short time Fourier transform
VQ	vector quantization
WPT	wavelet packet transform
WSS	wide sense stationary
WT	wavelet transform
WWW	world wide web

This thesis was prepared with LaTeX 2_ε under Debian GNU/Linux 2.2.

Contents

Resumen iii

Résumé v

Abstract vii

Glossary of abbreviations ix

List of Figures xiii

List of tables xviii

1 Introduction 1

1.1 Scope of the thesis 2

1.2 Proposed Coding System . 3

1.3 Thesis Overview 4

2 Background 7

2.1 Speech Compression 7

2.1.1 Waveform coding 9

2.1.2 Parametric coding 9

2.2 Audio Compression 9

2.3 State of the art 11

2.3.1 Toll quality 12

2.3.2 Wide-band speech audio 13

2.3.3 High quality coding 15

2.4 Subband Coding 18

2.5 Transform coding 21

2.5.1 Optimal Orthogonal Transform 23
2.6 Quantization 24
2.6.1 Additive noise model 28
2.7 Optimum bit allocation 29
2.8 Huffman Coding 31

3 Wavelets Fundamentals 33
3.1 Fourier Analysis 34
3.1.1 Windowed Fourier Transform 35
3.2 Continuous Wavelet transform . 36
3.3 Discrete Parameter Wavelet transform 42
3.4 Multiresolution Analysis 46
3.4.1 Relationship Between Multiresolution Analysis and the DPWT 48
3.4.2 The idea of multiresolution 51
3.5 Perfect Reconstruction Paraunitary Filter Banks 53
3.5.1 Polyphase representation 54
3.5.2 Two-band FIR Paraunitary Filter Banks 56
3.5.3 Tree Structured Filter Banks 60
3.6 Time-Frequency Auditory Mapping 62
3.6.1 Orthonormal Wavelet Packet Transform 62
3.6.2 Filter Bank Delay 63
3.6.3 Human Auditory System Modeling 65

4 Auditory System Modeling 71
4.1 General Aspects of Hearing 72
4.1.1 Anatomy of the ear 73
4.2 Cochlear Mechanics . 74
4.2.1 Tone Behavior 77
4.3 Masking 79
4.3.1 Absolute Threshold of Hearing 81
4.3.2 Masking of Pure Tones by Broad-Band White Noise 81
4.3.3 Masking of Pure Tones by a Narrow-Band White Noise 85
4.3.4 Masking of Pure Tones by Tones 89
4.3.5 Masking of Narrow-Band White Noise 90
4.3.6 Nonsimultaneous Masking 92

4.4 Perceptual Entropy 92
4.5 Masking Threshold Implementation 95

5 Perceptual Coding of Wide-Band Speech Audio 101
5.1 Description of the Encoder 101
5.1.1 Optimal Bit Allocation of Transform Coefficients 103
5.1.2 Masking Threshold Quantization 106
5.2 Description of the Decoder 106
5.3 Coder Results 107

6 Coder Performance Evaluation 109
6.1 Perceptual Objective Measures 109
6.1.1 Segmental Signal-to-Noise Ratio 111
6.1.2 Itakura-Saito Distortion 111
6.2 Subjective Measures 112
6.2.1 Mean Opinion Score 112
6.3 Evaluation of the Proposed System 113

7 Conclusions 117
7.1 Conclusions 117
7.2 Possible Extensions and Future Research 118

A High Resolution Hypothesis 121

B Bit Allocation 125

C Continous Wavelet Transform 127

D Noble Identities 129

E ISO/IEC MPEG Psychoacoustic Model 131

BIBLIOGRAPHY 136

List of Figures

1.1 *Proposed wide-band speech audio encoder and decoder.* 4

2.1 *Estimated relationship between speech coders bit rate and coding technique.* 8

2.2 *Generic perceptual audio encoder.* 10

2.3 *Code excited linear predictor (CELP) scheme. The goal is to minimize* $Y_k(n)$ *by selecting the best codebook entry.* 12

2.4 *Encoding principle of the ITU-T G.729 CS-ACELP encoder.* 14

2.5 *Structure of the ITU-T G.722 audio coder.* 15

2.6 *Block diagram of MPEG-1 Layer I and Layer II encoder.* 16

2.7 *Block diagram of MPEG-1 Layer III encoder.* 17

2.8 *Magnitude response of a uniform M-band filter bank.* 18

2.9 *Uniform M-band maximally decimated analysis-synthesis filter bank.* 20

2.10 *Structure of the Transform Coding principle.* 22

2.11 *Block diagram of a quantizer.* 25

2.12 *Midtread quantizer characteristic.* 26

2.13 *Additive noise model of the quantizer.* 28

3.1 *(a)Time-frequency boxes (Heisenberg rectangles) representing the energy spread of two Gabor atoms. (b) STFT tiling of the time-frequency plane.* 37

3.2 *Division of the frequency domain (a) for the STFT and (b) for the WT.* 39

3.3 *(a) Time-frequency boxes of two wavelets* $\psi_{u,s}$ *and* ψ_{u_0,s_0}*. When the scale* s *decreases the time support is reduced, but the frequency spread increases and covers an interval that is shifted towards the high frequencies. (b) Wavelet transform tiling of the time-frequency plane.* 41

3.4 *The dyadic sampling grid in the time-scale plane. Each dot corresponds to a wavelet basis function $\psi_{m,n}(t)$.* 45
3.5 *Mallat's multiresolution analysis scheme.* 46
3.6 *Reconstruction from subband decomposition.* 47
3.7 *The first stage decomposition of the MRA.* 48
3.8 The mth stage decomposition of the MRA. 50
3.9 *Spectrum of subspaces.* 52
3.10 *Decomposition of the frequency spectrum into succesive subspaces (octave bands). There is a scaling factor for $V_j(\omega)$ and $W_j(\omega)$ by $2^{j/2}$, not depicted, to make subspaces of unit norm.* 53
3.11 *M channel paraunitary transform.* 54
3.12 *A two-band analysis/synthesis filter bank.* 56
3.13 *Two-band filter bank polyphase decomposition stages.* 58
3.14 *Four-band tree structured filter bank (a) Analysis (b) Synthesis.* . 61
3.15 *Four-band filter bank. Parallel version.* 61
3.16 *Three stage wavelet decomposition. (a) Filter tree structure (Mallat's decomposition scheme). (b) Time-frequency tiling. (c) Idealized magnitud response of the filter bank.* 63
3.17 *Three-stage wavelet packet decomposition. (a) Filter tree structure. (b) Time-frequency tiling. (c) Idealized magnitud response of the filter bank.* 64
3.18 *Delay contributions of one tree branch at stage j.* 64
3.19 *WPT tree structure to approximate the critical band decomposition of the human auditory system.* 66
3.20 *Critical band rate approximation. Comparison between the DWPT and the human auditory system model.* 67
3.21 *Critical bandwidth approximation. Comparison between the DWPT and the human auditory system model.* 68
3.22 *Magnitude responses of the chosen wavelet packet transform for a) the mother wavelet DB5 (10 coefficients) b) the mother wavelet DB10 (20 coefficients).* 69
3.23 *Time-frequency tiling of the 21-critical band decomposition using the discrete wavelet packet transform. In the figure, j stands for decomposition stage, and k for the critical band number.* 70

4.1 *Anatomy of the ear. The ear is mainly composed of three parts: the outer ear, the middle ear and the inner ear.* 73

4.2 *Anatomy of the cochlea. This image is a magnification view of a transversal cut of one turn of the coil.* 75

4.3 *The basilar membrane. The vibration of the stapes produces vibratory waves in the fluid environment of the cochlea. The figure presents an uncoiled illustration of the cochlea and a wave traveling.* 76

4.4 *Envelope of the traveling wave appearing on the basilar membrane for* f = 19000, 4900, 990 *and* 190 *Hz (from left to right).* 78

4.5 *Envelope of the traveling wave appearing on the basilar membrane versus input frequency for* x =*4.675, 12.9, 22.1 and 29.9 mm (from left to right).* 79

4.6 *Envelope of the traveling wave appearing on the basilar membrane for a combination of* f = 19000, 4900, 990 *and* 190 *Hz tones.* 80

4.7 *Idealized absolute threshold of hearing.* 82

4.8 *Detection of pure sine waves in wide-band white noise. The dashed curve represents the absolute threshold of hearing.* 83

4.9 *Critical bandwidth vs. center frequency.* 85

4.10 *Critical band rate vs. center frequency.* 86

4.11 *Masking of pure tones by narrow-band white noise. The masker noise is centered at 1000 Hz.* 86

4.12 *Masking threshold vs. frequency (top) and critical band rate (bottom) for 150, 1000, 4000 and 12000 Hz center frequencies (from left to right).* 88

4.13 *Masking of pure tones by tones.* 89

4.14 *Behaviour of temporal masking. Pre-masking occurs prior to masker onset and lasts only a few milliseconds; post-masking may persist up to 200 ms after masker removal.* 93

4.15 *Masking pattern (dotted line) for a narrow-band white noise signal (solid line) located in the* $14^{th.}$ *critical band (2320–2700 Hz). The average power of the noise signal is 65 dB.* 97

4.16 *Masking of a tone by a narrow-band noise. The noise signal is located in the* $14^{th.}$ *critical band (2320–2700 Hz), its average power is 65 dB. The tone signal has a carrier frequency of 4000 Hz and a power of 32 dB. The masking pattern (dotted line) generated by the noise signal completely masks the tone signal.* 98

4.17 *Masking of a tone by another tone. The acoustic signal is formed by three tones of frequency 1500, 3500 and 4750 Hz (from left to right) with a power of 70, 85 and 30 dB respectively. In this case the upper tone is barely perceptible.* 99

4.18 *Power spectrum of a voiced frame of female speech. The masking threshold (dotted line) shows that almost half of the signal is rendered inaudible due to masking.* 99

4.19 *Power spectrum of a voiced frame of female speech. The masking threshold (dotted line) shows that almost half of the signal is rendered inaudible due to masking.* 100

4.20 *Power spectrum of an audio (music) signal. The spectral nature of audio signals is more complex than the nature of speech, but still a considerable part of the signal is rendered partially or totally inaudible due to masking.* 100

5.1 *Block diagram of the encoder.* 102

5.2 *Schematic representation of simultaneous masking. Ideally, we can equate the SNR to the signal-to-mask ratio (SMR) and keep m as small as possible.* 103

5.3 *Histogram presenting the Perceptual Entropy values of several wide-band speech audio sources. This result suggests a lower bound of about 2.1 bits/sample for transparent audio coding.* 104

5.4 *Histogram presenting the number of bits required to encode several wide-band speech audio sources. The upper bound, represented by the vertical line, indicates that 99.4% of the coefficients can be encoded using a maximum of 6-bits. Additionally, 92.3% of the coefficients can be encoded using a maximum of 3-bits.* 105

5.5 *Block diagram of the decoder.* 107

D.1 *Noble identities in multirate systems, for subsamplers (top) and downsamplers (bottom).* 129

List of Tables

3.1 *Two-band filter bank relations for perfect reconstruction.* 60

4.1 *Idealized critical band distribution.* 84
4.2 *Typical values of the spreading function parameters.* 89
4.3 *Meaning of the parameters presented in Equation 4.20.* 95

5.1 *Vector quantization arrangement of the masking threshold coefficients* δ_i. 106
5.2 *List of source material and their coded bit rate. The source material has been band limited to the range 50–7,000 Hz, sampled at 16 kHz and quantized using 16-bit linear PCM.* 108

6.1 *Five-point MOS scale.* 113
6.2 *Objective measures for the proposed coding system.* 114
6.3 *Subjective listening test result: transparency test.* 114

E.1 *Calculation partition table. This table is valid only at a sampling rate of 16 kHz.* 135

Chapter 1

Introduction

Perceptual audio coding algorithms are lossy compression schemes that minimize the number of bits required to represent audio signals while trying to maintain transparent quality. This is accomplished through two processes known as *irrelevancy reduction* and *redundancy removal.* Irrelevancy reduction is achieved by shaping the coding distortion (quantization noise) such that it cannot be perceived. The main idea is to perform a time-frequency distribution of the quantization noise such that it is not perceivable by the human auditory system. The method employed in this thesis for quantization noise shaping is involved with time-varying, signal-dependent allocation of bits for quantization of signal components.

Perceptually relevant signal components are accurately represented using a larger number of bits; perceptually unimportant or imperceptible components, in contrast, receive very few bits and in some cases are altogether discarded. For example, frequency components that fall below the threshold of hearing can be safely discarded. Nowadays, the irrelevancy reduction (or distortion control) step is essential to the success of perceptual coding schemes. Redundancy removal is also of vital importance to the perceptual audio coder. Whereas irrelevancy reduction exploits the properties of auditory perception, redundancy removal identifies and removes statistical redundancies. In the proposed coder irrelevancy reduction is applied after redundancy removal.

Perceptual coders employ a number of signal-processing tools in pursuit of both irrelevancy reduction and redundancy removal. Of particular importance are filter banks, used to decompose the audio signal into a set of time-frequency components. Given such a set, it is possible to discriminate between the perceptually relevant and irrelevant elements. Then, several quantization techniques can be applied to represent the relevant

time frequency components with as little precision as possible, without introducing perceptible distortion. Choice of an innappropriate filter bank can result in lower output quality or in a need for higher bit rate in the audio coder stream. Ideally, the properties of the filterbank must be matched to the characteristics of the input signal. For example, harmonic sounds produced by a bagpipe or spectrally complex signals such as the sound produced by a harpsicord demand for a filter bank with fine frequency resolution and coarse time resolution. In contrast, the sounds containing sharp attacks or abrupt transients such as those produced by percussive instruments like the castanets, the drums, or the triangle demand for a filter bank with a good time resolution.

Perceptual coders rely upon models of human auditory perception in order to discriminate between relevant and irrelevant signal components extracted by the filter bank. Most models seek to exploit masking phenomena, both simultaneous and nonsimultaneous. Both types of masking describe a process in which the presence of a sound hides the presence of a weaker sound. Essentially, the task of auditory models in perceptual audio coding is to estimate the amount of masking power present in the signal and then to use this information to determine the way bits should be allocated to the quantized representation of the time frequency components generated by the analysis filter bank. The idea is to allocate bits such that the quantization noise falls below the threshold of audibility. The threshold of audibility is estimated in terms of the masking power in the signal at a given time instant. Perceptual models, therefore, attempt to model the signal processing that takes place in the cochlea (inner ear).

1.1 Scope of the thesis

Audio coding usually refers to the compression of high fidelity audio signals, i.e., with 15- or 20-kHz bandwidth for consumer hi-fi, professional audio including motion picture for HDTV audio, and various multimedia systems. Wide-band speech coding refers to the compression of signals having a 50-7000 Hz bandwidth, usually for teleconferencing applications where an increased intelligibility of speech is required [28]. For the present work the term *wide-band speech audio* was wedged to refer to generic audio signals in the frequency band 50-7000 Hz sampled at 16 kHz. Traditionally, most of the effort in sound compression has been focused on the usual telephone bandwidth of roughly 3.1 kHz (300–3400 Hz[1]). There has also been a very large increase in research

[1]Bandwidth in Mexico and Europe

and development in the coding of high fidelity audio signals for transmission and storage of CD-quality music. Interest in wideband speech audio coding has increased during the last years; specially for applications like mobile radio communications, videoteleconferencing, multimedia database access, internet broadcasting and narrow band ISDN.

In the context of audiovisual communications, the quality of telephone-bandwidth speech is acceptable for some telephony services and to maintain backward compatibility. Nowadays, higher bandwidths are required to improve the intelligibility and naturalness of speech, allowing also the transmission/storage of non-speech audio signals. In addition, the excelent quality offered by high fidelity audio coding has a major drawback: its bit rate requirements are too high for some applications. For these reasons wide-band speech audio coding has emerged as a low cost alternative to provide good audio quality at reasonable bit rates. Further information concerning coding bit rates can be found in Chapter 2.

The objective of the present work is to design a variable-rate coding system for wideband speech audio signals. The proposed scheme makes use of filter banks to simulate the selectivity of the ear and also of auditory modeling to mask the errors introduced by the quantization process. The achievements are measured in terms of average bit rate and audio quality.

1.2 Proposed Coding System

In this thesis, a coding system for wide-band speech audio based on a transform working directly on an auditory scale is developed. This transform is the discrete orthogonal wavelet packet transform. For a well chosen decomposition and length of its basis functions, it provides a time frequency mapping that possesses temporal and spectral resolutions close to those achieved by the human ear. The coder developed, like most of the existing coding systems, takes into account the auditory masking model only in the frequency domain. So far, the temporal properties of masking have not been taken into consideration in the design.

The block diagram of the proposed coder is shown in Figure 1.1. The original audio signal is denoted by $x(n)$ and the decoded (reconstructed) signal by $\hat{x}(n)$. The wavelet packet transform of $x(n)$ is given by the coefficients X_i. T_i are the coefficients of the frequency masking threshold. The wavelet packet coefficients are uniformly quantized, according to the computed masking threshold, and coded using a lossless Huffman

algorithm.

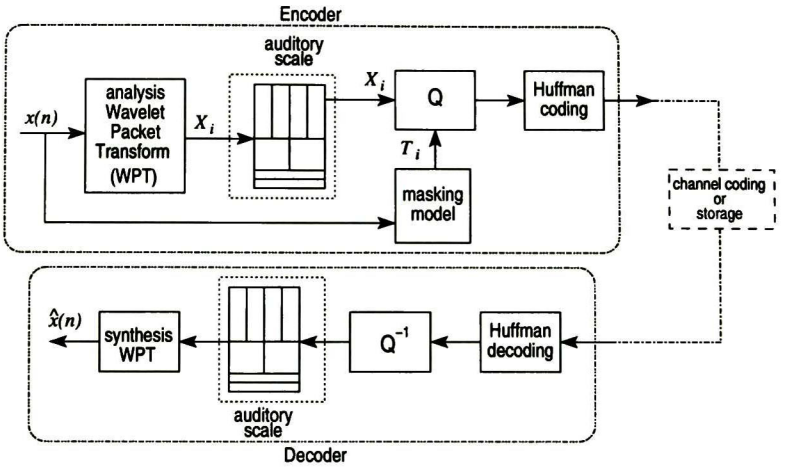


Figure 1.1: *Proposed wide-band speech audio encoder and decoder.*

1.3 Thesis Overview

This thesis is organized as follows: Chapter 2 presents the essential tools such us subband coding, transform coding, quantization, etc. to understand speech and audio compression algorithms. A brief background on state of the art methods for speech and audio coding is also presented.

Chapter 3 outlines the wavelets fundamentals required to develop the proposed coder. Subband, transform coding and its extension to perfect reconstruction parauni-tary transforms, also called orthonormal transforms, are addressed. The design of two channel orthonormal transform and their extension to tree structures is also discussed. Notions of wavelet and wavelet packet transforms are also presented and their connection to tree structure transforms are given. A particular wavelet packet decomposition that approaches the time-frequency resolution of the human ear is also presented.

Chapter 4 deals with the modeling of the human auditory system. First, a general overview of the physiological components of the ear system is broached. Then, auditory masking in the time-frequency domain is presented. The concept of perceptual entropy is provided. Finally, the frequency masking model is adjusted to the needs demanded by the audio coder through a psychoacoustic procedure.

Chapter 5 provides a description of the proposed wide-band speech audio coding/decoding system. Optimal allocation of bits and Huffman coding are also addressed. Coder results are also presented.

Chapter 6 presents the performance and the results of the subjective and objective tests. Chapter 7 concludes this thesis by providing a summary of the main developments, achievements and conclusions. A brief outline of possible future research directions is also discussed.

Chapter 2

Background

Despite the rapid progress in mass-storage density and digital communication–system performance, demand for data transmission bandwidth and storage capacity continues to outstrip the capabilities of available technologies. Particularly, the growth of data-intensive digital audio applications and the increasing use of bandwidth limited media (such as mobile phones, radio links, narrow band ISDN) have not only sustained the need for more efficient ways to encode analog signals, but have made signal compression central to digital communication and signal-storage technology. In this chapter, we provide a brief background on the essentials of speech and audio compression e.g., subband coding, transform coding, quantization. We also review different state of the art coding methods for speech and audio. With this chapter we intend to locate the reader in the context of the presently available speech and audio coders. In subsequent chapters this information will be very useful to know where, among the existing ones, lies the coder we have proposed.

2.1 Speech Compression

The speech signal is a slowly time varying stochastic process whose characteristics, when examined over a sufficiently short period of time (between 10 and 50 msec), are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristics change to reflect the different speech sounds being spoken. There are several ways of classifying events in speech. The simplest and most straightforward is via the state of the speech production source (the vocal cords). We use the three state representation of [67], in which the states are: (1) silence, where

no speech is produced; (2) unvoiced, in which the vocal cords are not vibrating, so the resulting speech waveform is aperiodic or random in nature; and (3) voiced, in which the vocal cords are tensed and therefore vibrate periodically as air flows from the lungs, so the resulting speech waveform is quasi-periodic.

The problem of signal compression, speech in this case, is to achieve a low bit rate in the digital representation of an input signal with a minimum perceived loss of signal quality. For speech this is achieved by eliminating the redundancy between signal samples and by reducing their numerical precision. The function of compression is very often referred to as low bit rate coding, or coding, for short. With very few exceptions in digital speech storage/transmission, speech is generally band-limited to 4 kHz (or 3.4 kHz) and sampled at 8 kHz.

In general, speech coders can be classified according to their bit rate or the technique they use [47, 73]. In addition, their bit rate is usually closely related to their coding technique. In a gross manner, we can classify the speech coder bit rate as high, medium, low or very low. Furthermore, their coding technique can be classified as waveform coding, parametric coding or hybrid. Traditionally, the speech coders that use waveform coding operate within the medium to high range (bit rates of 16 kbps or higher). The hybrid coders operate within the low to medium range (bit rates between 2.4 kbps and 16 kbps). Finally, the parametric coders operate within the very low to low range (bit rates below 2.4 kbps). Figure 2.1 summarizes the bit rates and techniques mentioned above. These techniques do not have to operate necessarily within the previously mentioned ranges, these are only estimates. For example, it means that we can find parametric coders that do not lie in the range of very low bit rate, as in the case of the Mixed Excitation Linear Predictor (MELP) at 4.8 kbps proposed by [55].

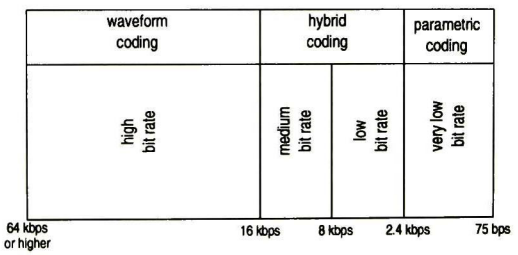


Figure 2.1: *Estimated relationship between speech coders bit rate and coding technique.*

2.1.1 Waveform coding

The waveform coding technique attempts to directly exploit the temporal and/or spectral characteristics of the signal. Basic waveform coders rarely exploit the constraints imposed by the human vocal tract on the speech waveform. Waveform coders are generally more robust than parametric coders. The following coding systems belong to this technique: pulse code modulation (PCM), differential PCM (DPCM), adaptive differential PCM (ADPCM), sinusoidal coders, sub-band and transform coding. Waveform coders can represent non-speech sounds (e.g., music, background noise) accurately, but do so at a higher bit rate than that achieved by efficient speech-specific coders [30, 41, 47, 60].

2.1.2 Parametric coding

Parametric coders rely on speech specific analysis-synthesis which is mostly based on the source-system model. Models that represent the human speech production mechanism (articulatory system) of the vocal tract have been proposed, i.e., distinct human voice-production organs are modeled explicitly. Usually, auto regressive (AR) modeling is employed. This leads us to the use of Linear Predictive Coding (LPC), in which an estimation of the vocal tract transfer function and of the excitation waveform is calculated. This estimation is performed over quasi-stationary speech segments. There exist many different parametric coders based in LPC analysis-synthesis, we usually refer to them as vocoders. Unlike waveform coding, parametric coders are very sensitive to non-speech sounds [30, 41, 47, 60].

2.2 Audio compression

Audio signals are non-stationary stochastic processes. Unlike speech signals, audio signals (such as music or mixed music and speech) cannot be characterized by a common production model. Audio compression algorithms are used to obtain compact digital representation of high-fidelity audio signals for the purpose of efficient transmission or storage. The central objective in audio coding is to represent the signal with a minimum number of bits while achieving transparent[1] signal reproduction [39]. The introduction of the compact disc (CD) in the early 1980's brought to the fore all the advantages of

[1] when the coded signal cannot be distinguished from the original of the coded speech or music.

digital audio (high-fidelity, dynamic range, robustness). These advantages came at the expense of high data rates. Conventional CD audio is typically band limited to 20 kHz and sampled at 44.1 kHz using PCM with a 16-bit sample resolution. This results in uncompressed data rates of 705.6 kbps for a mono aural channel.

There are several classes of analysis-synthesis audio compression algorithms. Within each algorithm class, either lossy or lossless compression is possible. A lossless coding system is able to reconstruct perfectly the samples of the original signal. In contrast, a lossy scheme is incapable of perfect reconstruction from the coded representation of the signal. Lossy coders, also known as perceptual coders, exploit the psychoacoustic principles and statistical redundancies to achieve bit rate reduction. The coder we have proposed in the previous chapter and that we will describe further in subsequent chapters uses a lossy compression scheme. Most of the perceptual audio coding algorithms are based on the generic architecture shown in Figure 2.2.

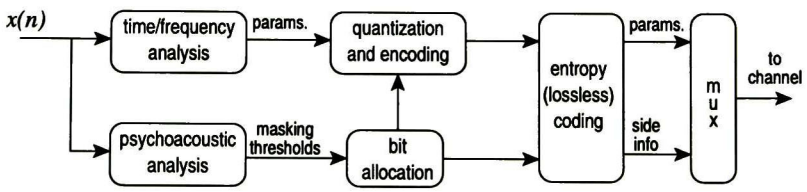


Figure 2.2: *Generic perceptual audio encoder.*

These coders typically segment the input signal into quasi-stationary frames ranging from 2 to 50 msec in duration. Then, a time-frequency analysis estimates the temporal and spectral components of each frame [60]. Often, the time-frequency mapping is matched to the analysis properties of the human auditory system, although this is not always the case. The main goal of this is to extract from the input audio a set of time-frequency parameters. These parameters are quantized and encoded in accordance with a perceptual distortion metric. Depending on the system objectives and design philosofy, the time-frequency representation may contain one of the following methods [60]:

- unitary transform;
- time-invariant bank of critically sampled, uniform, or nonuniform bandpass filters;

- time-varying (signal-adaptive) bank of critically sampled, uniform, or non uniform band pass filters;
- harmonic/sinusoidal analyzer;
- source-system analysis (LPC/multi pulse excitation);
- hybrid transform/filter bank/sinusoidal/LPC signal analyzer.

The choice of time frequency-analysis methodology always involves a fundamental tradeoff between time and frequency resolution requirements. The psychoacoustic analysis delivers masking thresholds that quantify the maximum amount of allowable distortion such that the quantization of the time-frequency parameters does not include audible artifacts. Generally, the quantization and encoding sections exploit statistical redundancies through classical techniques (ADPCM, VQ, PDF-optimized). Once a quantized compact parametric set has been formed, remaining redundancies are typically removed through entropy (lossless) coding (arithmetic, Huffman, Lempel-Ziv). Since the output of the psychoacoustic distortion control model is signal dependent, most of the algorithms are inherently variable rate. Fixed rate requirements are usually satisfied through buffer feedback systems [60].

2.3 State of the art

As we have stated before, speech and audio coding techniques can be roughly classified into one of three categories, depending on the bandwidth of the considered signals:

1. *Toll quality.* Normally used for speech. Limited to the frequency range of 300–3400 Hz and sampled at 8 kHz.
2. *Wide-band speech quality.* Intended for speech and audio signals in the frequency range of 50–7000 Hz and sampled at 16 kHz.
3. *High quality.* Used for speech and audio signals within the frequency range of 20-20000 Hz. The sampling frequency can be either 44.1 kHz (for CD audio quality) or 48 kHz (for DAT audio quality).

In what follows, we will briefly overview the coding techniques in these three domains through specific examples of some of the most extensively used standards: the ITU

G.729 CS-ACELP [36]; the ITU G.722 audio coding within 7 kHz [35]; and the ISO/IEC MPEG-1 audio coder for high quality applications [37].

2.3.1 Toll quality

The standard wireline quality of a telephone call is nowadays achievable at 8 kbps by making use of the ITU-T Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP) coding standard [36]. The coder is designed to operate with a toll bandwidth digital signal, sampled at 8 kHz with 16-bit linear PCM samples. The coder is based on the Code-Excited Linear-Prediction (CELP) coding method [57, 29, 73, 30]. The CELP is an analysis-by-synthesis method for encoding speech signals. Figure 2.3 presents the block diagram of the CELP encoder.

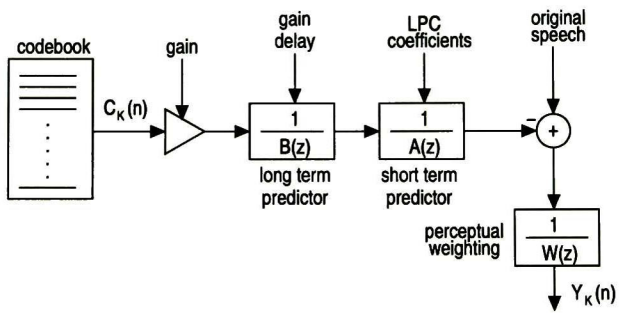


Figure 2.3: *Code excited linear predictor (CELP) scheme. The goal is to minimize $Y_k(n)$ by selecting the best codebook entry.*

The block diagram of the CS-ACELP encoder is shown in Figure 2.4. The coder operates on speech frames of 10 msec (80 samples) each. The input signal is high pass filtered and scaled in the pre-processing block. The pre-processed signal serves as the input signal for all subsequent analysis. LP analysis is done once per 10 msec frame to compute the LP filter coefficients. These coefficients are converted to Line Spectrum Pairs (LSP) and quantized using a two-stage predictive VQ with 18-bits. The excitation signal is chosen by using an analysis-by-synthesis search procedure in which the error between the original and reconstructed speech is minimized according to a perceptually weighted distortion measure. This is done by filtering the error signal with a perceptual weighting filter, whose coefficients are derived from the unquantized LP filter. The fixed and adaptive codebook parameters are determined per subframe of 5 msec each. The

quantized and unquantized LP filter coefficients are used for the second subframe, while in the first subframe interpolated LP coefficients are used. An open-loop pitch delay is estimated once per 10 msec frame based on the perceptually weighted speech signal. The following operations are repeated for each subframe. The target signal, $x(n)$, is computed by filtering the preprocessed voice signal through the weighted synthesis filter $W(z)/\hat{A}(z)$. The initial states of these filters are updated by filtering the error between the LP residual and excitation. The impulse response $h(n)$ of the weighted synthesis filter is computed. Open-loop analysis is performed to find a coarse estimation of the pitch, using the impulse response, $h(n)$, and the target signal, $x(n)$. Then, the pitch is refined by a closed-loop analysis to find the adaptive-codebook delay and gain. The pitch delay is encoded with 8-bits in the first subframe and differentially encoded with 5-bits in the second subframe. The target signal, $x(n)$, is updated by substracting the (filtered) adaptive-codebook contribution. The new target signal $x'(n)$ is used in the fixed-codebook search to find the optimum excitation. The gains of the adaptive and fixed-codebook contributions are vector quantized with 7-bits. Finally the filter memories are updated using the determined excitation signal.

2.3.2 Wide-band speech audio

In 1986 the ITU proposed the standard G.722 for 7 kHz audio coding in the form of a subband-ADPCM coder [35, 58, 59, 56]. It was initially developed for narrow band ISDN teleconferencing systems, multi point interactive audiovisual communications and loudspeaker telephony. Figure 2.5 illustrates the structure of the ITU G.722 audio coder.

The ITU G.722 standard supports bit rates of 64, 56, and 48 kbps. This coder is based on two identical Quadrature Mirror Filters (QMF) that divide the 16 kHz sampled 14-bit PCM signal into two critically subsampled (8 kHz sampled) components called low subband and high subband. The filters overlap, and aliasing will occur because of the critically subsampling. This problem is overcome by the synthesis QMFs in the receiver, which ensure that the aliasing products are canceled, in addition they also have a stopband attenuation of 60 dB. The coding of the subband signals is based on a modified version of the 32 kbps ITU G.721 ADPCM speech coder. Input samples are adaptively predicted, the prediction error signal (difference) is quantized and transmitted. The predictor coefficients are updated sample-wise under the control of the coded difference signal, that is also available at the decoder. The quantizer is also adaptive and can

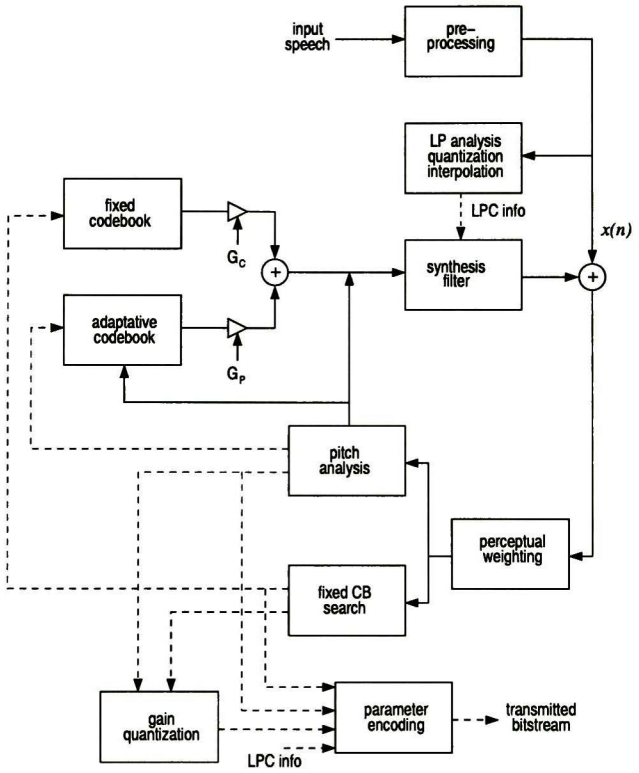


Figure 2.4: *Encoding principle of the ITU-T G.729 CS-ACELP encoder.*

rapidly adapt itself to the changing statistics of the audio signal. High quality coding is provided by a fixed bit allocation where the low and high subband ADPCM coders use a 6-bits per sample and 2-bits per sample quantizer, respectively. In the low subband the signal resembles the narrow band signal in most of its properties. A reduction of the quantizer resolution to 5- or 4-bits/sample is possible to support transmission at lower rates or auxiliary data at rates of 8 kbps and 16 kbps. Embedded coding is used in the low subband ADPCM coding, i.e., the adaptations of predictor and quantizer are always based only on the four most significant bits of each ADPCM codeword. Hence, a stripping of one or two least significant bits from the ADPCM codewords do not affect the adaptation processes. A transparent audio quality has been reported for the G.722 coder at 64 kbps [58]. Two different MOS values have been reported for the ITU G.722, in [58] the author suggests a MOS of 4.1 and in [56] the author suggests a MOS of 3.3. For the present work, we will take the second as the MOS value for the G.722.

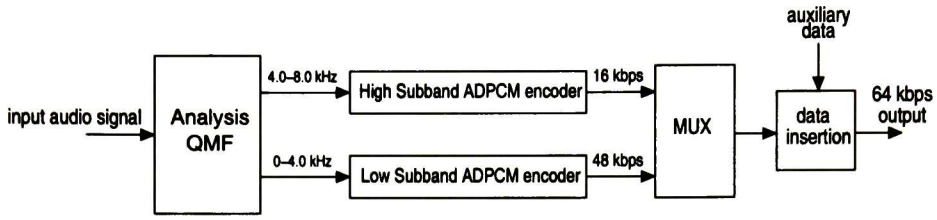


Figure 2.5: *Structure of the ITU-T G.722 audio coder.*

This measure will be found very useful for comparison in subsequent chapters. The G.722 also proved to be error robust at the higher bit rates [58], 64 kbps and 56 kbps respectively.

2.3.3 High quality coding

The goal of this section is to give a short introduction of the MPEG audio compression scheme [37]. The MPEG audio compression is the first international standard for compression of high-fidelity audio. The MPEG audio standard, adopted in 1992, results from more than three years of work by an international committee of high-fidelity audio compression experts within the Moving Pictures Experts Group (MPEG/audio). Although perfectly suitable for audio applications only, the MPEG/audio is actually one of a three-part compression standard that also includes video and systems [37, 61]. The MPEG audio coder is essentially derived from the adaptive spectral perceptual entropy coding (ASPEC) [8] and the masking-pattern universal subband integrated coding and multiplexing (MUSICAM) [24] algorithms. MPEG/audio is a generic audio compression algorithm, which makes no assumption about the nature of the audio source. Instead, the coder exploits the perceptual limitations of the human auditory system. Much of the compression results from the removal of perceptually insignificant parts of the audio signal. MPEG/audio offers a choice of three independent layers of compression. This provides a wide range of trade-offs between CODEC complexity and compressed audio quality. Layer I, the simplest, best suits bit rates above 128 kbps per channel. For example, Philips' Digital Compact Cassette (DCC) uses Layer I compression at 192 kbps. Layer II has an intermediate complexity and targets bit rates around 128 kbps. Possible applications for this layer include Digital Audio Broadcasting (DAB) audio coding, and the storage of synchronized video-and-audio sequences on CD-ROM. Layer III is the most complex, but offers the best ratio between audio quality and data rate, particu-

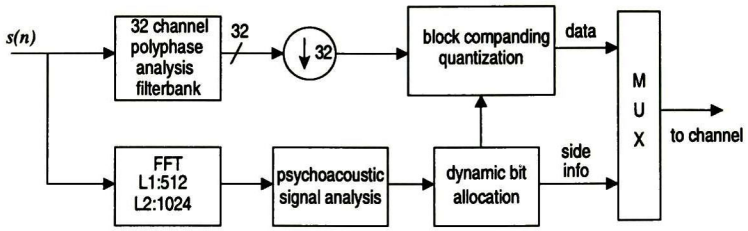


Figure 2.6: *Block diagram of MPEG-1 Layer I and Layer II encoder.*

larly for bit rates around 64 kbps per channel. This layer has been very successful in many applications. For example, MPEG-1 Layer III has become the most widely used standard for transmission and storage of compressed audio for both WWW and hand held media [9, 78].

Layers I and II, presented in Figure 2.6, work as follows. The input signal is first decomposed into 32 critically subsampled subbands using a polyphase realization of a PQMF (Pseudo-Quadrature Mirror Filter) bank [77, 60, 82]. The channels are equally spaced such that a 44.1 kHz sampled input signal is split into subbands of approximately 690 Hz, with the subbands decimated 32:1. A 511th-order prototype filter was chosen such that the inherent overall PQMF distortion remains inaudible. Moreover, the prototype filter was designed for very high sidelobe attenuation (96 dB) to insure that the intraband aliasing due to quantization noise remains negligible. For the purpose of psychoacoustic analysis and determination of JND (Just Noticeable Distortion) thresholds, a 512 (Layer I) or 1024 (Layer II) point FFT is computed in parallel with the subband decomposition for each decimated block of 12 input samples (8.7 msec at 44.1 kHz). Next, the subbands are normalized by a scale factor such that the maximum sample amplitude in each block is unity. Then, an iterative bit allocation procedure applies the JND thresholds to select an optimal quantizer from a predetermined set for each subband. In each subband, scale factors are quantized using 6 bits each and quantizer selections are encoded using 4 bits each. For Layer I encoding, decimated subband sequences are quantized and transmitted to the receiver in conjunction with side information [37, 61].

Layer II improves three portions of layer I in order to realize enhanced output quality and reduce bit rates at the expense of greater complexity and increased delay. Particularly, the perceptual model relies on higher resolution FFT. The maximum sub-

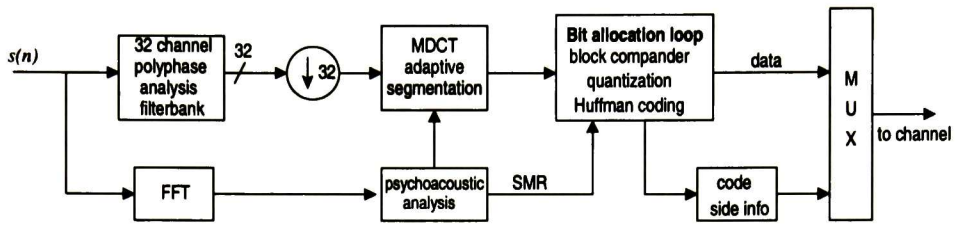


Figure 2.7: *Block diagram of MPEG-1 Layer III encoder.*

band quantizer resolution is increased, and the scale factor side information is reduced while exploting temporal masking by considering properties of three adjacent 12-sample blocks and optionally transmitting one, two, or three scale factors. Average MOS's of 4.7 and 4.8 were reported in [58] for one channel Layer I and Layer II operating at 192 and 128 kbps respectively.

Layer III MPEG (Figure 2.7) architecture achieves performance improvements by adding several important mechanisms on top of the Layer I/II foundation. A hybrid filter bank is introduced to increase frequency resolution and thereby better approximate critical band behavior. The hybrid filter bank includes adaptive segmentation to include pre-echo control. Sophisticated bit allocation and quantization strategies that rely upon non-uniform quantization, analysis-by-synthesis, and entropy coding are introduced to allow reduced bit rates and improved quality. The hybrid filter bank is constructed by following each subband filter with an adaptive Modified Discrete Cosine Transform (MDCT) [77, 51, 60]. This practice allows higher frequency resolution and pre-echo control. For example, an 18-point MDCT improves frequency resolution to 38.3 Hz per spectral line. The adaptive MDCT switches between 6–18 points to allow improved pre-echo control. Shorter blocks provide temporal premasking of pre-echoes during transients; longer block during steady-state periods improve coding gain, while reducing side information and hence bit rates. Bit allocation and quantization of the spectral lines are realized in a nested loop procedure that uses both nonuniform quantization and Huffman coding. The inner loop adjusts the nonuniform quantizer step sizes for each block until the number of bits required to encode the transform components falls within the bit budget. The outer loop evaluates the quality of the coded signal (analysis-by-synthesis) in terms of quantization noise relative to the JND threshold. Average MOS of 3.1 and 3.7 were reported in [58] for one channel Layer II and Layer III codecs operating at 64 kbps.

2.4 Subband coding

Signal compression, one of the main applications of digital signal processing, uses signal expansions as a major component. When the channels of a filter bank are used for coding, the resulting scheme is known as *subband coding* [82]. Driven by applications like speech and image compression, subband coding was proposed by Croisier, *et al.* [20] using a special class of filters called Quadrature Mirror Filters (QMF) in the late 1970's. Due to the strong relationship between subbband coding and filter banks we will refer to them as a single term during this short introduction.

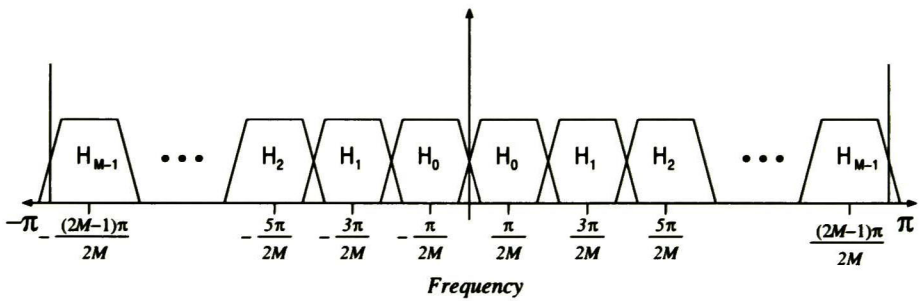


Figure 2.8: *Magnitude response of a uniform M-band filter bank.*

The time-frequency tool most commonly employed for mapping the time-domain input signal to a set of quantizable frequency parameters is the filter bank. The filter bank divides the entire spectrum into frequency subbands and generates a time-indexed series of coefficients representing the frequency localized signal power within each band. The filter bank provides us essential information about the distribution of the signal and hence masking power over the time-frequency plane. The filter bank plays a very important role in the perceptual irrelevancies identification when used in conjuntion with a perceptual model. Time-frequency parameters generated by the filter bank provide a signal mapping that is manipulated to shape the coding distortion in order to match the time-frequency distribution of the masking power. By decomposing the signal into its frequency components, the filter bank also helps in the reduction of statistical redundancies. For example, Figure 2.8 shows the magnitude response of a uniform bandwidth M-channel filter bank. The M analysis filters have normalized center frequencies $(2k+1)/2M$ and are characterized for individual impulse and frequency responses ($h_k(n)$ and $H_k(\omega)$ respectively, for $0 \leq k < M$).

Filter banks for audio coding are better described in terms of an analysis-synthesis framework, as shown in Figure 2.9. The input signal $x(n)$ is processed at the encoder by a parallel bank of $(L-1)$th order FIR bandpass filters $H_k(z)$. The bandpass analysis outputs

$$v_k(n) = h_k(n) \star x(n) = \sum_{m=0}^{L-1} x(n-m)h_k(m), \qquad k = 0, 1, ..., M-1 \tag{2.1}$$

are decimated by a factor of M, yielding the subband sequences

$$\begin{aligned} y_k(n) &= v_k(Mn) \\ &= \sum_{m=0}^{L-1} x(Mn-m)h_k(m), \qquad k = 0, 1, ..., M-1 \end{aligned} \tag{2.2}$$

which comprise a *critically sampled* or *maximally decimated* signal representation, i.e., the number of subband samples is equal to the number of input samples. Because it is impossible to achieve perfect magnitude responses with finite order bandpass filters [66], there is unavoidable aliasing between the decimated subband sequences. Quantization and coding are performed on the subband sequences $y_k(n)$ to yield $\hat{y}_k(n)$. Then, the decoder receives the subband samples $\hat{y}_k(n)$, where they are upsampled by M to form the intermediate sequences

$$w_k(n) = \begin{cases} \hat{y}_k(n), & n = 0, M, 2M, 3M, ... \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

In order to eliminate the imaging distortions introduced by the upsampling operations, the sequences $w_k(n)$ are processed by a parallel bank of synthesis filters, $g_k(n)$. Then, the filter outputs are combined to form the overall output $\hat{s}(n)$. The filterbank's encoding/decoding mathematical description, in the absence of quantizers, is shown in the following equations

$$\begin{aligned} V_k(z) &= H_k(z)X(z), \qquad k = 0, 1, ..., M-1 \\ Y_k(z) &= \frac{1}{M} \sum_{\alpha=0}^{M-1} H_k(z^{\frac{1}{M}} e^{\frac{2\pi i \alpha}{M}}) X(z^{\frac{1}{M}} e^{\frac{2\pi i \alpha}{M}}) \\ W_k(z) &= \frac{1}{M} \sum_{\alpha=0}^{M-1} H_k(z\, e^{\frac{2\pi i \alpha}{M}}) X(z\, e^{\frac{2\pi i \alpha}{M}}) \\ \hat{X}(z) &= \frac{1}{M} \sum_{k=0}^{M-1} \sum_{\alpha=0}^{M-1} H_k(z\, e^{\frac{2\pi i \alpha}{M}}) X(z\, e^{\frac{2\pi i \alpha}{M}}) G_k(z) \end{aligned} \tag{2.4}$$

The z-transforms of $v(n) = (\downarrow M)x(n)$ and $u(n) = (\uparrow M)v(n)$ are provided in [77, 80]. For perfect reconstruction filterbanks, as long as there is no quantization, the output $\hat{x}(n)$ will be identical to the input $x(n)$ with a delay, i.e., $\hat{x}(n) = x(n - n_0)$.

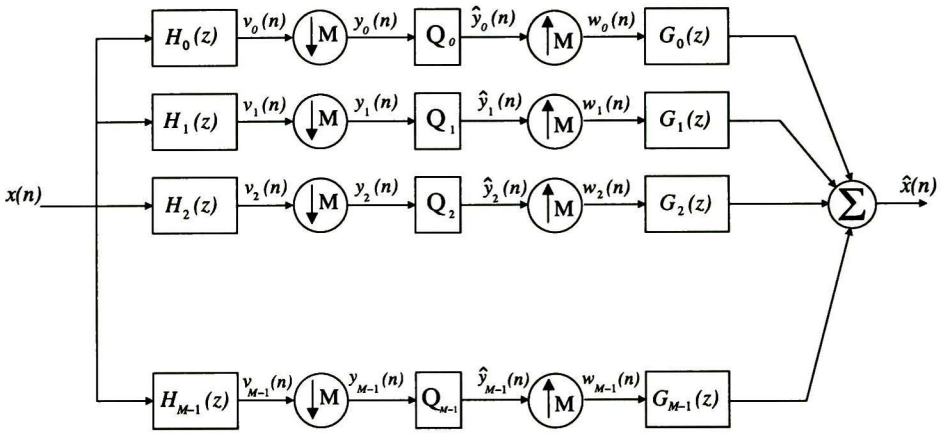


Figure 2.9: *Uniform M-band maximally decimated analysis-synthesis filter bank.*

As we have stated before, efficient coding performance depends heavily on adequately matching the properties of the analysis filter bank to the characteristics of the input signal. During the filter bank structure design for perceptual audio coding applications, it is necessary to face the difficult tradeoff between time and frequency resolution, because no single resolution is optimal for all signals. Audio signals are highly non-stationary and contain significant tonal and non-tonal energy, as well as both steady-state and transient intervals. Due to this fact, an ideal coder should contain an adaptive time-frequency resolution, i.e., a time varying filter bank whose structure adapts according to the signal being processed. In short, a number of highly desirable characteristics for audio coding using filter banks are

- Signal adaptive time-frequency tiling;
- Efficient resolution switching;
- good channel separation;
- strong stopband attenuation;
- perfect reconstruction;
- critical sampling;

- availability of fast algorithms

We will review additional information about filter banks in the next chapter. We will also address the strong connection between filter banks and discrete time wavelets. For a deeper analytical development about filter banks the reader is referred to [77, 80, 68, 82, 79].

2.5 Transform coding

As we have previously mentioned in Section 2.1.1, in the area of waveform coding there exists a very popular technique called *transform coding*. The name of transform coding was first used by Jayant and Noll [41], but the idea was introduced by Kramer and Mathews in 1956 [46]. Transfom coding can also be seen as a special case of subband coding [80]. During this description, as in most of the cases, we will suppose the transformation is orthogonal and linear. Figure 2.10 illustrates the general structure of the transform coding principle, where $\mathbf{T}$ is a $M \times M$ invertible matrix that performs the linear transformation [41, 29, 57, 70, 77, 82].

Let us suppose we have a block of consecutive samples of a stationary random process that we wish to efficiently code using a specified number of bits. Let $\mathbf{X}$ denote the length M sample vector

$$\mathbf{X}(n) = [x(nM) \;\; x(nM-1) \;\; x(nM-2) \;\; \cdots \;\; x(nM-M+1)]^T \tag{2.5}$$

These samples will tipically have substantial correlation and due to the stationarity of the process they also have the same PDF; the same mean $\mu = 0$, unless otherwise stated; and the same variance σ_X^2. The idea of transform coding is to perform a suitable linear transformation on the input vector, $\mathbf{X}$, and at the output we obtain a new vector, $\mathbf{Y}$ also with M components, usually called transform coefficients.

A very important characteristic of $\mathbf{Y}$ is that its coefficients are much less correlated than the original samples. In addition, the information may be much more compact in the sense of being concentrated in only a few of the transform coefficients. Once the statistical redundancy has been removed, we are able to quantize these coefficients in a more efficient way[2] It is very important to point out that every coefficient needs a

[2]There exists no general theorem that states that uncorrelated variables can be more efficiently quantized than correlated variables. But as a matter of practice and experience Tansform Coding has proved to be a simple and effective way of obtaining good compression [29].

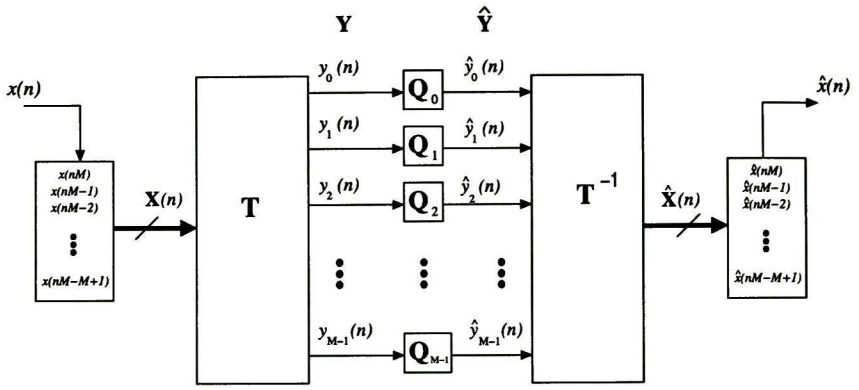


Figure 2.10: *Structure of the Transform Coding principle.*

different quantizer, since the transform coefficients may, in general, have different PDFs [28, 57]. The reconstructed approximation to the original vector, $\hat{\mathbf{X}}$, is obtained by performing the corresponding inverse operation on the quantized transformed vector, $\hat{\mathbf{Y}}$, as follows

$$\hat{\mathbf{X}} = \mathbf{T}^{-1}\hat{\mathbf{Y}} \tag{2.6}$$

and

$$\mathbf{Y} = \mathbf{T}\,\mathbf{X} \tag{2.7}$$

$\mathbf{T}^{-1}$ is called the *inverse transform*, which satisfies the condition $\mathbf{T}^{-1}\mathbf{T} = \mathbf{T}\,\mathbf{T}^{-1} = \mathbf{I}$, where $\mathbf{I}$ is the $M \times M$ identity matrix. In addition, we can define an orthogonal transformation as a linear operation in which the transformation matrix, $\mathbf{T}$, satisfies the orthogonality condition

$$\mathbf{T}^{\mathsf{T}} = \mathbf{T}^{-1} \tag{2.8}$$

where $\mathbf{T}^{\mathsf{T}}$ denotes the transpose of $\mathbf{T}$. One of the most important advantages of orthogonal transforms is that they prevent the propagation of the quantization noise [29]. If we express $\mathbf{T}$ in terms of its component vectors, $\mathbf{T}^{\mathsf{T}} = [\mathbf{t}_0\ \mathbf{t}_1\ \cdots\ \mathbf{t}_{M-1}]$, the reconstructed signal, $\hat{\mathbf{X}}$, can be written as the linear combination of the orthogonal columns of $\mathbf{T}^{\mathsf{T}}$

$$\hat{\mathbf{X}} = \sum_{i=0}^{M-1} y_i(n)\mathbf{t}_i \tag{2.9}$$

The $\mathbf{t}_i$ vectors are known as *basis functions* of the transform. They form the rows of $\mathbf{T}$ and the columns of $\mathbf{T}^\mathsf{T}$. We say the transformation is *orthonormal* if the basis functions have unity norm. An essential property of an orthonormal transform is its capacity to preserve the power after the tansformation, also known as isometry property, where the average variance of the transform (output) coefficients equals the variance of the input [57]

$$\frac{1}{M}\sum_{i=0}^{M-1}\sigma_{y_i}^2 = \frac{1}{M}\sum_{i=0}^{M-1}\mathrm{E}\{y_i^2(n)\} = \frac{1}{M}\mathrm{E}\{\mathbf{Y}^\mathsf{T}\mathbf{Y}\} = \frac{1}{M}\mathrm{E}\{\mathbf{X}^\mathsf{T}\mathbf{T}^\mathsf{T}\mathbf{T}\mathbf{X}\}$$
$$= \mathrm{E}\{\mathbf{X}^\mathsf{T}\mathbf{X}\} = \sum_{i=0}^{M-1}\mathrm{E}\{x_i^2(n)\} = \frac{1}{M}\sum_{i=0}^{M-1}\sigma_X^2(n) = \sigma_X^2 \tag{2.10}$$

2.5.1 Optimal Orthogonal Transform

As we have previously mentioned, the components of $\mathbf{X}$ are correlated with one another. However, it is possible to select an orthogonal matrix $\mathbf{T}$, for a given PDF describing $\mathbf{X}$, that will make $\mathbf{Y} = \mathbf{TX}$ to have pairwise uncorrelated components. The matrix that makes this linear transformation possible is called the *discrete time Karhunen-Loève Transform (KLT)* or the *Hotelling Transform* [29, 41, 51, 57, 70, 77, 82]. Let us denote the autocorrelation matrix of the input vector, $\mathbf{X}$, as $\mathbf{R}_\mathbf{X} = \mathrm{E}[\mathbf{X}\mathbf{X}^\mathsf{T}]$. Let $\mathbf{u}_i$ denote the eigenvectors of $\mathbf{R}_\mathbf{X}$ (normalized to unit norm) and λ_i the corresponding eigenvalues. Since the correlation matrix is symmetric and nonnegative definite, there are M orthogonal eigenvectors and the corresponding eigenvalues are real and nonnegative. In general, during the indexing we put them in descending order

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M \geq 0$$

the Karhunen-Loève transform matrix is defined as $\mathbf{T} = \mathbf{U}^\mathsf{T}$ where

$$\mathbf{U} = [\mathbf{u}_1\, \mathbf{u}_2\, \cdots\cdots\, \mathbf{u}_M] \tag{2.11}$$

that is, the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{R_X}$. Then, the auto correlation matrix of $\mathbf{Y}$ is given by

$$\begin{aligned} \mathbf{R_Y} &= \mathrm{E}[\mathbf{Y}\,\mathbf{T}^{\mathsf{T}}] = \mathrm{E}[\mathbf{U}^{\mathsf{T}}\mathbf{X}\,\mathbf{X}^{\mathsf{T}}\mathbf{U}] \\ &= \mathrm{E}[\mathbf{U}^{\mathsf{T}}\mathbf{R_X}\mathbf{U}] = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_M \end{pmatrix} \end{aligned} \tag{2.12}$$

Thus we see that the KLT does indeed decorrelate the coefficients of the input vector. The variances of the transform coefficients are the eigenvalues of the autocorrelation matrix $\mathbf{R_X}$. However, the KLT has the disadvantage of being signal-dependent, which means that it must be recomputed for each different input signal. Furthermore, for nonstationary signals the transformation matrix, $\mathbf{T}$, will be different from one frame to another. There exists no fast algorithm for the implementation of the KLT. It means that the product given by Equation 2.7 must be computed in the traditional way. All these reasons make the KLT very inefficient. However, the KLT serves as a theoretical bound and reference for transform coding performance. For practical purposes, other transforms possessing fast algorithms and efficient implementations are employed. They all try to approximate the performance of the KLT. Some of the most common orthogonal transforms are the Discrete Hadamard Transform (DHT) [41, 70], the Discrete Fourier Transform (DFT) [66, 51, 41], the Discrete Cosine Transform (DCT) [41, 51, 57, 70, 77], and the Discrete Wavelet Transform [18, 19, 34, 51, 70, 77, 82].

2.6 Quantization

In a digital system all signals are represented as fixed point binary fractions. This is a b-bit binary representation of the signal. We usually refer to b as the *wordlength* of the digital system. In a real life application, it is necessary to transform a time indexed continuous signal or parameter (generally a random variable), $s(n)$, into a value, $\hat{s}(n)$, taken from a finite set of possible values. In signal processing literature, we usually refer to this operation as *quantization*[3] Thus, we say that a *scalar quantizer* is a device

[3]The dictionary definition of quantization is the division of a quantity into a discrete number of small parts, often assumed to be integral multiples of a common quantity. This definition seems alike to the previously stated one

which takes an arbitrary real number and converts it to a b-bit fraction according to some arithmetic rules. Figure 2.11 presents the block diagram of a scalar quantizer.

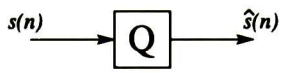


Figure 2.11: *Block diagram of a quantizer.*

There exist two necessary conditions that all quantizers must accomplish in order to achieve optimality

1. Given a dictionary $\{\hat{s}^1 \cdots \hat{s}^L\}$, the best partition is that which satisfies

$$P^i = \{s : (s - s^i)^2 \leq (s - s^j)^2 \qquad \forall j \in \{1 \cdots L\}\}$$

   This condition is also known as the nearest neighbor rule.

2. Given a partition $\{P^1 \cdots P^L\}$, its best representatives are obtained by using the centroid (or center of gravity) condition. This is calculated using the PDF within the region P^i

We can describe the quantization process analytically as $\hat{s}(n) = Q(s(n))$. Unless otherwise specified, we will assume that $x(n)$ is a wide sense stationary process (WSS), whose mean value is zero and with a dynamic range restricted to $-A \leq s(n) < A$. Furthermore, we will consider that we have a roundoff scalar quantizer that possesses a *midtread characteristic*, where zero is one of the output representation levels, i.e., the origin is in the middle of a tread [29]. The output representation levels of a midtread quantizer are depicted in Figure 2.12.

The quantization step of a midtread quantizer is given by

$$\delta = \frac{2A}{2^b - 1} \tag{2.13}$$

Midtread quantizers possess an odd number of levels, $L = 2^b - 1$, and are often preferred to midrise quantizers[4], since quantized values are not affected by small fluctuations around zero or "silence" intervals. As mentioned above, the scalar quantization

[4]In a *midrise quantizer* the origin is a boundary point that is equidistant from two adjacent output points.

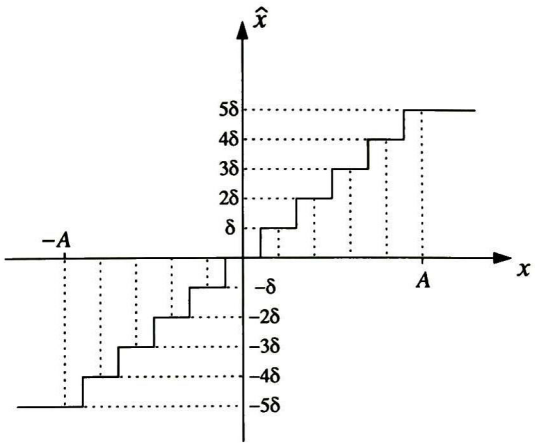


Figure 2.12: *Midtread quantizer characteristic.*

procedure consists in mapping an incoming value $s_i(n)$ onto an output value which is an integer multiple of the quantization step

$$\hat{s}_i = \text{sign}(s_i(n))\, N_i \delta \tag{2.14}$$

where N_i is the number of levels necessary to represent s_i and it is within the range $-2^{b-1} - 1 \leq N_i \leq 2^{b-1} - 1$. This number of levels is calculated using the expression

$$N_i(n) = \left\lfloor \frac{|s_i(n)|}{\delta} + \frac{1}{2} \right\rfloor \tag{2.15}$$

where $\lfloor y \rfloor$ stands for the integer part of y. As we can see, this is a roundoff operation, since $\text{round}(y) = \lfloor y + \frac{1}{2} \rfloor$. During this process, some error is introduced and we will denote it as $q(n)$. This error is usually called *quantization noise* and it is a random variable commonly assumed to be WSS, uniformely distributed over the interval $-\delta/2 \leq q \leq \delta/2$. Its value is

$$q(n) = s(n) - \hat{s}(n) \tag{2.16}$$

It is necessary to point out that these assumptions are no longer valid when b is small. If we assume that the number of quantization levels, L, is large, it is possible to obtain an analytical expression for the power (variance) of the quantization noise. This expression is a function only in terms of the PDF, $p_s(x)$. This assumption is known as the *high resolution hypothesis* [57, 29]. It states that the PDF can be supossed constant

within the interval $[t^{i-1}, t^i]$ and its representative value can be taken from the middle of the interval. Then, we can write

$$\begin{aligned} p_s(x) \approx p_s(\hat{s}^i) \qquad & \text{for} \qquad x \in [t^{i-1}, t^i] \\ \hat{s}^i \quad \approx \quad & \frac{[t^{i-1} + t^i]}{2} \end{aligned} \tag{2.17}$$

We call

$$\Delta(i) = t^i - t^{i-1} \tag{2.18}$$

the length of the interval is $[t^{i-1}, t^i]$ and the probality that $s(n)$ belongs to this interval is given by

$$P_r(i) = P_r\{S \in [t^{i-1}, t^i]\} = p_s(\hat{s}^i)\Delta(i) \tag{2.19}$$

The variance of the quantization noise is given by the expression

$$\sigma_q^2 = \sum_{i=1}^{L} p_s(\hat{s}^i) \int_{t^{i-1}}^{t^i} (x - \hat{s}^i)^2 dx \tag{2.20}$$

Using some algebra we can write

$$\int_{t^{i-1}}^{t^i} (x - \hat{s}^i)^2 dx = \int_{-\Delta(i)/2}^{\Delta(i)/2} x^2\, dx = \frac{\Delta^3(i)}{12} \tag{2.21}$$

then, using the previous expression with Equation 2.19, we obtain

$$\sigma_q^2 = \sum_{i=1}^{L} P_r(i) \frac{\Delta^2(i)}{12} = \frac{1}{12} \mathrm{E}\{\Delta^2\} \tag{2.22}$$

As we can see in the last expression, the variance of the quantization noise *depends only* on the interval length, $\Delta(i)$. Now, let us suppose that $s(n)$ has a uniform PDF ($p_s(x) = \frac{1}{2A}$, $\mu_s = 0$, $\sigma_s^2 = A^2/3$). Then, using Equation 2.22 we can show (see Appendix A) that the quantization noise variance for a unifom PDF is given by

$$\sigma_q^2 = \frac{\sigma_x^2}{L^2} = \frac{A^2}{3\,L^2} \tag{2.23}$$

This result will be very useful in Chapter 5, for quantizing the transform coefficients with a noise variance below the Just Noticeable Distortion (JND) threshold.

In the practice, we do not know the PDF of the signal, $p_s(x)$. To design a quantizer, we have to use empirical data and a learning database. The learning database must be

composed with a large number of the source representative samples. To overcome this problem and build a quantizer based on empirical data we use the Lloyd-Max iterative algorithm. An explanation of this algorithm is out of the scope of this thesis, but a detailed description is provided in [29, 57, 70].

By grouping source outputs together and encoding them as a single block, we can obtain efficient lossy compression algorithms. This algorithm is called Vector Quantization (VQ). We can say that VQ is a generalization of the scalar quantization. While scalar quantization is primarily used for analog-to-digital conversion, VQ is more used with sophisticaded DSP algorithms. A deep explanation of VQ is not among the goals of this thesis, but the reader is referred to the excellent bibliography that exists in that field [29, 57, 70].

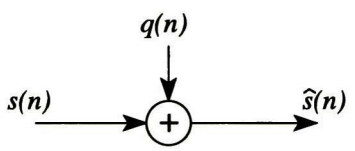


Figure 2.13: *Additive noise model of the quantizer.*

2.6.1 Additive noise model

As we have already seen in Equation 2.16, the quantization error can be considered as additive and the quantized signal is $\hat{s}(n) = s(n) + q(n)$, as presented in Figure 2.13. Thus, by squaring $\hat{x}(n)$ and taking the mathematical expectation we find

$$\sigma_{\hat{s}}^2 = \sigma_s^2 + \sigma_q^2 + \mathrm{E}\{s\,q\} \tag{2.24}$$

If we maintain the previously mentioned assumption that the quantization noise, $q(n)$, is white (this assumption is valid only if L is large, i.e., δ is small) then, as we know, it is uncorrelated with $y(n)$, which implies that $\mathrm{E}\{s\,q\} = 0$. Then, using Equation 2.23, the quantization process can be expressed by the aditive and input independent noise model

$$\sigma_{\hat{s}}^2 = \sigma_s^2 + \frac{\delta^2}{12} \tag{2.25}$$

This means that the quantizer increases the variance of the input signal, $s(n)$, by the amount $\delta^2/12$. At low bit rates (1 or 2 bits/sample), the noise component is no longer

strictly additive and it is correlated with the input signal, $s(n)$. However, the additive components of $q(n)$ are still the main source of perceptual degradation in the quantization process [12, 41].

2.7 Optimum bit allocation

Let us suppose that we have calculated a time invariant linear transformation, $\mathbf{Y}(n) = \mathbf{T}\,\mathbf{X}(n)$. The output of this transformation is a vector of length M containing the samples $\mathbf{Y}(n) = [Y_0(Mn)\ Y_1(Mn-1) \cdots Y_{M-1}(Mn-M+1)]$ and we have bM bits to quantify these transform coefficients. We will also suppose that these M coefficients have mean $\mu_i = 0$, variance $\mathrm{E}\{Y_i^2\} = \sigma_{Y_i}^2$, and that we know the PDF of every Y_i. On the average, every coefficient can be represented with a resolution of b bits, i.e., we have b bits available for each coefficient. It is often desirable in practice that b be an integer, but it is sufficient if the product of the available bits per coefficient multiplied by the number of quantizers (bM) be an integer. As a distortion measure, suppose that we know for every coefficient the expression for the mean squared error $\sigma_{q_i}^2 = \mathrm{E}\{|Y_i - \hat{Y}_i|^2\}$, given by the optimum scalar quantization of every $Y_i(n)$ and using b_i bits for each coefficient. The function $\sigma_{q_i}^2$ tells us how we can reduce the average distortion by increasing the resolution, in other words, we have a "price-performance" trade-off available. The "price" is the number of bits we allocate and the "performance" is determined by the mean distortion that results.

Before starting, let us suppose that all the random variables are identically distributed ($[Y_0(Mn) \cdots Y_{M-1}(Mn-M+1)]$), that we know all the statistical properties of each the random process and that the high resolution hypothesis developed in Section 2.6 is acomplished. The bit allocation problem consists in determining the optimal values of $b_1, b_2, \cdots, b_k$ subject to a fixed given quota, bM, of available bits, in order to minimize the power of the total distortion given by

$$D = \frac{1}{M} \sum_{i=0}^{M-1} \sigma_{q_i}^2(b_i) \tag{2.26}$$

under the constraint that

$$\sum_{i=0}^{M-1} b_i \leq bM \tag{2.27}$$

According to the high resolution property (see Appendix A) we can write

$$\sigma_{q_i}^2 = h_i \sigma_{Y_i}^2 2^{-2b_i} \tag{2.28}$$

where h_i is a constant that only depends of the PDF of each coefficient Y_i (see Appendix A). In order to solve this problem, we make an assumption: the transform coefficients have a Gaussian PDF, thus all the constants h_i will be identical and they will not be modified during the linear transformation. If the reader is interested, the general solution to this problem is presented in [29]. Therefore, the goal of the optimal bit allocation is to minimize the expression

$$D = \frac{h}{M} \sum_{i=0}^{M-1} \sigma_{Y_i}^2 2^{-2b_i} \qquad \text{where } h = \frac{\sqrt{3}}{2}\pi \tag{2.29}$$

The result that takes into account the constraint imposed by Equation 2.27 and minimizes the distortion is given by (see Appendix B)

$$b_i = b + \frac{1}{2}\log_2 \frac{\sigma_{Y_i}^2}{\rho^2} \tag{2.30}$$

Where b is the average number of bits per coefficient and ρ is the geometric mean of the coefficients variance

$$\rho^2 = \left(\prod_{i=0}^{M-1} \sigma_{Y_i}^2 \right)^{\frac{1}{M}} \tag{2.31}$$

The minimum overall distortion attained with this solution is given by

$$D = h \left(\prod_{i=0}^{M-1} \sigma_{Y_i}^2 \right)^{\frac{1}{M}} 2^{-2b} \tag{2.32}$$

The overall distortion is the same that would be achieved if each random variable had variance ρ^2 and each quantizer were allocated b bits. Note that the optimal bit allocation is independent of the PDF of the random variables. Equation 2.30 shows that the allocation to quantizer i exceeds the average allocation b if the variance of Y_i is greater than the geometric average ρ^2 of the variances. Conversely, if σ_i^2 is less than ρ^2 the bit allocation is less than the average allocation and if σ_i^2 is sufficiently small, the allocation can be even negative. If we see in a different way Equation 2.32 it means that, the optimum bit allocation solution consists in equalizing the variances of

the quantization noise sources. This simple solution offers a very useful tool for efficient bit allocation that is widely used in signal compression schemes [57, 29].

However, several remarks concerning this bit allocation procedure ought to be highlighted here:

1. The resulting b_i may not always be integers. In that case, they must be rounded to integers.
2. Some b_i may be negative, in which case they must be set to zero.
3. The two previous remarks would lead to a reallocation procedure in order to satisfy the constraint of a constant average bit rate of b bits.
4. Small values of b_i violate the noise model assumptions (white noise, uncorrelated and uniformly distributed). In that case, the derived results may not be valid. For instance, if $\sigma^2_{Y_i} = 0$ for any i, then expression 2.32 becomes zero.

2.8 Huffman Coding

Traditionally, the last step in transform/subband coding is entropy coding. After the output coefficients have been quantized, they take values drawn from a finit set a_i. The goal of entropy coding is to find a mapping of the symbols a_i to a new set of symbols b_i, such that the average bits per symbol are minimized [29, 41, 70, 77, 82]. If the quantized coefficients have a stationary PDF, as we have previously assumed, fixed mapping techniques such Huffman coding can be used [70, 82]. If the statistics of the coefficients evolve over time, more sophisticated adaptive methods, such as adaptive Huffman coding or arithmetic coding can be used. For the present work we decided to employ Huffman codes as the entropy coding method.

The Huffman codes are prefix codes and are optimum for a given model (set of probabilities). The Huffman procedure is based on two observations regarding optimum prefix codes:

1. In an optimum code, symbols that occur more frequently, i.e., have a higher probability of occurrence, will have shorter codewords than symbols that occur less frequently.
2. In an optimum code, the two symbols that occur least frequently will have the same length.

The Huffman procedure is obtained by adding a simply requirement to these two observations, the codewords of the lowest probability symbols differ only in in the last bit. That is, if α and β are the least probable symbols in the alphabet, and if the codeword for α was $m * 0$ the codeword for β would be $m * 1$. Where m stands for a string of ones and zeros and "$*$" denotes concatenation. One of the ways of building a Huffman code is to use the fact that such codes, due to the property of being a prefix code, can be represented as a binary tree in which the external nodes or leaves correspond to the symbols [29, 70]. The Huffman code for any symbol can be obtained by traversing the tree corresponding to such symbol adding a "0" every time the traversal takes us over a predetermined direction (for example, left) and a "1" every time the traversal takes us over the opposite direction. We build the binary tree starting with the leaf nodes. We know that the codewords for symbols with the smallest probability are identical except for the last bit.

Chapter 3

Wavelets Fundamentals

In recent years, few subjects have attracted as much attention in engineering and mathemathics as wavelets have. There are various reasons for this: the topic itself is very old and very new. Indeed, wavelets can be traced back to 1910 and Alfred Haar, though they were not called wavelets at that time. Its underlying idea of linear signals expansion and its connection with the Fourier-based techniques date back to the early part of 19th century and the great work done by the French scientist Joseph Fourier. Although seeming dormant most of the part of this century, thanks to various mathematical advances (such as harmonic analysis, which has many things in common with wavelet analysis) wavelets made their *debut* in the mid-1980's. Wavelet based techniques have been used in signal compression for almost two decades, sometimes under a different name (subband coding) and somewhat in disguise. The wavelets subject area is connected to older ideas in many other fields like pure and applied mathematics, physics, computer science and engineering. If the reader is interested about wavelets history, [23] provides a good start up. Wavelets is a field that can be seen as a tree with roots reaching deeply and in many directions. Obviously, in this chapter we do not try to cover the entire field, but the elements that will be necessary to accomplish the goals established in Chapter 1. Perhaps, in some cases the reader will notice a lack of formality or clarity, but references providing a deeper coverage of the subject are provided.

3.1 Fourier Analysis

In our common life, our attention is clearly attracted by transients and movements as opossed to stationary stimuli, which we soon ignore. Concentrating on transients is probably a strategy for selecting important information from the overwhelming amount of data recorded by our senses. In spite of that, classical signal processing has devoted most of its efforts to the design of time invariant operators, that modify stationary signal properties. This has led to the hegemony of the Fourier transform, which we will briefly present in this section.

The Fourier transform rules over linear time-invariant signal processing because sinusoidal waves $e^{i\omega t}$ are eigenvectors of linear time invariant operators. A linear time-invariant operator L is entirely specified by the eigenvalues $H(\omega)$:

$$\forall \omega \in \mathbb{R}\ ,\ Le^{i\omega t} = H(\omega)e^{i\omega t} \tag{3.1}$$

To calculate $Lg(t)$, a signal $g(t)$ is decomposed as a sum of sinusoidal eigenvectors $\{e^{i\omega t}\}_{\omega\in\mathbb{R}}$:

$$g(t) = \int_{-\infty}^{+\infty} G(\omega)e^{i\omega t}\, d\omega \tag{3.2}$$

If $g(t)$ has finite energy, the theory of fourier integrals [51] proves that the amplitud $G(\omega)$ of each sinusoidal wave $e^{i\omega t}$ is the Fourier transform of $g(t)$.

$$G(\omega) = \int_{-\infty}^{+\infty} g(t)e^{-i\omega t}\, dt \tag{3.3}$$

Applying the operator L to $g(t)$ in Equation 3.2 and inserting the eigenvector expression in Equation 3.1 gives

$$Lg(t) = \int_{-\infty}^{+\infty} H(\omega)G(\omega)e^{i\omega t} d\omega \tag{3.4}$$

The operator L amplifies or attenuates the frequencies of each sinusoidal component of $e^{i\omega t}$ by $H(\omega)$. It is a frequency *filtering* of $g(t)$. As long as we are satisfied with linear time-invariant operators, the Fourier transform provides simple answers to most questions. Its richness makes it suitable for a wide range of applications such as signal transmission or stationary signal processing. However if we are interested in transient phenomena (a musical instrument playing at a particular time, a ball in one of the corners of an image), the Fourier transform becomes an inadequate and cumbersome tool.

3.1.1 Windowed Fourier Transform

Heisenberg's uncertainty principle states that the energy spread of a function and its Fourier transform cannot be simultaneously arbitrarily small. Motivated by quantum mechanics, in 1946 the Hungarian Nobel prize winner physicist Dennis Gabor defined elementary time-frequency *atoms* as waveforms that have minimal spread in the time-frequency plane. To measure time-frequency "information" content, he suggested the decomposition of signals over these elementary atomic waveforms. By showing that such decompositions are related to the human sensitivity of sounds, and that they exhibit important structures in speech and musical recordings, Gabor demonstrated the importance of localized time-frequency signal processing.

Gabor atoms are constructed by translating in time and frequency a time window, which we will call w

$$w_{u,\xi}(t) = w(t-u)e^{i\xi t} \tag{3.5}$$

In the time domain, the energy of $w_{u,\xi}(t)$ is concentrated in the neighborhood of u over an interval of size σ_t, known as *time resolution*[1] Its Fourier transform is a translation by ξ of the Fourier transform $W(\omega)$ of $w(t)$

$$W_{u,\xi}(\omega) = W(\omega-\xi)e^{-iu(\omega-\xi)} \tag{3.6}$$

The energy of $W_{u,\xi}(\omega)$ is therefore localized near the frequency ξ over an interval of size σ_ω, that is also known as *frequency resolution*[2] This interval measures the domain where $W_{u,\xi}$ is non neglegible. In a time frequency plane (t,ω) the energy spread of the atom $w_{u,\xi}(t)$ is symbolically represented by the Heisenberg rectangle depicted in Figure 3.1-(a). This rectangle is centered at (u,ξ) and has a time width σ_t and frequency width σ_ω. The uncertainty principle proves that its area satifies

$$\sigma_t\sigma_\omega \geq \frac{1}{2} \tag{3.7}$$

For a detailed demonstration of the uncertainty principle the reader is referred to [16, 51, 62, 77, 82]. This area is minimum when the window w is Gaussian, in which case

[1]The time resolution is defined as the capacity to discriminate two pulses in time only if they are more than σ_t apart. Where, $\sigma_t^2 = \frac{\int t^2|w(t)|^2dt}{\int|w(t)|^2dt}$

[2]The frequency resolution is defined as the capacity to discriminate two sinusoids in frequency only if they are more than σ_ω apart. Where, $\sigma_\omega^2 = \frac{\int f^2|W(\omega)|^2d\omega}{\int|W(\omega)|^2d\omega}$

the atoms $w_{u,\xi}$ are called *Gabor functions.* The windowed Fourier transform[3] defined by Gabor correlates a signal g with each atom $w_{u,\xi}$:

$$Sg(u,\xi) = \int_{-\infty}^{+\infty} g(t) w_{u,\xi}(t) dt = \int_{-\infty}^{+\infty} g(t) w(t-u) e^{-i\xi t} dt \tag{3.8}$$

It is a Fourier integral that is localized in the neighborhood of u by the window $w(t-u)$. The transform $Sg(u,\xi)$ depends only on the values of $g(t)$ and $G(\omega)$ in the time and frequency neighborhoods, where the energies of $w_{u,\xi}(t)$ and $W_{u,\xi}(\omega)$ are concentrated. Gabor interprets this as a "quantum information" over the time-frequency rectangle presented in Figure 3.1-(a).

When listening to music, we perceive sounds that have a frequency that varies in time. In measuring time varying harmonics the windowed Fourier transform becomes a very important tool in audio and speech applications. A spectral line of g creates high amplitude windowed Fourier coefficients $Sg(u,\xi)$ at frequencies $\xi(u)$ that depend on time u. The time evolution of such spectral components is therefore analyzed by following the location of large amplitude coefficients. A real life application of the previously described windowed Fourier transform in speech compression is presented in [54].

3.2 Continuous Wavelet transform

In the late 1970's, Jean Morlet, a French geophysical engineer working for the oil company Elf Acquitaine, came up with an alternative for the STFT. In reflection seismology, Morlet knew that modulated pulses sent underground have a duration that is too long at high frequencies to separate the returns of fine, closely-spaced layers. The signals that he wanted to analyze had different features in time and frequency that he wanted to separate. The traditional STFT were not useful for Morlet, in order to gain time resolution for the high frequency transients, he could choose to do a wide-band STFT; on the other hand, he also wanted good frequency resolution for the low frequency components and that would require a narrow-band STFT. But he wanted to achieve both aims simultaneously (i.e., with one single transform). Morlet came up with the idea to generate the transform functions using a signal (he used a Gaussian windowed cosine wave) and compressing it in time to obtain a higher frequency function. In order to investigate what happened at different times, these functions were also

[3]The windowd Fourier transform is also known as the Short Time Fourier Transform (STFT).

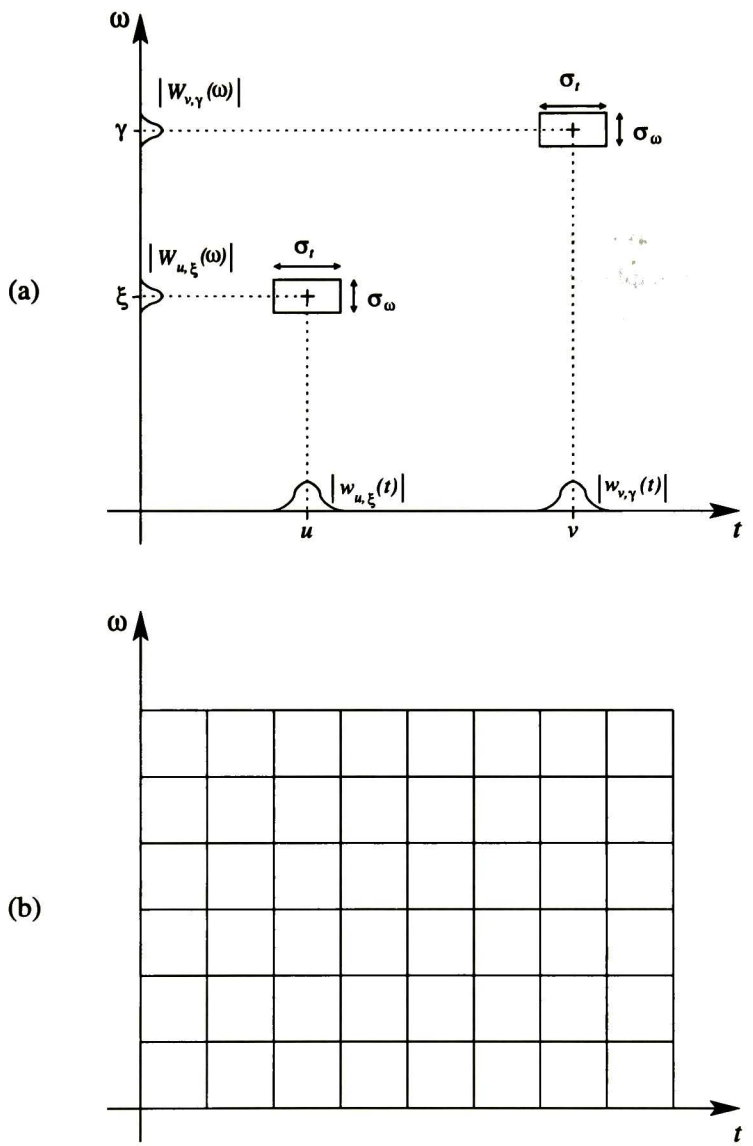


Figure 3.1: *(a)Time-frequency boxes (Heisenberg rectangles) representing the energy spread of two Gabor atoms. (b) STFT tiling of the time-frequency plane.*

shifted in time. The transform function depended on two parameters: the time location and their degree of compression. Thus, the transform waveforms are simply obtained by scaling and translating in time a single waveform that Morlet called *une ondelette* or a wavelet [23]. This is how the Wavelet Transform (WT) was born, and the resolution limitations of the STFT were overcome by the WT.

Now, the WT lets us vary the time and frequency resolutions (σ_t and σ_ω respectively) in the time-frequency plane in order to obtain a *multiresolution analysis.* Intuitively, we can see the analysis as a filter bank structure. Under the restriction imposed by the Heisenberg's principle, we can increase the time resolution with the central frequency of the analysis filters. If we impose that σ_ω must be proportional to ω ($\frac{\sigma_\omega}{\omega} = \kappa$, where κ is a constant) the analysis filter bank is composed of bandpass filters with constant relative bandwidth. It means that, the frequency responses of the analysis filters instead of being regularly spaced over the frequency axis (like in the STFT case), they are regularly spread in a logarithmic scale, as depticed in Figure 3.2-(b). This property of the wavelet transform will be very useful later in this chapter for modeling the response of human aditory system. It is important to remark that the constant relative bandwidth analysis satisfies the Heisenberg's uncertainty principle, but now the time resolution becomes arbitrarily good at high frequencies, while the frequency resolution becomes arbitrarily good at low frequencies (see Figure 3.3-(a)). This kind of analysis works best if the signal is formed by high frequency components of short duration and low frequency components of long duration, which is generally the case for speech and audio signals [69].

The Continuous Wavelet Transform (CWT) exactly follows the previously mentioned ideas. To analize structures of very different sizes, it is necessary to use time-frequency atoms with different time supports. The CWT decomposes signals over scaled[4] and translated wavelets. Mathematically, we define a wavelet as a function $\psi \in \mathbf{L}^2(\mathbb{R})$ [5] with a zero average

$$\int_{-\infty}^{+\infty} \psi(t)\, dt = 0 \tag{3.9}$$

It is normalized $\|\psi\| = 1$, and centered in the neighborhood of $t = 0$. As mentioned, a family of time frequency atoms is obtained by scaling ψ by s and translating it by u,

[4] All the wavelets are defined as stretched or compressed versions of the same prototype.

[5] We say that $x(t)$ belongs to the finite energy functions set $\mathbf{L}^2(\mathbb{R})$ if $\int_{-\infty}^{+\infty} |x(t)|^2\, dt < +\infty$.

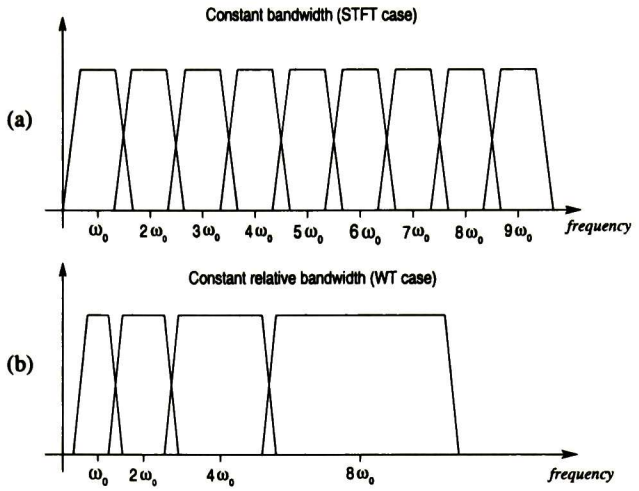


Figure 3.2: *Division of the frequency domain (a) for the STFT and (b) for the WT.*

where $s, u \in \mathbb{R}$ and $s \neq 0$

$$\psi_{u,s}(t) = \frac{1}{\sqrt{|s|}}\psi\left(\frac{t-u}{s}\right) \tag{3.10}$$

These atoms remain normalized: $\|\psi_{u,s}\| = 1$. The CWT of a function $g \in \mathbf{L}^2(\mathbb{R})$ at time u and scale s is defined as

$$Wg(u,s) = \langle g, \psi_{u,s}\rangle = \int_{-\infty}^{+\infty} g(t)\frac{1}{\sqrt{|s|}}\psi^*\left(\frac{t-u}{s}\right)dt \tag{3.11}$$

where α^* stands as the complex conjugate for α. In accordance with the linear filtering case, the CWT can be seen also as a convolution product

$$Wg(u,s) = \int_{-\infty}^{+\infty} g(t)\frac{1}{\sqrt{|s|}}\psi^*\left(\frac{t-u}{s}\right)dt = g \star \tilde{\psi}_s(u) \tag{3.12}$$

where

$$\tilde{\psi}_s = \frac{1}{\sqrt{|s|}}\psi^*\left(\frac{-t}{s}\right) \tag{3.13}$$

The Fourier transform of $\tilde{\psi}(t)$ is given by

$$\tilde{\Psi}_s(\omega) = \sqrt{|s|}\Psi^*(s\omega) \tag{3.14}$$

Since $\tilde{\Psi}(0) = \int_{-\infty}^{+\infty} \psi(t)\, dt = 0$, it appears that Ψ is the transfer function of a band pass filter. Thus, we can say that the convolution of Equation 3.12 computes the wavelet transform with dilated bandpass filters.

Like the STFT, the WT can measure the time evolution of frequency transients. This requires using a complex analytic wavelet, which can separate amplitude and phase components. In contrast, real wavelets are often used to detect sharp signal transitions. For the present work, we will focus our attention only in real wavelets. If the reader is interested in analytic wavelets, useful information can be found in [51]. In the time frequency plane, a wavelet atom $\psi_{u,s}$ is symbolically represented by a rectangle centered at $(u, \eta/s)$, where η stands for the center frequency of $\Psi(\omega)$. The time and frequency spread are respectively proportional to s and $1/s$. When s varies, the height and width of the rectangle change, but its area remains constant, as illustrated in Figure 3.3-(a).

A real wavelet transform is complete and maintains an energy conservation, as long as the wavelet satisfies a weak admissibility condition, specified by the Theorem 3.1 [51, 82, 77]. This theorem was proved independently by the mathematician Calderón in 1964, and years later by Grossman and Morlet, who were not aware of the previous work. The demonstration of this theorem is provided in Appendix C.

Theorem 3.1 (CALDERÓN-GROSSMAN-MORLET) *Let $\psi \in \mathbf{L}^2(\mathbb{R})$ be a real function such that*

$$C_\psi = \int_0^{+\infty} \frac{|\Psi(\omega)|^2}{\omega}\, d\omega < +\infty, \tag{3.15}$$

Any $g \in \mathbf{L}^2(\mathbb{R})$ satisfies

$$g(t) = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} Wg(u,s) \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-u}{s}\right) du \frac{ds}{s^2} \tag{3.16}$$

and

$$\int_{-\infty}^{+\infty} |g(t)|^2\, dt = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} |Wg(u,s)|^2\, du\, \frac{ds}{s^2} \tag{3.17}$$

Like the STFT, the WT of a signal, $Wg(u,s)$, is a two-dimensional representation of a one-dimension signal g. Thus, the WT is a redundant representation whose redundancy is characterized by a reproducing kernel equation. Inserting the reconstruction

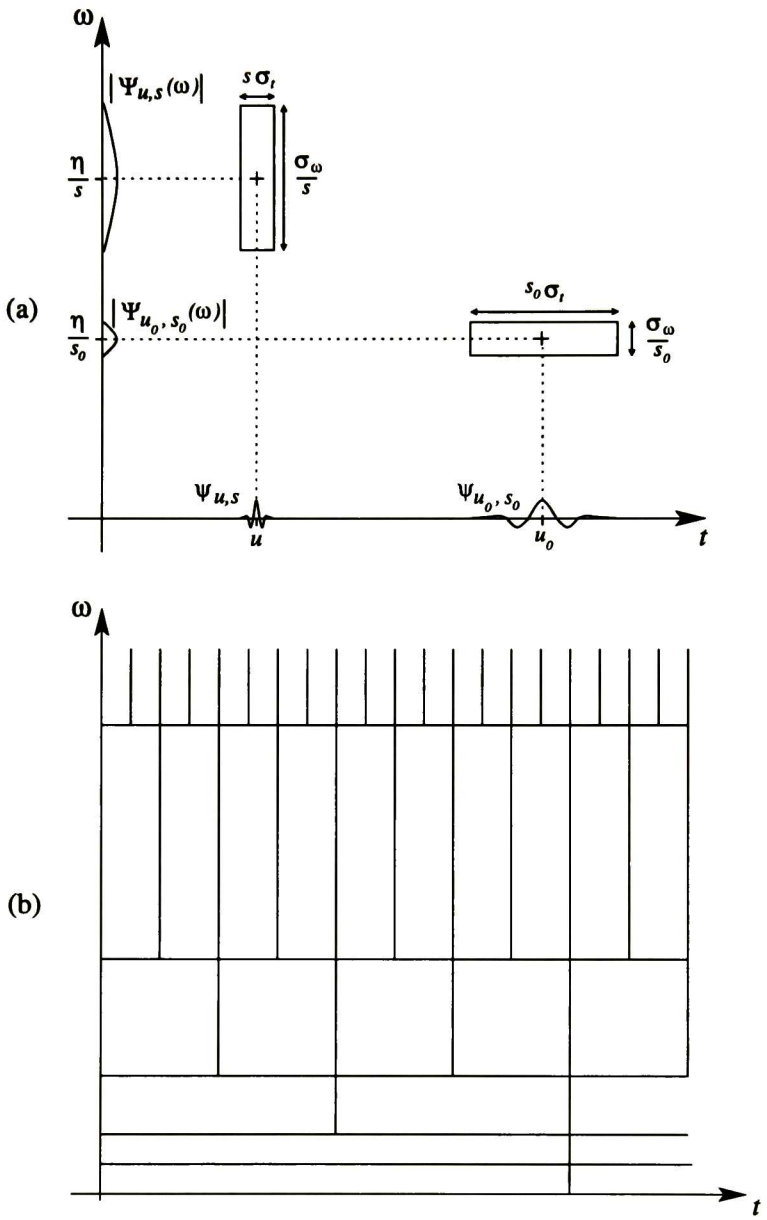


Figure 3.3: *(a) Time-frequency boxes of two wavelets $\psi_{u,s}$ and ψ_{u_0,s_0}. When the scale s decreases the time support is reduced, but the frequency spread increases and covers an interval that is shifted towards the high frequencies. (b) Wavelet transform tiling of the time-frequency plane.*

formula (Equation 3.16) into the definition of the WT (Equation 3.11) yields

$$Wg(u_0, s_0) = \int_{-\infty}^{+\infty} \left(\frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} Wg(u,s) \frac{1}{\sqrt{|s|}} \psi \left(\frac{t-u}{s} \right) du \frac{ds}{s^2} \right) \psi^*_{u_0,s_0}(t)\, dt \tag{3.18}$$

Interchanging these integrals gives

$$Wg(u_0, s_0) = \frac{1}{C_\psi} \int_{-\infty}^{+\infty} K(u, u_0, s, s_0) Wg(u,s)\, du \frac{ds}{s^2} \tag{3.19}$$

where

$$K(u, u_0, s, s_0) = \langle \psi_{u,s}, \psi_{u_0,s_0} \rangle \tag{3.20}$$

The reproducing kernel $K(u, u_0, s, s_0)$ measures the correlation of two wavelets $\psi_{u,s}$ and ψ_{u_0,s_0} [51, 82].

When $Wg(u,s)$ is known only for $s < s_0$, to recover g we need complementary information corresponding to $Wg(u,s)$ for $s > s_0$. This is obtained by introducing a *scaling function* ϕ that is an aggregation of wavelets at scales larger than 1. The modulus of its Fourier transform is defined by

$$|\Phi(\omega)|^2 = \int_1^{+\infty} |\Psi(s\omega)|^2 \frac{ds}{s} = \int_\omega^{+\infty} \frac{|\Psi(\xi)|^2}{\xi} d\xi \tag{3.21}$$

with $\xi = s\omega$. The complex phase of $\Phi(\omega)$ can be arbitrarily chosen [51]. In the next section, we will verify that $\|\phi\| = 1$ and from Equation 3.15 that [51]

$$\lim_{\omega \to 0} |\Phi(\omega)|^2 = C_\psi \tag{3.22}$$

The scaling function can thus be interpreted as the impulse response of a low-pass filter. The low frequency approximation of g at scale s is

$$Lg(u,s) = \left\langle g(t), \frac{1}{\sqrt{|s|}} \phi \left(\frac{t-u}{s} \right) \right\rangle \tag{3.23}$$

3.3 Discrete Parameter Wavelet transform

As we have previously seen, in Equation 3.11 the parameters u, s are continuous variables and, as mentioned, there exists redundancy in the CWT representation of a signal. That means that there is no need to compute the CWT for all possible values of

u, s. Aditionally, it is of practical necessity that u, s take only a finite number of values. In a different way, we could say that the wavelet family $\psi_{u,s}(t)$ behaves in the analysis and synthesis of functions like an orthonormal basis of $\mathbf{L}^2(\mathbb{R})$. If we appropriately discretize the time-scale parameters u, s and *depending* on the type of $\psi(t)$, we can obtain a finite basis of functions that achieves perfect reconstruction. Though, sometimes duals are required [69, 16, 68, 51, 77, 82]. The time-scale parameters can take any values, but a special case occurs when u, s are samples of a dyadic grid, when certain $\psi(t)$ can produce orthonormal $\psi_{u,s}(t) = \frac{1}{\sqrt{|s|}}\psi\left(\frac{t-u}{s}\right)$, with u, s discrete. Consequently, a signal $x(t)$ can be exactly synthesized as a weighted sum of these orthonormal basis functions, as we shall see later.

When time-scale parameters u, s are discrete and given by

$$s = s_0^m, \qquad u = n\, u_0\, s_0^m, \quad \text{where } m, n \in \mathbb{Z} \tag{3.24}$$

the discrete parameter wavelet transform (DPWT) is defined as

$$c_{m,n} = \int_{-\infty}^{+\infty} x(t)\psi_{m,n}(t)\, dt \tag{3.25}$$

where the corresponding wavelets are

$$\psi_{m,n}(t) = s_0^{-\frac{m}{2}}\psi(s_0^{-m}t - n\, u_0) \tag{3.26}$$

From Equation 3.26, we can see that $\psi_{0,0}(t) = \psi(t)$, and that s_0, t_0 are constants that determine the sampling intervals. It is important to note that we are still working with continuous functions of time. An analogy of Equation 3.25 is the following [69]: first one chooses the magnification, that is, s_0^{-m} Then one moves to the chosen location. If one looks at very small details, the magnification is large and corresponds to m negative and large. Then $s_0^m\, n$ corresponds to small steps, which are used to catch small details. From this previous coments, it is necessary to analize different circumstances in which the type of $\psi(t)$ and the sampling intervals for u, s permit to achieve perfect reconstruction

$$x(t) = C \sum_m \sum_n c_{m,n}\, \psi_{m,n}(t) \tag{3.27}$$

where C is a constant that only depends on $\psi(t)$.

Let us suppose that there exists oversampling. Then, Equation 3.27 will not only hold, but moreover it is possible to have a non-unique representation of $x(t)$ with respect

to the same mother wavelet $\psi(t)$. Let $\tilde{c}_{m,n}$ be the DPWT due to $\tilde{u}, \tilde{s}$ and $\breve{c}_{m,n}$ due to $\breve{u}, \breve{s}$, we have that

$$\begin{aligned} x(t) &= \tilde{C} \sum_m \sum_n \tilde{c}_{m,n}\, \tilde{\psi}_{m,n}(t) \\ &= \breve{C} \sum_m \sum_n \breve{c}_{m,n}\, \breve{\psi}_{m,n}(t) \end{aligned} \tag{3.28}$$

where $\tilde{\psi}_{m,n}(t)$ and $\breve{\psi}_{m,n}(t)$ are as defined in Equation 3.26. Redundancy from oversampling permits two distinct sets of wavelets, derived from the same mother wavelet, that give exact but distinct synthesis of $x(t)$.

If the sampling is sparse, Equation 3.27 is not valid. In this case, the theory of frames[6] [22, 21, 68, 16, 51, 82] describes the reconstruction conditions for the DPWT. The objective is to develop conditions for $\psi_{m,n}(t)$ that permit perfect reconstruction, via duals if necessary. For convenience, let $\langle \psi_{m,n}(t), \psi_{m,n}(t) \rangle = 1$, and let

$$c_{m,n}(t) = \langle x(t), \psi_{m,n}(t) \rangle = \int_{-\infty}^{+\infty} x(t) \psi_{m,n}(t)\, dt \tag{3.29}$$

Now, let us suppose that the $\psi_{m,n}$ do not constitute an orthonormal basis, i.e., Equation 3.27 is no longer valid. In this case, the conditions on the $\psi_{m,n}(t)$ that would allow the alternate reconstruction

$$x(t) = \sum_m \sum_n c_{m,n} \tilde{\psi}_{m,n}(t) \tag{3.30}$$

are: the $\psi_{m,n}(t)$ must be elements of a frame and the $\tilde{\psi}_{m,n}(t)$ must be elements of the dual frame. If the $\psi_{m,n}(t)$ form a frame, they must obey the inequality

$$A \|x(t)\|^2 \leq \sum_m \sum_n \|\langle x(t), \psi_{m,n}(t) \rangle\|^2 \leq B \|x(t)\|^2 \tag{3.31}$$

with $0 < A \leq B < +\infty$. The constants A, B are the frame bounds, and are dependent only on the $\psi_{m,n}(t)$, additionally $x(t) \in \mathbf{L}^2(\mathbb{R})$.

The bounds in Equation 3.31 ensure that the reconstruction is numerically stable, in the sense that if $\langle s_1(t), \psi_{m,n}(t) \rangle$ and $\langle s_2(t), \psi_{m,n}(t) \rangle$ are approximately equal, then necessarily the functions $s_1(t)$ and $s_2(t)$ should be approximately equal as well. It is possible to select u_0, s_0 so that $A \approx B$. In that case the formula for duals is [16, 82]

$$x(t) \approx \frac{2}{A+B} \sum_m \sum_n \langle x(t), \psi_{m,n}(t) \rangle\, \psi_{m,n}(t) \tag{3.32}$$

[6] The theory of frames provides the representation of a signal in terms of a set of basis functions that are not necessarely orthonormal. Such functions must span the signal space.

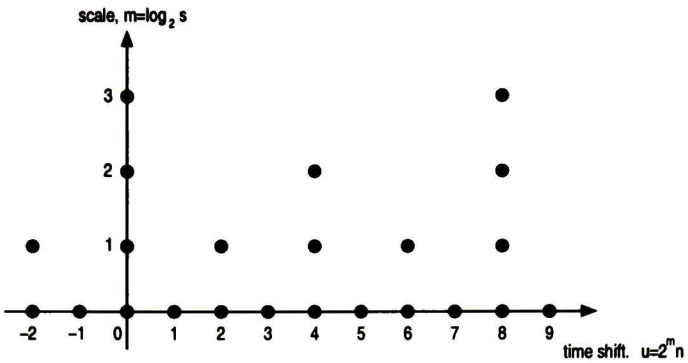


Figure 3.4: *The dyadic sampling grid in the time-scale plane. Each dot corresponds to a wavelet basis function* $\psi_{m,n}(t)$.

The closer is A to B the better is the approximation. In general, the computation of A and B is difficult. Estimates of the frame bounds for certain special wavelets are available in [22, 51]. There exists no procedure for selecting the time-scale parameters that leads to $A \approx B$ for any $\psi(t)$. By definition [16, 51, 77, 82], a frame is tight if $A = B$, then

$$x(t) = \frac{1}{A} \sum_m \sum_n \langle x(t), \psi_{m,n}(t) \rangle \, \psi_{m,n}(t) \tag{3.33}$$

Additionally, if $A = B = 1$ the basis functions are orthonormal.

As mentioned, a practical sampling scheme is $s = 2^m$, $u = n\,2^m$, i.e., $s_0 = 2$, $u_0 = 1$, then

$$\psi_{m,n}(t) = 2^{-\frac{m}{2}} \psi(2^{-m}t - n) \tag{3.34}$$

with this octave time scaling and a dyadic translation, the sampled values of u, s are as depicted in Figure 3.4. Since the Fourier transform of $\frac{1}{\sqrt{|s|}}\psi(st)$ is $\frac{1}{s\sqrt{|s|}}\psi\left(\frac{\omega}{s}\right)$, the center frequency and bandwidth of a wavelet are both scaled by $1/s$ for a time scaling of s. Thus, as mentioned in the previous section (see Figure 3.2), the relative bandwidht of all the derived wavelets is constant

$$Q = \frac{\text{center frequency}}{\text{bandwidth}} = constant \tag{3.35}$$

This analysis is also known as "constant-Q analysis".

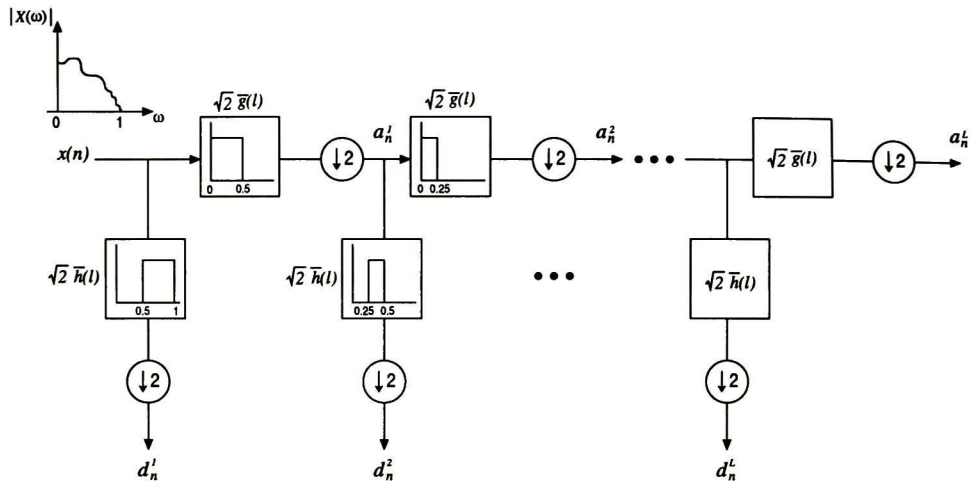


Figure 3.5: *Mallat's multiresolution analysis scheme.*

The search for discrete parameters orthonormal wavelets has been a subject of intense research [21] due to their obvious desirable properties and potential for a wide range of applications. In the signal processing literature [16, 51, 77, 82] it is known that the only non-trivial ones are those of Daubechies [21] constructed by recursion. There exists no analytic expression for them.

3.4 Multiresolution Analysis

The idea of multiresolution analysis (MRA) is very related to subband decomposition and coding, where for coding efficiency, a signal is divided into a set of frequency bands. In Figure 3.5, the sequence $x(n)$, for $n = 0, 1, \cdots, N-1$, is bandlimited from 0 to 1. With lowpass ($\bar{g}(l)$), highpass ($\bar{h}(l)$) filters and decimation, the scheme presented in Figure 3.5 decomposes $x(n)$ into subband components d_n^1, d_n^2, $\cdots$, d_n^L and a_n^L, representing the coarsest, the dc component, as $L \to \infty$. Assuming the absence of quantizers, Figure 3.6 illustrates the signal reconstruction from a subband decomposition. The symbols $\uparrow$ 2 and $\downarrow$ 2 stand for subsampling and upsampling by two, respectively. Subsampling by two is dropping every other sample in the sequence and renumbering the sequence. Upsampling by two is inserting zeros between the samples of a sequence and then renumbering. The scheme presented in Figure 3.5 produces frequency bands in contiguous octaves, as depicted in Figure 3.2-b. This octave division

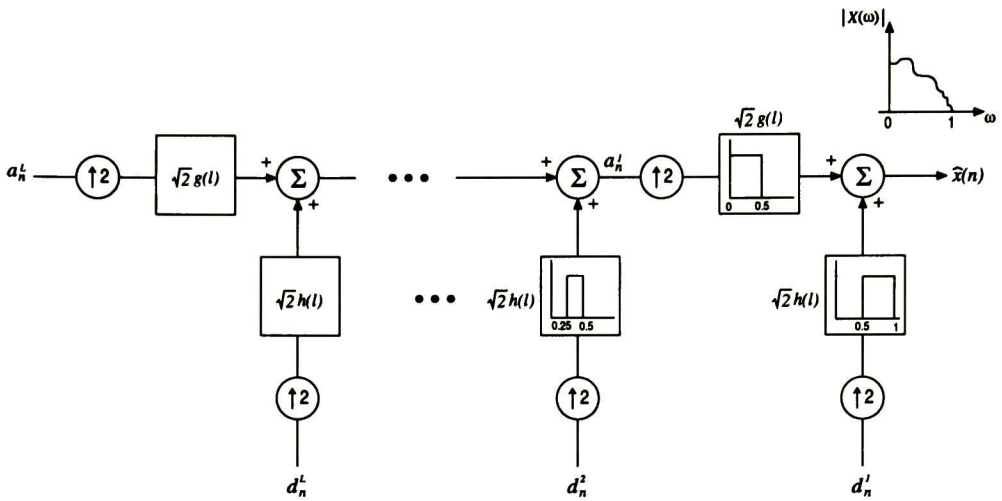


Figure 3.6: *Reconstruction from subband decomposition.*

corresponds to the MRA scheme of Mallat [52]. Besides giving a constant-Q analysis it is also critically sampled[7] The resolution of a signal is a qualitative term related to its frequency content For a low pass signal, the lower its frequency content, the smaller is its resolution. The frequency content or resolution of the *approximation* sequence a_n^m, output of the lowpass filters, decreases as m increases, until it reaches the dc component a_n^∞ of lowest resolution. The highpass (or bandpass) filters output d_n^m are called the high frequency *detail* or *difference*, which is the difference in resolution between a_n^{m-1} and a_n^m

Downsampling by two a sequence is equivalent to compressing it by two. However, since lowpass filtering must precede[8] downsampling, it also halves the signal bandwidth and reduces its resolution. The DPWT with dyadic sampling on the u, s parameters also scales a sequence successively by two, thus we can establish a link between MRA and the DPWT of a signal [16]. The d_n^m of $x(t)$ are the DPWT coefficients, $c_{m,n}$, defined by Equation 3.25, with respect to an orthonormal, compactly supported set of wavelets, $\psi_{m,n}(t)$. It is very important to remark that lowpass and highpass filter coefficients determine the $\psi(t)$, not the reverse [21, 16].

[7]The sum of data rates at all outputs (d_n^1, d_n^2, $\cdots$, d_n^L and a_n^L) equals the input data rate.

[8]This is not necessarily true. [77, 80] provide the *noble indentities* which allow the exchange of the up-/downsampling operators with the filters to permit a highly efficient implementation.

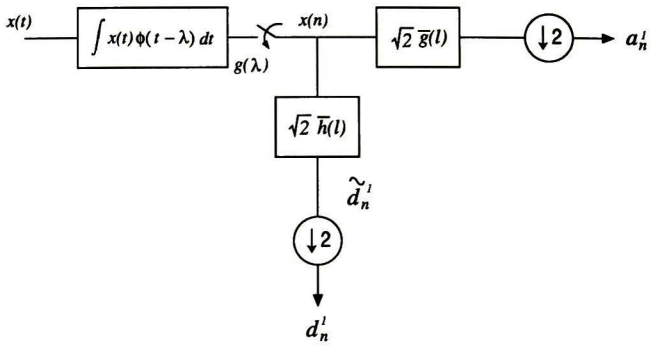


Figure 3.7: *The first stage decomposition of the MRA.*

3.4.1 Relationship Between Multiresolution Analysis and the DPWT

To establish a link between MRA and the DPWT, we must analyze the first stage of the MRA decomposition of Figure 3.5, presented in Figure 3.7. Let $x(n)$ be the samples of $g(\lambda)$, which is the inner product of $x(t)$ and $\phi(t-\lambda)$. This inner product puts the $x(n)$ into the proper subspace, spanned by $\phi(t-n)$, so the d_n^m are indeed the coefficients of the DPWT, $c_{m,n}$, of $x(t)$.

$$x(n) = g(\lambda = Tn) = \int_{-\infty}^{+\infty} x(t)\phi(t-n)\,dt \tag{3.36}$$

where $T = 1$. We will demonstrate that if Equation 3.36 holds[9] then $d_n^m = c_{m,n}$. From Figure 3.7 we see that

$$\tilde{d}_n^1 = \sqrt{2}\sum_l \bar{h}(l)\,x(n-l) \tag{3.37}$$

After subsampling and renumbering the sequence

$$d_n^1 = \tilde{d}_{2n+p-1}^1 = \sqrt{2}\sum_l \bar{h}(l)x(2n+p-1-l) \tag{3.38}$$

where the delay $p-1$ is the filter order. Using Equation 3.36 with Equation 3.38 results in

$$d_n^1 = \sqrt{2}\sum_l \bar{h}(l)\int_{-\infty}^{+\infty} x(t)\phi(t-2n-p+1+l)\,dt \tag{3.39}$$

[9]In practice this is not true because $x(n)$ usually comes from a direct sampling of $x(t)$, already a lowpass signal. Thus, Equation 3.36 is not valid and $d_n^m \neq c_{m,n}$ of $x(t)$. Instead, the d_n^m equal the $c_{m,n}$ of function $\tilde{x}(t) = \sum_n x(n)\phi(t-n)$. In spite of this, Equation 3.36 is a close approximation under certain restrictions [16].

By definition of the DPWT (Equation 3.25), for $m = 1$ with respect to a wavelet $\psi(t)$ we have

$$c_{1,n} = \int_{-\infty}^{+\infty} x(t)\psi_{1,n}(t)\, dt \tag{3.40}$$

and from Equation 3.26

$$\psi_{1,n}(t) = \frac{1}{\sqrt{2}}\psi\left(\frac{t}{2} - n\right) \tag{3.41}$$

The condition for d_n^1 comes from inserting Equation 3.41 into Equation 3.40 and equating the result with -3.39

$$\frac{1}{\sqrt{2}}\psi\left(\frac{t}{2} - n\right) = \sqrt{2}\sum_l \bar{h}(l)\phi(t - 2n - p + 1 + l) \tag{3.42}$$

Letting $t/2 - n = t$ gives

$$\psi(t) = 2\sqrt{2}\sum_l \bar{h}(l)\phi(2t - p + 1 + l) \tag{3.43}$$

Next,

$$c_{1,n} = \sqrt{2}\sum_i \bar{g}_i s(2n + p - 1 - i) \tag{3.44}$$

From Figure 3.8 for $m = 2$ we have

$$\begin{aligned} d_n^2 &= \sqrt{2}\sum_l \bar{h}(l)a_{2n+p-1-l}^1 \\ &= 2\sum_l\sum_i \bar{h}(l)\bar{g}(i)x(4n + 3p - 3 - 2l - i) \end{aligned} \tag{3.45}$$

since

$$c_{2,n} = \frac{1}{2}\int_{-\infty}^{+\infty} x(t)\psi\left(\frac{t}{4} - n\right)\, dt \tag{3.46}$$

equating Equation 3.45 with -3.46 and using Equation 3.36

$$\frac{1}{2}\psi\left(\frac{t}{4} - n\right) = 2\sum_l\sum_i \bar{g}(i)\bar{h}(l)\phi(t - 4n - 3p + 3 + 2l + i) \tag{3.47}$$

Now, using Equation 3.43 we obtain

$$\begin{aligned} \sum_l \bar{h}(l)\phi\left(\frac{t}{2} - 2n - p + 1 + l\right) &= 2\sum_l\sum_i \bar{g}(i)\bar{h}(l) \\ &\quad \phi(t - 4n - 3p + 3 + 2l + i) \end{aligned} \tag{3.48}$$

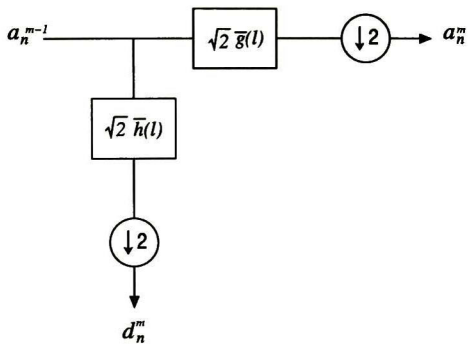


Figure 3.8: The mth stage decomposition of the MRA.

With $t/2 - 2n - p + 1 + l = t$ this simplifies to

$$\phi(t) = 2\sum_i \bar{g}(i)\phi(2t - p + 1 + i) \tag{3.49}$$

Equations 3.43 and 3.49 are two-scale equation that define the functions $\phi(t)$ and $\psi(t)$, via the decomposition filter coefficients. For perfect reconstruction, the synthesis filters coefficients must satisfy the relationship (in the next section we will address the design of perfect reconstruction filters.)

$$\bar{g}(l) = g(p - 1 - l) \tag{3.50}$$

$$\bar{h}(l) = h(p - 1 - l) \tag{3.51}$$

Putting these into Equations 3.43 and 3.49 we obtain the classical two-scale equations

$$\phi(t) = 2\sum_l g(l)\phi(2t - l) \tag{3.52}$$

$$\psi(t) = 2\sum_l h(l)\psi(2t - l) \tag{3.53}$$

As previously stated, $\phi(t)$ is the scaling function associated with the low passfilter $g(l)$, so called because it serves to time-scale the sequence $x(n)$. These two scale difference equations are fundamental in the generation of orhonormal, compact support, discrete parameters wavelets [21, 16, 51, 77, 82]. In preceeding paragraphs we have shown that if Equation 3.36 holds, then $c_{m,n} = d_n^m$ for $m = 1, 2$. We assumed the this result is valid for all m without a proper demonstration. If the reader is interested [16] provides a demonstration by mathematical induction.

So far, we have only seen what happens in the time domain. Let $\Psi(\omega)$, $\Phi(\omega)$ be the Fourier transforms of the wavelet and scaling functions respectively, and let

$$G(\omega) = \sum_l g(l)e^{j\omega l} \tag{3.54}$$

$$H(\omega) = \sum_l h(l)e^{j\omega l} \tag{3.55}$$

The Fourier transform of Equations 3.52 and 3.53 are

$$\Phi(\omega) = G\left(\frac{\omega}{2}\right)\Phi\left(\frac{\omega}{2}\right) \tag{3.56}$$

$$\Psi(\omega) = H\left(\frac{\omega}{2}\right)\Phi\left(\frac{\omega}{2}\right) \tag{3.57}$$

Iterating Equations 3.56 and 3.57 to infinity results in

$$\Phi(\omega) = \Phi(0)\prod_{k=1}^{\infty} G\left(\frac{\omega}{2^k}\right) \tag{3.58}$$

$$\Psi(\omega) = \Phi(0)H\left(\frac{\omega}{2}\right)\prod_{k=1}^{\infty} G\left(\frac{\omega}{2^{k+1}}\right) \tag{3.59}$$

If we normalize the scaling function by $\int \phi(t)\, dt = 1$, leads to $\Phi(0) = 1$, and due to Equation 3.58 $G(0) = 1$. Additionally, because of Equations 3.9 and 3.57 $H(0) = 0$.

3.4.2 The idea of multiresolution

Multiresolution analysis [51] is the decomposition of a signal $x(t)$ into components of different scales (frequencies). Associated with each scale (frequency band) there is a subspace $\mathbf{V}_m$. Thus, there is a *piece* of $x(t)$ in each subspace. For audio signals this scales are usually octaves. These subspaces are time functions which satisfy conditions 1 to 4 [77, 16]:

1. Containment. $\mathbf{V}_j \subset \mathbf{V}_{j+1}$, $\bigcap \mathbf{V}_j = \{\emptyset\}$ and $\overline{\bigcup \mathbf{V}_j} = \mathbf{L}^2$ The subspaces begin with the null space and expand in scales of two to reach the space of all square integrable functions. If a function $x(t)$ is in $\mathbf{V}_j$, then $x(2t)$ is in $\mathbf{V}_{j+1}$, and vice versa.

2. Existence of orthonormal scaling functions. There exists a scaling function $\phi(t) \in \mathbf{V}_0$ such that the set

$$\left\{\phi_{m,n}(t) = 2^{-\frac{m}{2}}\phi(2^{-m}t - n) : \; n \in \mathbb{Z}\right\} \tag{3.60}$$

is an orthonormal basis that spans $\mathbf{V}_m$

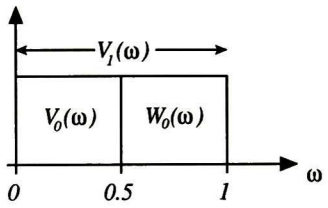


Figure 3.9: *Spectrum of subspaces.*

3. Basis functions defined by two-scale difference equations. Since $\phi_{0,n}(t)$ spans $\mathbf{V}_0$ and $\phi_{1,n}(t)$ spans $\mathbf{V}_1$ and $\mathbf{V}_1$ contains $\mathbf{V}_0$, then $\phi_{0,0}(t)$ is a linear combination of $\phi_{1,n} = \sqrt{2}\phi(2t-n)$, i.e.,

$$\phi(t) = 2\sum_l g(l)\phi(2t-l) \tag{3.61}$$

which is a two-scale difference equation. Consider the spectra $\mathbf{V}_1$ and $\mathbf{V}_0$ in Figure 3.9. The bandwidth of $V_0(\omega)$ is one half that of $V_1(\omega)$; in other words $V_1(\omega)$ contains $V_0(\omega)$. The subspaces $\mathbf{W}_0$ and $\mathbf{V}_0$ are orthogonal because they have no common frequency components and they are complementary to each other in forming $V_1(\omega)$. Now since $\mathbf{W}_0 \subset \mathbf{V}_1$, then, the wavelet function $\psi(t)$ whose translate $\psi(t-n)$ span $\mathbf{W}_0$, can also be written as a linear combination of $\psi_{1,n}$ which span $V_1(\omega)$. Thus,

$$\psi(t) = 2\sum_l h(l)\phi(2t-l) \tag{3.62}$$

The generalization of this process is presented in Figure 3.10. Note that for clarity purposes, the figure does not illustrate the unit norm subspaces.

4. Existence of orthonormal wavelet functions. If we substitute Equation 3.62 into Equation 3.34, yields

$$\begin{aligned} \psi_{m,n}(t) &= 2^{1-\frac{m}{2}}\sum_l h(l)\phi(2^{1-m}t-2n-l) \\ \psi_{m,n}(t) &= \sqrt{2}\sum_l h(l)\phi_{m-1,2n+1}(t) \end{aligned} \tag{3.63}$$

provided that $\phi_{m,n}(t)$ and $\psi_{m,n}(t)$ are also orthonormal.

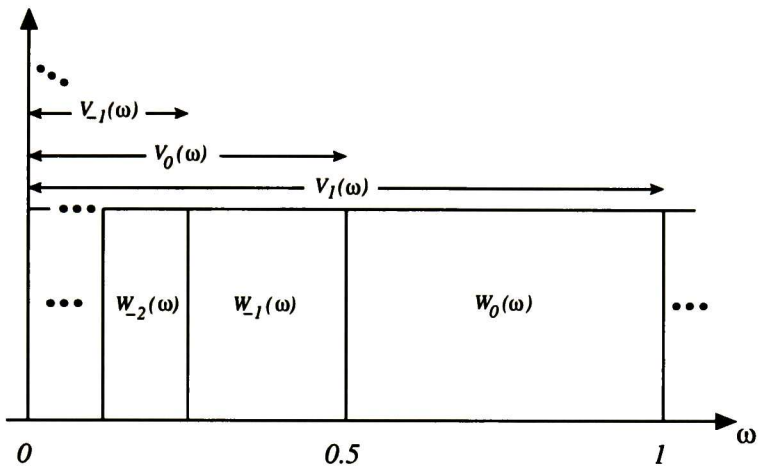


Figure 3.10: *Decomposition of the frequency spectrum into succesive subspaces (octave bands). There is a scaling factor for $V_j(\omega)$ and $W_j(\omega)$ by $2^{j/2}$, not depicted, to make subspaces of unit norm.*

3.5 Perfect Reconstruction Paraunitary Filter Banks

In this section we will present a special kind of Perfect Reconstruction (PR) filter banks that satisfy the *lossless* or *paraunitary* property, which is basic to the generation of the *orthonormal wavelet basis*.

An audio coding system is designed to produce signals that are finally perceived by a human listener. Additionally, most models of perception are based on a frequency domain formulation and assume that any frequency response can be obtained by combining and weighting a given number of subband channels [10]. For this to be true, the corresponding subband filters must be as selective as possible, approaching ideal filters. Thus, the filter bank selectivity must be increased by some means. The most natural solution to this problem is to let the impulse response of the filters be longer than in the orthogonal transform[10] case (Section 2.5). This will change the size of the $M \times M$ orthogonal transform matrix $\mathbf{T}$ to an $L \times M$ matrix $\mathbf{A}$, whose properties are well described in [79, 80]. The block scheme of the system is depicted in Figure 3.11, where $\mathbf{B}$ represents the synthesis transform matrix. Under this situation, the length of the basis functions is longer than the number of transform channels ($L \geq M$). At the analysis, a temporal overlapping of $L - M$ samples is introduced. In the absence

[10] As stated in the previous chapter, transform coding can be seen as a special case of subband coding.

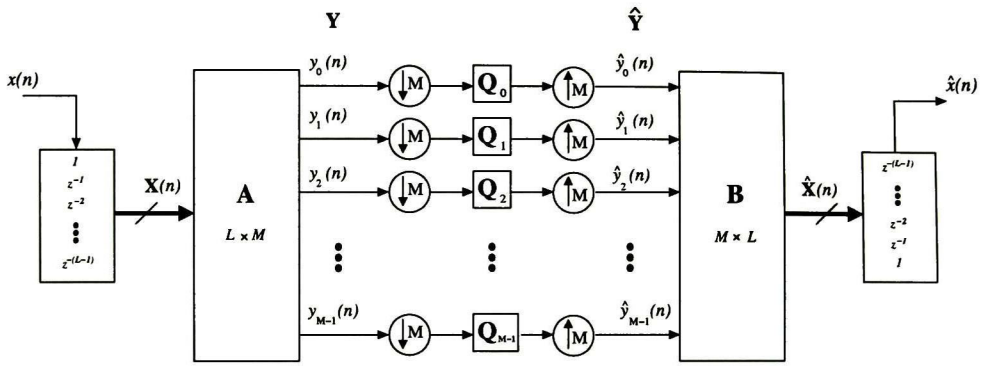


Figure 3.11: *M channel paraunitary transform.*

of quantizers, the operations performed by the system are presented in the following expression

$$\hat{X}(z)\begin{bmatrix} 1 \\ z^{-1} \\ z^{-2} \\ \vdots \\ z^{-(L-1)} \end{bmatrix} = [\mathbf{b}_0 \ \mathbf{b}_1 \ \cdots \ \mathbf{b}_{M-1}] \begin{bmatrix} \mathbf{a}_0 \\ \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_{M-1} \end{bmatrix} \begin{bmatrix} 1 \\ z^{-1} \\ z^{-2} \\ \vdots \\ z^{-(L-1)} \end{bmatrix} X(z) \qquad (3.64)$$

where $\mathbf{a_i}$ are the analysis basis functions, $\mathbf{b_i}$ the synthesis basis functions and $\hat{x}(n)$ a partially reconstructed window of $x(n)$.

3.5.1 Polyphase representation

An important advancement in multirate signal processing is the invention of the *polyphase representation.* This permits great simplifications of theoretical results and also leads to computational efficient implementations of decimation/interpolation filters. If the reader is interested, [77, 79, 80] provide a deep coverage about polyphase representation.

If we write the analysis filter as $A_i(z)$ and the synthesis filter as $B_i(z)$, for $0 \leq i \leq M-1$; we can express these transfer functions as the M-term sums

$$A_i(z) = \sum_{k=0}^{M-1} E_{ik}(z^M) z^{-k} \qquad (3.65)$$

and

$$B_i(z) = \sum_{k=0}^{M-1} R_{ki}(z^M) z^{-(M-1-k)} \tag{3.66}$$

with this notation, the analysis filters can be written as

$$\begin{bmatrix} A_0(z) \\ \vdots \\ A_{M-1}(z) \end{bmatrix} = \begin{bmatrix} E_{0,0}(z^M) & E_{0,1}(z^M) & \cdots & E_{0,M-1}(z^M) \\ \vdots & \vdots & & \vdots \\ E_{M-1,0}(z^M) & E_{M-1,1}(z^M) & \cdots & E_{M-1,M-1}(z^M) \end{bmatrix} \begin{bmatrix} 1 \\ z^{-1} \\ z^{-2} \\ \vdots \\ z^{-(M-1)} \end{bmatrix} \tag{3.67}$$

additionally, the synthesis filters can be written as

$$\begin{bmatrix} B_0(z) \\ \vdots \\ B_{M-1}(z) \end{bmatrix} = \begin{bmatrix} z^{-(M-1)} \\ z^{-(M-2)} \\ \vdots \\ 1 \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} R_{0,0}(z^M) & R_{0,1}(z^M) & \cdots & R_{0,M-1}(z^M) \\ \vdots & \vdots & \ddots & \vdots \\ R_{M-1,0}(z^M) & R_{M-1,1}(z^M) & \cdots & R_{M-1,M-1}(z^M) \end{bmatrix} \tag{3.68}$$

Using vector notation, we can rewrite Equations 3.67 and 3.68 in a more compact way

$$\mathbf{a}(z) = \mathbf{E}(z^M)\mathbf{d}(z) \tag{3.69}$$

$$\mathbf{b}^{\mathsf{T}}(z) = z^{-(M-1)}\tilde{\mathbf{d}}(z)\mathbf{R}(z^M) \tag{3.70}$$

where $\tilde{\mathbf{d}}(z) = \mathbf{d}^{\mathsf{T}}(z^{-1})$ is the paraconjugate[11] version of $\mathbf{d}(z)$. The delay $z^{-(M-1)}$ is necessary to ensure the causality of the system. $\mathbf{E}(z)$ is called the *analysis polyphase matrix* and $\mathbf{R}(z)$ the *synthesis polyphase matrix*. In the absence of quantizers, the analysis and synthesis cascade results in the polyphase matrix product $\mathbf{P}(z) = \mathbf{R}(z)\mathbf{E}(z)$. As for the orthogonal transform case, this product should be identical to the $M \times M$ identity matrix to satisfy perfect reconstruction of the input, i.e.,

$$\hat{x}(n) = x(n - L + 1) \tag{3.71}$$

[11] In [80] the paraconjugate is defined as $\tilde{\mathbf{d}}(z) = \mathbf{d}_*^{\mathsf{T}}(z^{-1})$, where '$*$' stands for the complex conjugate. In this case we are using filters with real coefficients, thus the complex conjugate is not necessary.

If $\mathbf{R}(z)$ is chosen as the paraconjugate of $\mathbf{E}(z)$ [10, 80], then

$$\mathbf{P}(z) = \tilde{\mathbf{E}}(z)\mathbf{E}(z) = \mathbf{I}_M \tag{3.72}$$

This result describes a PR scheme and is analogous to the one found in the orthogonal transform case. Since the synthesis polyphase matrix $\mathbf{R}(z) = \tilde{\mathbf{E}}(z)$ is noncausal, it can be rendered causal by choosing it as

$$\mathbf{R}(z) = z^{-K}\tilde{\mathbf{E}}(z) \tag{3.73}$$

with $K \geq \frac{L-M}{M}$. The paraunitariness of $\mathbf{E}(z)$ is a sufficient, but not necessary condition for PR.

3.5.2 Two-band FIR Paraunitary Filter Banks

As previously mentioned in Chapter 1, in this work we make use of the orthogonal wavelet transform. Since there exists a close relationship between this transform and two-band paraunitary filter banks [80, 77], we present this filter bank case here. A two-band analysis/synthesis filter bank is shown in Figure 3.12.

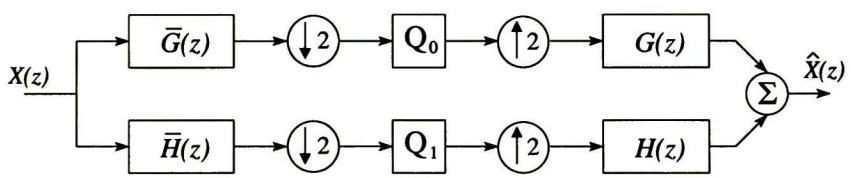


Figure 3.12: *A two-band analysis/synthesis filter bank.*

It can be demonstrated [77, 82, 80] that the reconstruction equation of this system can be expressed as

$$\hat{X}(z) = \frac{1}{2}[\bar{G}(z)G(z) + \bar{H}(z)H(z)]X(z) + \frac{1}{2}[\bar{G}(-z)G(z) + \bar{H}(-z)H(z)]X(-z) \tag{3.74}$$

In order to have a linear transfer function $T(z)$ for the whole system such that

$$T(z) = \frac{\hat{X}(z)}{X(z)} \tag{3.75}$$

the aliasing components produced by the downsampling operation, the terms with $X(-z)$, must be eliminated. This can be achieved by imposing the following conditions to the filters

$$G(z) = \bar{H}(-z) \tag{3.76}$$

$$H(z) = -\bar{G}(-z) \tag{3.77}$$

these choices are equivalent to the time domain expressions

$$g(n) = (-1)^n \bar{h}(n) \tag{3.78}$$

$$h(n) = (-1)^{n+1} \bar{g}(n) \tag{3.79}$$

Thus, $T(z)$ simplifies to

$$T(z) = \frac{1}{2}[\bar{G}(z)G(z) + \bar{H}(z)H(z)]. \tag{3.80}$$

The input signal, $x(n)$, is perfectly reconstructed if $T(z)$ is a constant magnitud function [77, 82, 80]. In other words, there is neither amplitude distortion nor phase distortion of $\hat{x}(n)$ with respect to $x(n)$. For this to be true, we must accomplish

$$T(z) = z^{-(L-1)} \qquad L \geq 0 \tag{3.81}$$

This expression is equivalent to the time domain relation given by Equation 3.71. The output is simply a delayed version of the input[12] It must be clear that this condition is imposed on the whole system, i.e., individual filters may have non-linear phase and the whole analysis/synthesis structure still preserves linear phase and satisfies Equations 3.81 and 3.71. If we assume that each of the analysis/synthesis filters can be decomposed into their odd and even sample terms, we can write the filter equations in the polyphase form given by Equations 3.65 and 3.66 with $M = 2$ and expressed as

$$\begin{bmatrix} \bar{G}(z) \\ \bar{H}(z) \end{bmatrix} = \begin{bmatrix} E_{0,0}(z^2) & E_{0,1}(z^2) \\ E_{1,0}(z^2) & E_{1,1}(z^2) \end{bmatrix} \begin{bmatrix} 1 \\ z^{-1} \end{bmatrix} \tag{3.82}$$

$$\begin{bmatrix} G(z) \\ H(z) \end{bmatrix} = \begin{bmatrix} z^{-1} \\ 1 \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} R_{0,0}(z^2) & R_{0,1}(z^2) \\ R_{1,0}(z^2) & R_{1,1}(z^2) \end{bmatrix} \tag{3.83}$$

[12]The output can also be a scaled version of the input of the form $\hat{x}(n) = cx(n - L - 1)$ with $c \in \mathbf{C}$ and $c \neq 0$.

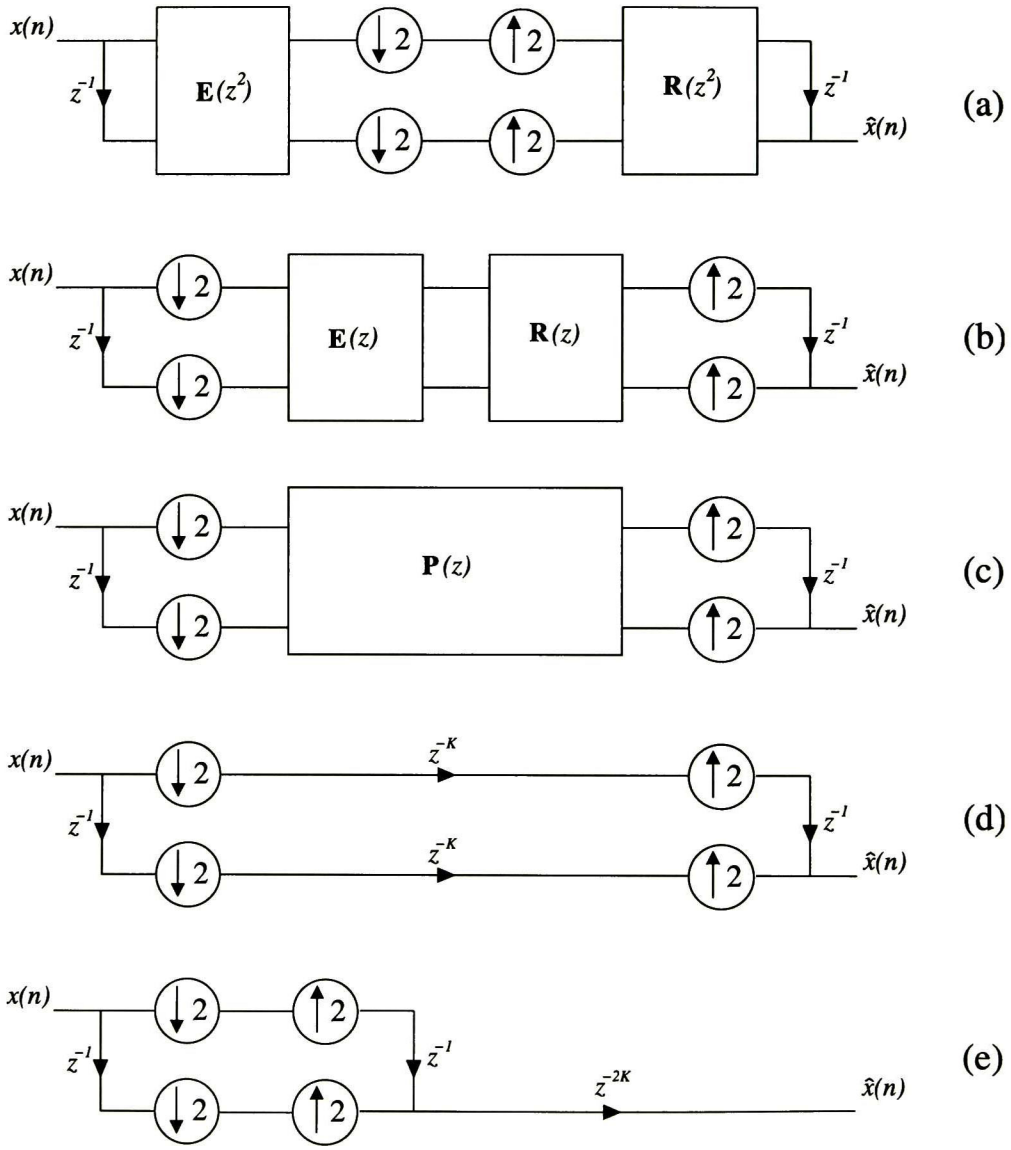


Figure 3.13: *Two-band filter bank polyphase decomposition stages.*

Using Equations 3.69 and 3.70, we may also express the filter in vector form as $[\bar{\mathbf{g}}\ \bar{\mathbf{h}}]^{\mathsf{T}} = \mathbf{E}(z^2)\mathbf{d}(z)$ and $[\mathbf{g}\ \mathbf{h}]^{\mathsf{T}} = z^{-1}\tilde{\mathbf{d}}(z)\mathbf{R}(z^2)$. This situation is presented in Figure 3.13(a). Then using the noble identities [77, 80], presented in Appendix D, we can move the polyphase block to the middle of the structure by reducing their power factors, as shown if Figure 3.13(b). The product of the analysis and synthesis polyphase matrices $\mathbf{P}(z)$, previously given, is presented in Figure 3.13(c). To achieve perfect reconstruction of the input signal $x(n)$, we can choose the polyphase matrix to be

$$\mathbf{R}(z) = z^{-K}\tilde{E}(z) \tag{3.84}$$

where $K = \frac{L-2}{2}$. This situation is illustrated in Figure 3.13(d). Using again the noble identities, the delays z^{-K} can be moved to the output of the filter bank, as depicted in Figure 3.13(e). This last block diagram is filter bank characterized by the following filter relations

$$\begin{aligned} \bar{G}(z) &= 1 \\ G(z) &= z^{-(2K+1)}, \\ \bar{H}(z) &= z^{-1} \\ H(z) &= z^{-2K} \end{aligned}$$

This can also be easily shown by using Equation 3.80:

$$T(z) = \frac{1}{2}[z^{-(2K+1)} + z^{-1}z^{-2K}] = z^{-(2K+1)} = z^{-(L-1)} \tag{3.85}$$

Thus, the paraunitary choice of Equation 3.84 leads to perfect reconstruction, satisfying Equation 3.71. The filters must satisfy the relation [10]

$$\bar{H}(z) = -z^{-(L-1)}\widetilde{\bar{G}}(-z) \tag{3.86}$$

which can be expressed in the domain as

$$\bar{h}(n) = (-1)^{n+1}\bar{h}(L-1-n) \tag{3.87}$$

Thus, with the help of Equations 3.76 to 3.79, the filter relations for the two-band paraunitary filter bank are given in Table 3.1 Each filter is represented as a function of $\bar{G}(z)$ in the z domain and in the time domain. Furthermore, the paraconjugation operation has been directly replaced by $z \to z^{-1}$ and, as mentioned before, the filters are assumed to be real. Additionally, it must be noted that $H(z) = z^{-(L-1)}\widetilde{\bar{H}}(z)$, meaning that the analysis and synthesis filters are the paraconjugate versions of one

z domain
$\bar{H}(z) = -z^{-(L-1)}\bar{G}(-z^{-1})$
$G(z) = z^{-(L-1)}\bar{G}(z^{-1})$
$H(z) = -\bar{G}(-z)$
Time domain
$\bar{h}(n) = (-1)^{n+1}\bar{g}(L-1-n)$
$g(n) = \bar{g}(L-1-n)$
$h(n) = (-1)^{n+1}\bar{g}(n)$

Table 3.1: *Two-band filter bank relations for perfect reconstruction.*

another. Filters satisfying such relations are also known as *conjugate quadrature filters* (CQF) since they are half-band filters that possess mirrored magnitudes with respect to $\omega = \pi/2$ and maintain a complex conjugate relation. Several techniques have been proposed for the design of the low pass filter $\bar{G}(z)$, the discussion of this techniques is out of the scope of this thesis. Among the most importants are *spectral factorization* [21, 77, 80], *autocorrelation optimization* [10], *lattice structure optimization* [77, 80] and *modulated filters* [77, 80, 82].

3.5.3 Tree Structured Filter Banks

Using the property that a cascade of paraunitary systems is also paraunitary [80], the two-band paraunitary filters, analyzed in the previous subsection, can be cascaded to form binary tree structures that still conserve the paraunitary property [80]. An example of such a tree with two stages is shown in Figure 3.14. Using the noble identities, we can move the sample rate converters and directly cascade the filters to obtain the filter bank structure presented in Figure 3.15. If the prototype low pass filter $\bar{G}(z)$ has length L and N is the number of bands of the tree structure (a power of

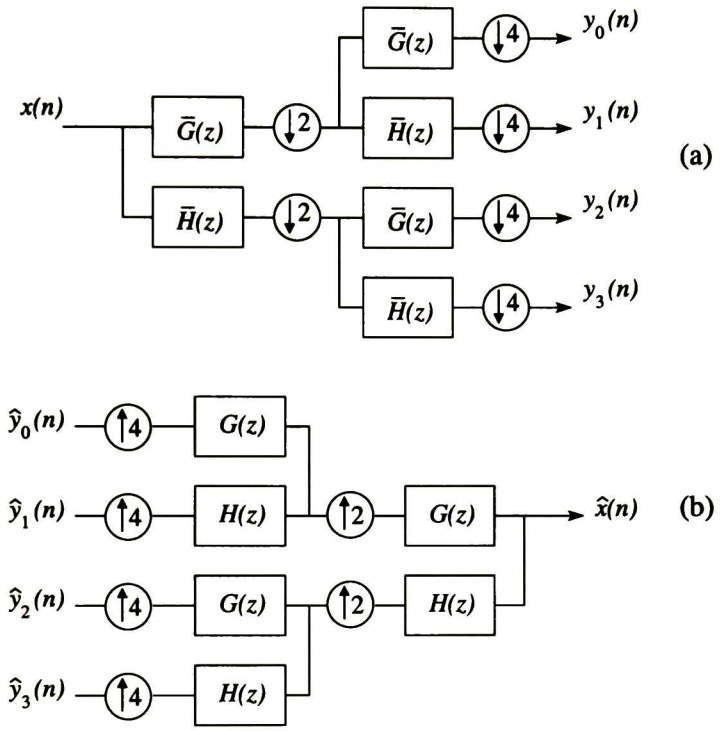


Figure 3.14: *Four-band tree structured filter bank (a) Analysis (b) Synthesis.*

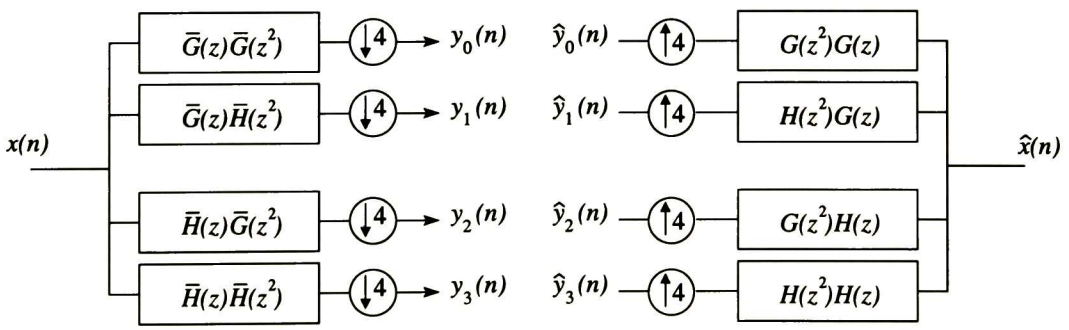


Figure 3.15: *Four-band filter bank. Parallel version.*

two, i.e. $N = 2^p$), the length of the tree filters is given by

$$M = (L-1)(N-1)+1 \tag{3.88}$$

At an intermediate stage j, for $0 \leq j \leq p-1$, the filter length can be computed as

$$M^{(j)} = (L-1)(2^{j+1}-1)+1 \tag{3.89}$$

Therefore, the total reconstruction delay of a paraunitary tree structured filter bank is M, and the output signal is given by $\hat{x}(n) = x(n-M+1)$.

3.6 Time-Frequency Auditory Mapping

As mentioned in Chapter 1, in this work we make use of a particular transform called Wavelet Packet Transform (WPT) [19, 18, 34, 51, 70] that can be seen as a generalization of the WT; which, as previously seen, performs an octave scale time-frequency decomposition. In [19] Coifman *et al.* use the WPT to perform a non uniform time-frequency decomposition of acoustic signals. Although wavelets and wavelet packets are still a field of intensive research, many aspects and properties are well documented nowadays [16, 17, 18, 19, 21, 22, 34, 51, 52, 68, 77, 82].

3.6.1 Orthonormal Wavelet Packet Transform

The Discrete orthonormal Wavelet Packet Transform (DWPT) can be implemented using subtrees of an N-channel tree structured paraunitary filter bank [72, 81]. The decomposition is simply obtained by pruning or suppressing some branches of the tree, depending on the desired time-frequency resolution. The tree must be of the type depicted in Figure 3.14.

Early in this chapter we presented the CWT. The case of the Discrete Wavelet Transform (DWT), obtained from an actave tree, is alike. In the tree of Figure 3.16-(a) the upper branches correspond to low pass prototype filters, at the outputs f_s stands for the sampling frequency. Another idealized time-frequency representation is the *tiling*, which can be used to present the resolutions in the time domain and in the frequency domain performed by the transform, as shown in Figure 3.16-(b). Finally, the idealized magnitude responses of the DWT are shown in Figure 3.16-(c).

The more general case of the DWPT is presented if Figure 3.17. The tree of Figure 3.17-(a) shows the arrangement of the filters in order to obtain a time-frequency tiling

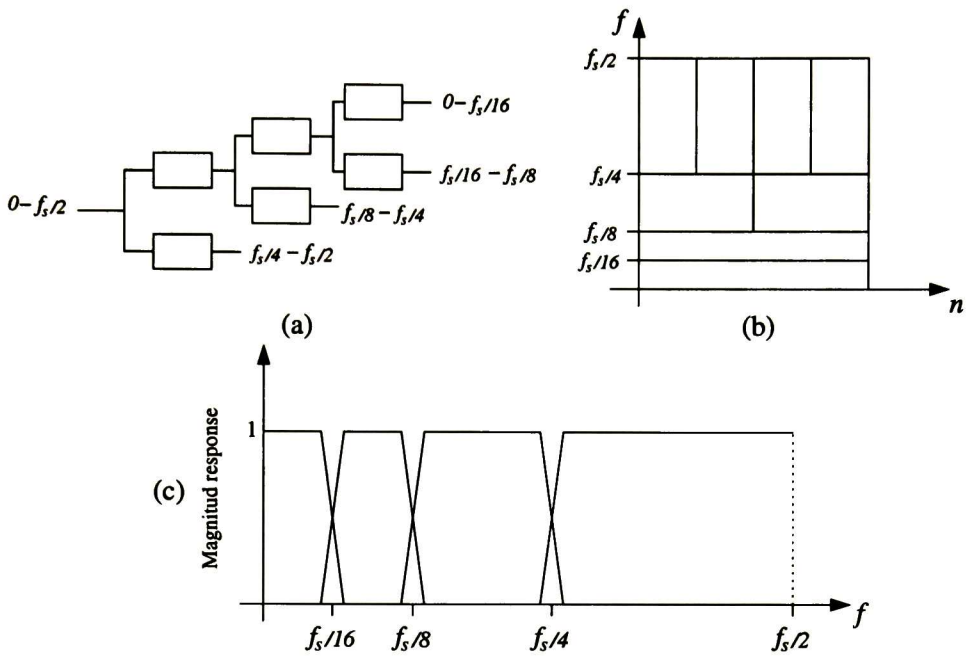


Figure 3.16: *Three stage wavelet decomposition. (a) Filter tree structure (Mallat's decomposition scheme). (b) Time-frequency tiling. (c) Idealized magnitud response of the filter bank.*

of the form depicted in 3.17-(b). The corresponding idealized magnitudes are presented in Figure 3.17-(c).

Wave packet representations have been proposed as an extension of the wavelet transform. In a usual wavelet transform only the approximation (output of the low pass filter) at a given scale is further decomposed. In contrast, in a wave packet decomposition [18], the 2-band wavelet filter banks are used to split both the low pass and high pass bands. As already presented, this type of decomposition is presented by a binary tree in which one has the freedom to stop or continue the decomposition at any node [18].

3.6.2 Filter Bank Delay

The tree structure decomposition, with p stages, introduces at each stage a processing delay that has two components, as depicted in Figure 3.18. One is due to the

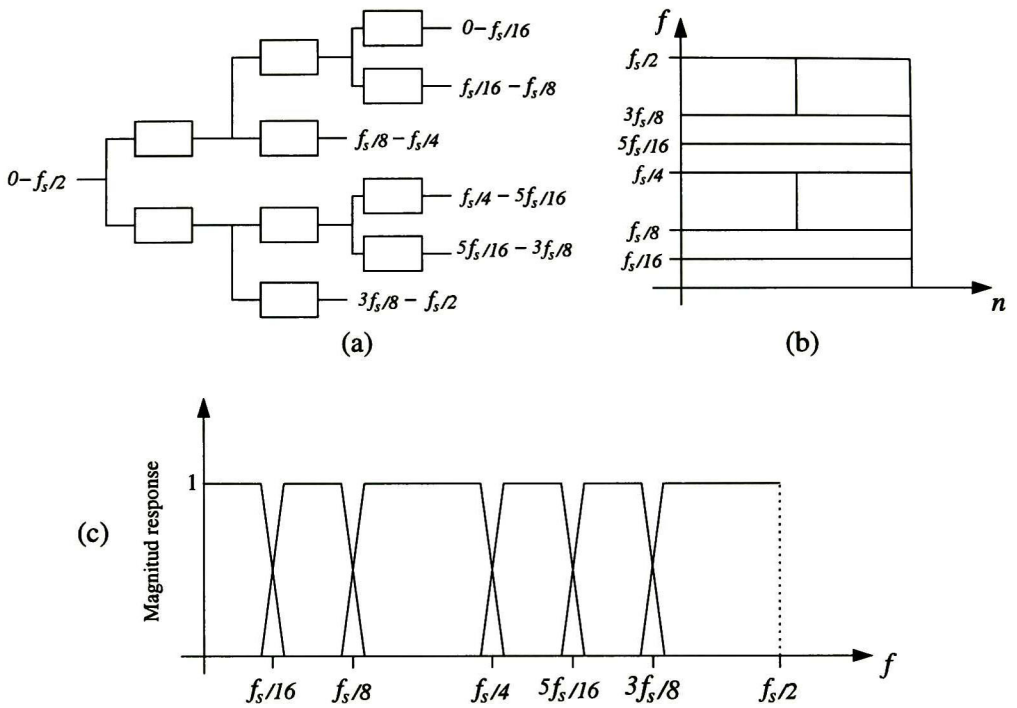


Figure 3.17: *Three-stage wavelet packet decomposition. (a) Filter tree structure. (b) Time-frequency tiling. (c) Idealized magnitud response of the filter bank.*

delay chain of the filter and the other to the powered polyphase matrix. Stage j, for $0 \leq j \leq p-1$, contributes to the global delay by introducing the term

$$\Delta_j(z) = z^{-2^j} z^{-2^{j+1}(\frac{L}{2}-1)} = z^{-2^j(L-1)} \tag{3.90}$$

which is related to the prototype filter length L. It can be easily verified by inspection that the local delay at stage $j+1$ is twice that at stage j.

If the tree structure is prunned, the branches removed must be replaced by compensation delays. This is only necessary in an analysis/synthesis system, since these delays

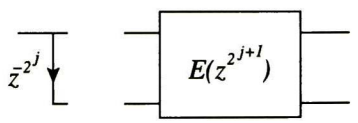


Figure 3.18: *Delay contributions of one tree branch at stage j.*

are required for the perfect reconstruction of the input signal. The accumulated delay $\Delta_{c,j}(z)$, due to the removal of several consecutive branches, depends on the stage remaining after pruning. If we assume this stage as being j, we have

$$\Delta_{c,j}(z) = \prod_{k=j+1}^{p-1} \Delta_j(z) \tag{3.91}$$

3.6.3 Human Auditory System Modeling

The Human Auditory System (HAS) can be seen as a filter bank with varying frequency and time resolutions [31, 39, 60, 83]. The WPT, previously presented, lets us perform such non uniform time-frequency decompositions. Our goal is to approximate the HAS subband analysis, called critical band analysis[13], using the WPT. In this work, the critical band analysis, also known as Bark mapping, is necessary along with the STFT for the estimation of the auditory masking thresholds.

As stated in Chapter 1, in this thesis we consider signals sampled at 16 kHz and band limited from 50 Hz to 7,000 Hz. Within this bandwidth there are approximately 21-critical bands. The critical bandwidth towards lower frequencies is around 100 Hz. By using a six-stage tree structure decomposition a frequency resolution of 125 Hz can be achieved. Thus, we will use a tree with 6 stages. The design of the prototype filter $\bar{G}(z)$ influences the temporal and spectral selectivities of the transform. The design procedure should consider two main aspects: the magnitude response of $\bar{G}(z)$ and the filter length. Although these two features are related, they must be selected to achieve sufficient frequency selectivity while avoiding too much temporal spreading. Excessive temporal spreading could cause pre echos during the coding process. Finally, the DWPT tree structure was selected to approximate as good as possible the critical band decomposition performed by the ear. The selected tree structure is presented in Figure 3.19, the comparison between the DWPT and the model of the HAS is presented in Figures 3.20 and 3.21 respectively.

In [45, 48, 65, 71] a deep study on several different mother wavelets is performed by the authors and they all conclude that the best wavelets for speech and audio coding applications are Daubechies wavelets. In [65] the authors restrict the optimality of

[13]Critical band analysis and other subjects related to perceptual modeling will be presented in Chapter 4.

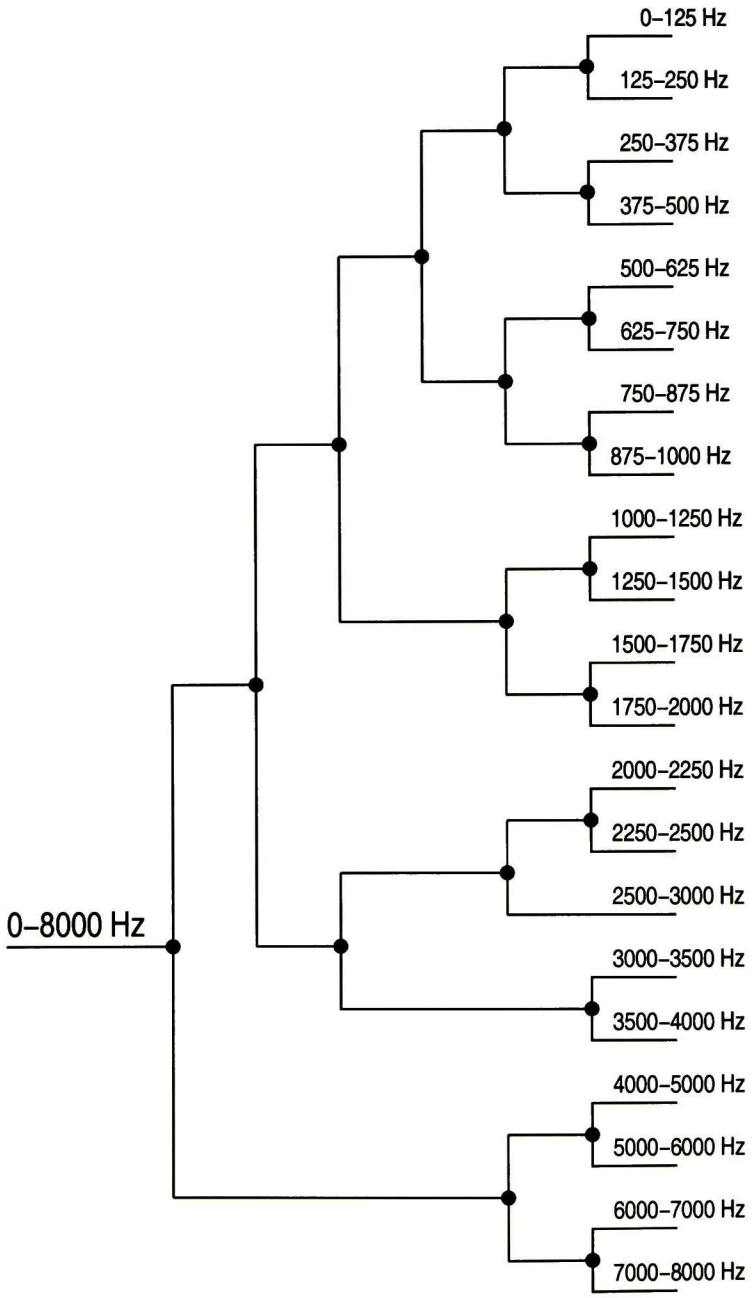


Figure 3.19: *WPT tree structure to approximate the critical band decomposition of the human auditory system.*

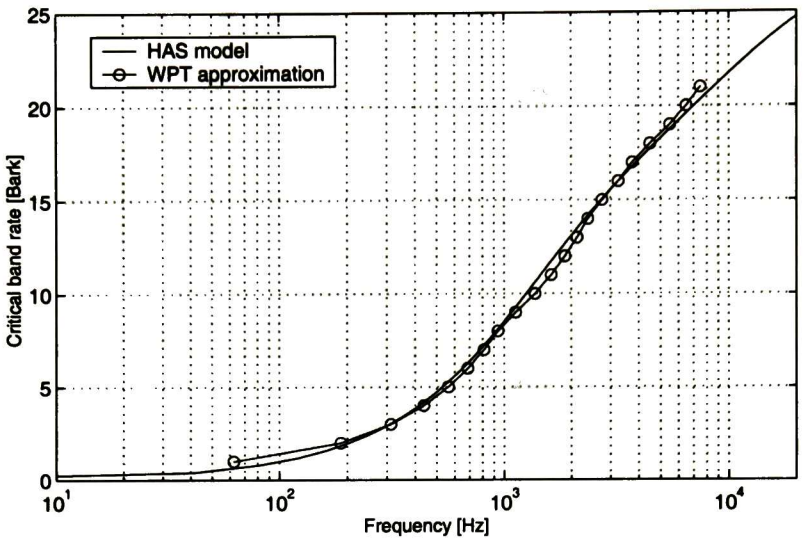


Figure 3.20: *Critical band rate approximation. Comparison between the DWPT and the human auditory system model.*

Daubechies wavelets only to the case where there exists delay constraints. The filters proposed by Daubechies [21] are the ones which preserve at best frequency selectivity as the number of stages of the DWPT increase, this is due to their regularity property. The wavelet DB5 is suggested by [12, 11, 13, 48, 75] because it allows to achieve a good frequency separation with a reasonable filter delay and time spread (it only has 10 coefficients). Other Daubechies wavelets permit better frequency separation at the expense of a higher filter delay or low time spread at the expense of poor frequency separation [48]. With the help of Equation 3.89, we can conclude that the chosen WPT tree structure introduces a processing delay of 35.5 ms.

The magnitude responses of the basis functions of the transform, corresponding to the 21 critical band filters, are shown in Figure 3.22-(a). It can be easily seen that the magnitude of the filters increases as they become narrower (see the first eight lobules of the graph). This is due to the orthonormality property of the transform. In addition, the magnitudes have also been computed from the mother wavelet DB10 (it has 20 coefficients) for comparison. Figure 3.22-(b) shows that even by increasing (doubling in this case) the length of the filters, the most important secondary lobes do not significantly decrease. It can be seen that for both kinds of filters an important

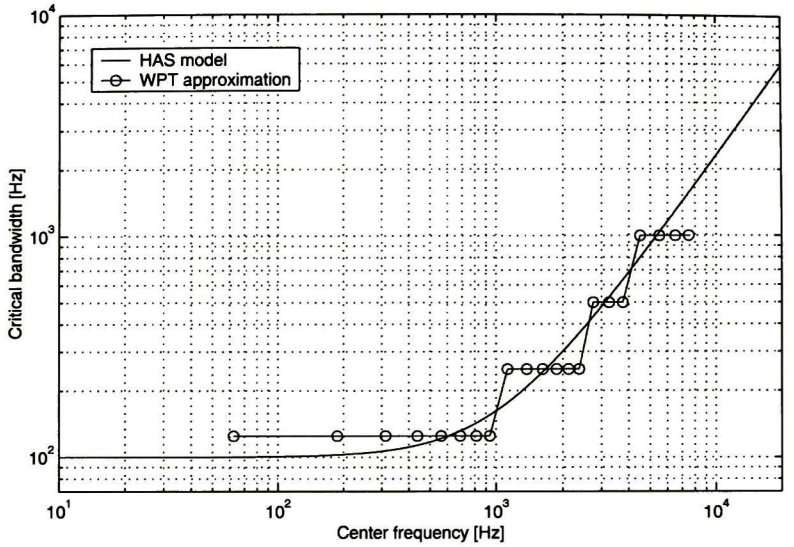


Figure 3.21: *Critical bandwidth approximation. Comparison between the DWPT and the human auditory system model.*

amount of overlapping does exist in the frequency domain between different subbands. This was consider to be a strong limitation to the use of such decomposition to calculate the masking thresholds. For this reason we decided to use the classical STFT method to calculate such thresholds, though there are authors that report successful results using the DWPT method [13, 75].

Once the transform has been chosen, its time-frequency resolution remains fixed. The time frequency tiling correspoding to this transform is shown in Figure 3.23.

The gain factor in each critical band depends on the stage of transform from which the coefficients X_i are extracted, where i stands for the coefficient number. Since we use an orthonormal transform, it is different at each stage. As we have previously highlighted, the effect can be observed in Figure 3.22. The two channel paraunitary transform introduces a gain of $\sqrt{2}$ in each of the subbands. Using the stage number j, depicted in Figure 3.23, this gain can be simply computed as

$$\Lambda_k(j) = \left(\sqrt{2}\right)^{j+1} \tag{3.92}$$

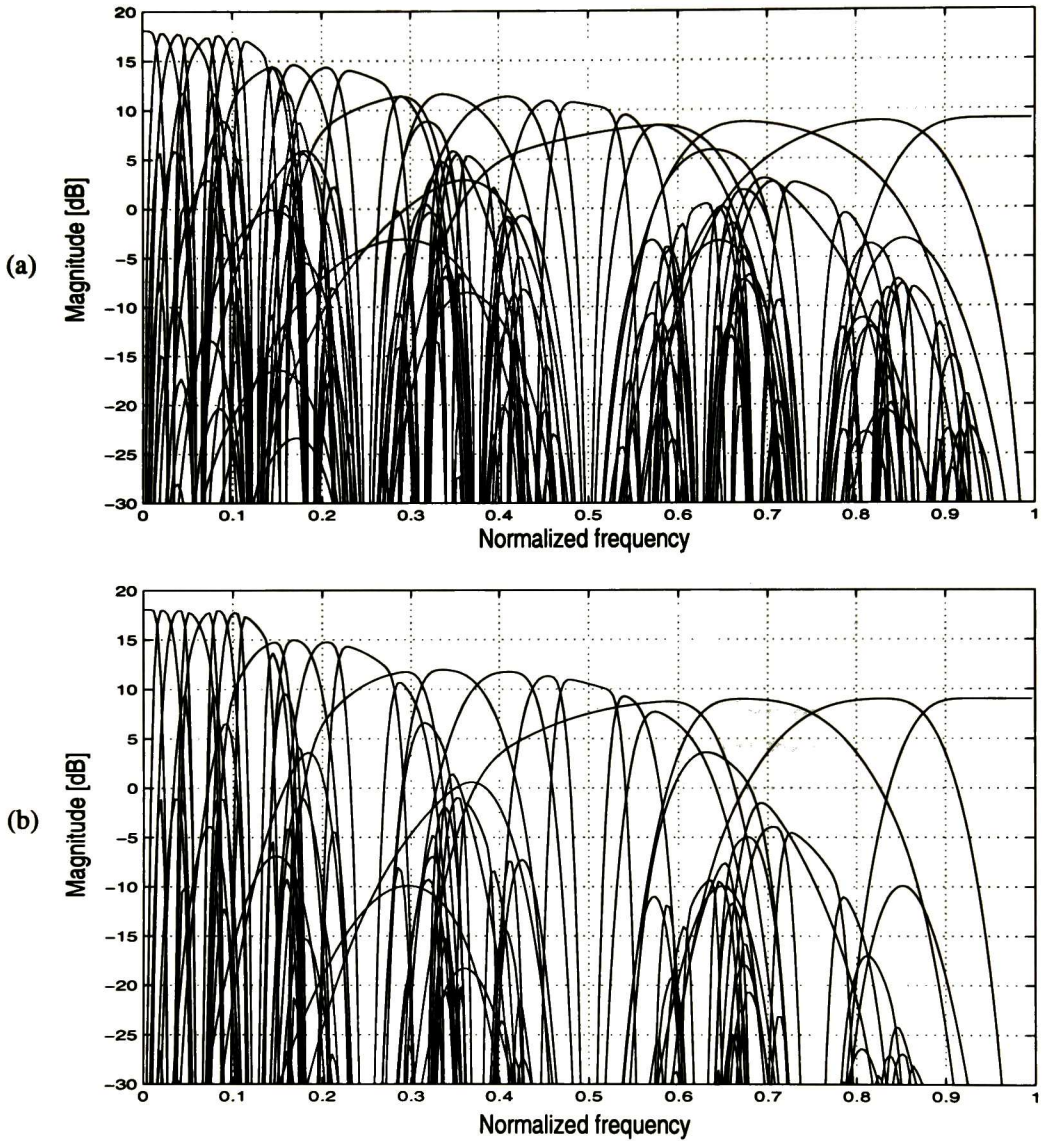


Figure 3.22: *Magnitude responses of the chosen wavelet packet transform for a) the mother wavelet DB5 (10 coefficients) b) the mother wavelet DB10 (20 coefficients).*

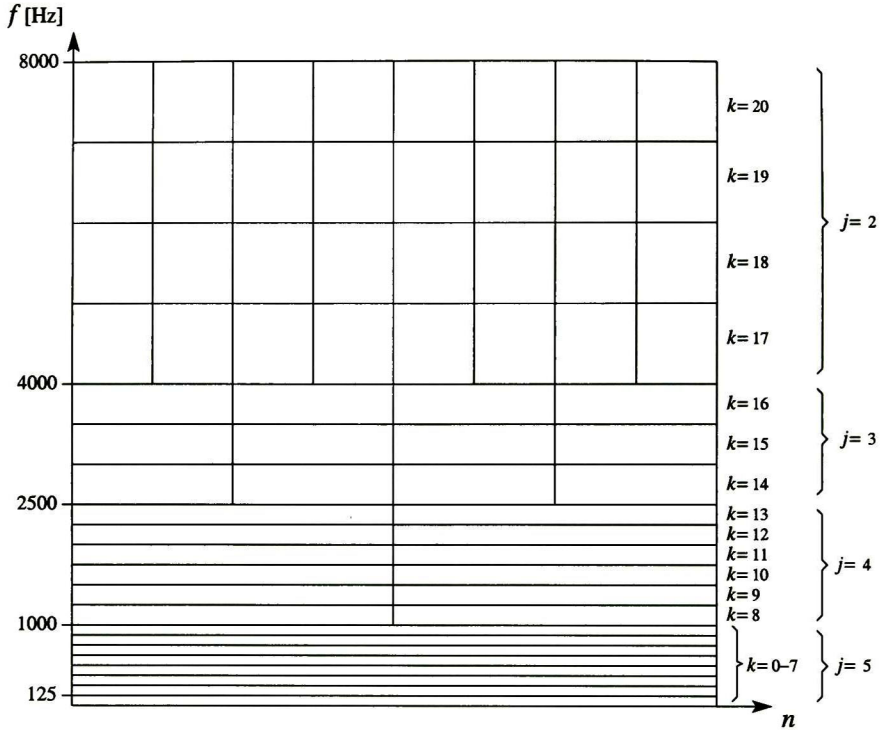


Figure 3.23: *Time-frequency tiling of the 21-critical band decomposition using the discrete wavelet packet transform. In the figure, j stands for decomposition stage, and k for the critical band number.*

Thus, for 16-bit quantized samples, the amplitude of the transform coefficients has the bounds

$$-\Lambda_k(j)\,2^{15} \leq X_i \leq \Lambda_k(j)\,2^{15} \tag{3.93}$$

Chapter 4

Auditory System Modeling

The goal of signal compression can be defined as the twofold search to achieve a low bit rate digital representation of an input signal while maintaining as minimum degradation as possible. During that search in the digital representation of a signal, it is essential that we design a coding algorithm that minimizes a perceptually meaningful measure of signal distortion, instead of the more traditional and more tractable criteria such as the mean squared difference between the waveforms of the input and output signals. A key issue of this idea is the notion of distortion (or noise) ***masking***, whereby the distortion that is inevitably introduced in the coding process, if properly distributed or shaped, is masked by the input signal itself. The masking can be partial or total, leading either to a system with increased quality compared to a system without noise shaping, or to perfect signal quality that is equivalent to that of the uncoded signal. In either case, the masking occurs because of the inability of the human perceptual mechanism to distinguish two signal components (one belonging to the signal, and the other one belonging to the noise). In the case of acoustic signals, this occurs in the same spectral or temporal locality. An importat consequence of this limitation is that the perceptibility of noise can be zero even if the objectively measured local signal-to-noise ratio (SNR) is low. Ideally, the noise level at all points in the signal space is exactly at the level of *just noticeable distortion* (JND). This corresponds to perfect signal quality at the lowest possible bit rate. This bit rate is a fundamental limit to which we can compress the signal with zero perceptible distortion. We call this limit the *perceptual entropy*. A signal coding algorithm that is based on the criterion of minimizing the perceived error is called a *perceptual coding* algorithm.

In this chapter we will review the main aspects of human audition and their useful

mathematical models. We start with some general aspects of hearing, the anatomic structure of the ear and modeling the behavior of the cochlea. Next, we present auditory masking by giving the notions of auditory scales and masking thresholds. Then, we explain the calculation procedure of the perceptual entropy and of the masking threshold.

4.1 General Aspects of Hearing

The auditory system allows human to detect different frequencies of sounds (*pitches*[1]) and to localize sounds in space. Sound is generated by mechanical vibrations that generate pressure waves in some medium (for example, air). These pressure waves travel through the medium in a way that depends on the physical characteristics of such medium. For example, sound travels through air at a speed[2] of 340 m/s. The pressure waves generated by sound set up vibrations in the tympanic membrane of the ear, which in turn are transmitted through the middle ear to the cochlea. Therefore, sound perception is essentially a special form of vibration sensitivity. Among the most important aspects of sound we have the spectral content and the intensity. The spectral content can be defined on the basis frequencies of the sinusoidal waves comprising a sound. As mentioned in previous chapters, humans are sensitive to sounds in the range of 20 to 20,000 Hz [39, 76, 83].

Sound intensity is defined on the basis of relative force. The measure is relative because it is with respect to the threshold for auditory perception. Sound intensity is referred to this threshold using decibels sound pressure level (SPL) as the unit of measurement. Human can distinguish differences in sound intensities ranging from 0 up to about 120 dB(SPL), i.e., the HAS has a dynamic range of 1 to 1 million [76]. Another important aspect of sound is *loudness*, which is a perceptual magnitude. Loudness can be defined as *that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud* [31]. There are two commonly used measures of loudness. One is *loudness level* (unit *phon*) and the other is *loudness* (unit *sone*) [31].

[1]The American Standards Association defines pitch as *that attribute of auditory sensation in which sounds may be ordered on a musical scale.* Pitch bears a close relationship with frequency: while frequency is an objective physical measure, pitch is a subjective perceptual measure [31].

[2]This speed varies and depends on the air temperature, pressure and humidity.

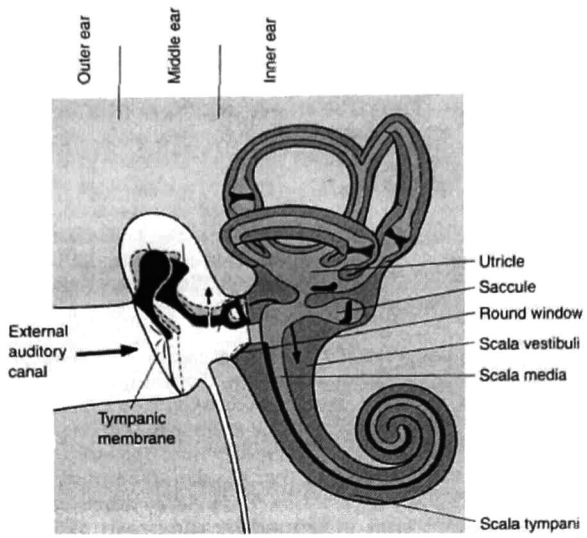


Figure 4.1: *Anatomy of the ear. The ear is mainly composed of three parts: the outer ear, the middle ear and the inner ear.*

4.1.1 Anatomy of the ear

The anatomical structure of the ear consists of three successive parts: the *outer ear*, the *middle ear* and the *inner ear*. Sound waves enter the ear through the *pinna* and the external *auditory canal*, causing the *tympanic membrane* (eardrum) to vibrate. The vibrations are conducted from the tympanic membrane to the *oval window* of the cochlea via the *ossicles*, the three small bones of the middle ear, which include the *malleus* (hammer), which is attached to the tympanic membrane; the *incus* (anvil); and the *stapes* (stirrups), which is pressed against the membrane covering the oval window of the cochlea, see Figure 4.1.

The transduction of sounds to neural activity occurs in the *cochlea*. The cochlea is part of a complex set of cavities in the petrous portion of the temporal bone called *vestibulo-cochlear labyrinth*. The cavities in the bone are calle the *bonny labyrinth*. Within these is a set of epithelial[3] membranes that form a *membranous labyrinth* that contains the sensory neuroepithelium [76].

[3] A membranous cellular tissue that covers a free surface or lines a tube or cavity of an animal body and serves especially to enclose and protect the other part of the body.

The components of the vestibulo-cochlear labyrinth are the *cochlea*, the *utricle*, the *sacule*, and three *semicircular canals*. The cochlea, depicted in Figure 4.2, is essentially a spiral tunnel through the temporal bone, which contains the specialized receptor apparatus. The cochlea is divided into three, parallel, fluid-filled compartments: *the scala vestibuli*, *scala tympani* and *scala media*. The overall appereance of the cochlea is like a snail's shell that coils two and a half times. The tunnel narrows from a large end where the oval and round windows are (the base) to a narrow end (the apex). The tunnel of the cochlea is divided into two more or less equal longitudinal chambers *scalae* by the basiliar membrane, this last has a length of approximately 35 mm. The two chambers are termed the scala tympani and the scala vestibuli. The *Reissner's membrane* attaches near the midpoint of the basilar membrane and extends to the wall of the tunnel, forming a third, small, triangular compartment called the *scala media*. The floor of this triangular compartment is formed by the *organ of Corti*, a highly specialized *neuroepithelium* that contains the auditory receptor cells (the cochlear *hair cells*). The scala vestibuli and the scala tympani are filled with a fluid termed *perilymphatic fluid*. The two chambers are connected via a small opening at the apex of the cochlea, termed the *helicotrema*, so the fluid composition of the two chambers is identical [76].

The external ear directs sound waves towards the eardrum. The tympanic membrane together with the ossicles in the middle ear transform the large-amplitude, low-force, airborne waves into small-amplitude, high force vibrations of the membrane at the oval window. The most relevant of the middle ear tasks is this *impedance matching* which consists in adapting sounds in the air to sounds in the cochlear fluid. Then, the vibration of the stapes, that is produced by sounds, puts oscillating pressure on the membrane covering of the oval window and sets up vibratory waves in the fluid in the scala vestibuli of the cochlea. These waves are transmitted throughout the cochlea, ultimately causing oscillations of the membrane of the round window of the scala tympani. These fluid oscillations cause the basilar membrane to vibrate, which in turn causes small deflections of the *stereocilia* of the cochlear hair cells [76].

4.2 Cochlear Mechanics

The stapes produces vibration of the membrane of the oval window, leading to vibratory waves in the fluid environment of the cochlea. The vibratory waves in turn produce a vibration of the basiliar membrane. Because of its properties, the maximum

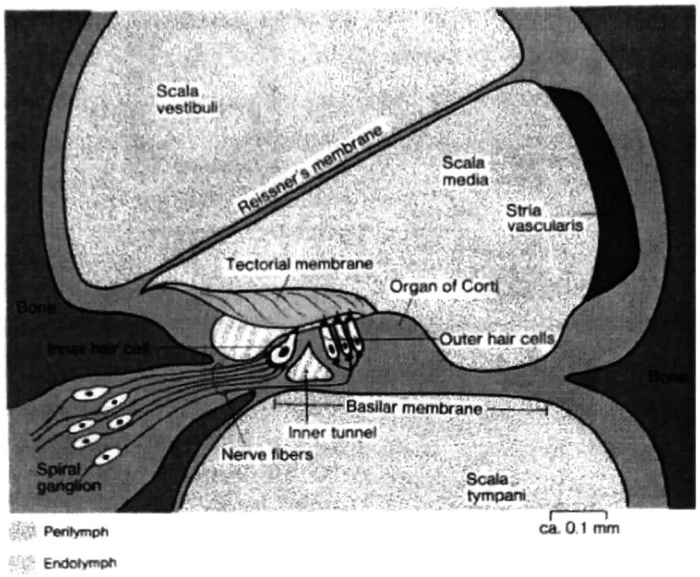


Figure 4.2: *Anatomy of the cochlea. This image is a magnification view of a transversal cut of one turn of the coil.*

vibration of the basiliar membrane occurs at specific locations depending on the sound frequency, see Figure 4.3.

The physical property that determines the basiliar membrane function is the membrane's width. The membrane is narrowest at beginning the of the cochlea nearest the round and oval windows and becomes progressively wider towards the apex. The differences in width result in differences of the tautness of the membrane per unit area. The more taut, the narrow portion of the basilar membrane resonates at high frequencies, whereas the less taut, the wide portion in the apex resonates at low frequencies. The actual spatial pattern of vibration is in the form of a *traveling wave* [76]. The basiliar membrane vibrates at the same frequency of a tone being heard. The deflections begin at the oval window and travel in a wavelike fashion down the cochlea toward the apex, hence the term traveling wave. The shape of the traveling wave, the distance that it travels, and its position of maximal deflection are all dependent of the sound frequency. High frequencies produce narrow waves that are maximal near the oval window and are quickly damped in more apical regions. Low frequencies produce much wider waves that travel further toward the apex and produce maximal deflections apically. From

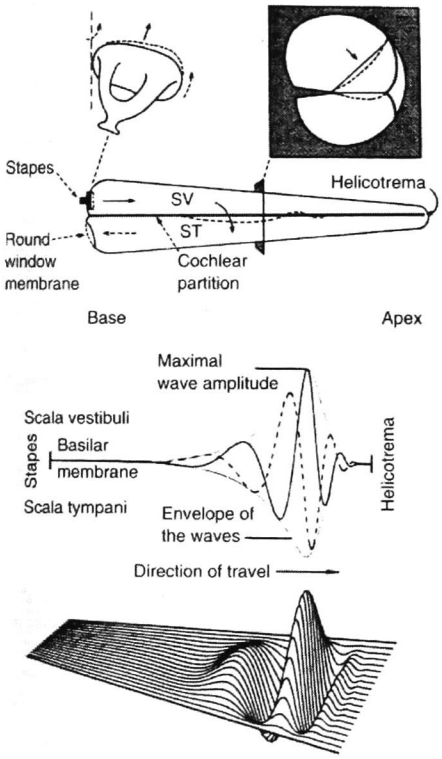


Figure 4.3: *The basilar membrane. The vibration of the stapes produces vibratory waves in the fluid environment of the cochlea. The figure presents an uncoiled illustration of the cochlea and a wave traveling.*

this we conclude the extraordinary operation of the basilar membrane as a mechanical acoustic spectrum analyzer, transforming the displacement of the oval window into a spatial array of basilar membrane deflections. Transduction of vibrational energy into neural activity occurs in the organ of Corti, a highly specialized neuroephitelial sheet made up of of sensory receptor cells, called *hair cells*. Hair cells are so called because of the *stereocilia* that extent from the apical surfaces of the hair cells to become embedded in the overlying *tectorial membrane*. There are two types of hair cells, *inner* and *outer*, that refer to the relative proximity to the *modiolus*. Inner cells are found on one row in one side of the row of inner pillar cells, the outer hair cells are found in three rows on the opposite side of the inner pillar cells. The stereocilia of hair cells are anchored to the gelatinous material of the tectorial membrane. As a result, when the basilar

membrane vibrates in response to sound, the hair cells are deflected. The morphology of the stereocilia varies along the length of the organ of Corti. The stereocilia at the basal end of the cochlea (high frequency) are very short, whereas the stereocilia at the apex (low frequency) are long. These physical differences correspond to frequencies of vibration that the stereocilia experience [76].

4.2.1 Tone Behavior

The basilar membrane is complex non linear structure. For clarity, it is convienient to separate its behavior in the presence of single-tone sounds (pure sinusoids) from that of multi-tone sounds (sum of sinusoids). We will attempt to simulate the mechanical behavior of the cochlea based on macro-mechanical models of this organ. In [10] the author mentions several different models of the cochlea. Here, we present one of those models which considers the basilar membrane as a two dimmensional tapered transmission line, i.e. an electrical transmission line, whose geometry and impedance changes with length [25].

Single-Tone Behavior

As previously mentioned, the basilar membrane length x varies between 0 and 35 mm and the human audition range is 20 to 20,000 Hz. Then, the displacements of the membrane, considered relative to the base in the oval window ($x = 0$), can be computed using the expression

$$d(f,z) = e^{f^2[5.765\cdot 10^{-10} - 0.0039(1+100z)e^{-100z}]}\sqrt{(0.0975\, f^2\, e^{-100z})^2 + (0.033125\, f\, e^{-20z})^2} \tag{4.1}$$

where

$$z = \sqrt{0.035 - x}$$

Here f is in Hz and x in m. Several relative envelope magnitudes are presented in Figure 4.4, for $f = 18000, 4900, 990, 150$ Hz sinusoidal tones. The curves have been normalized with respect to their maxima. They are strongly magnified in comparison with actual physiological membrane displacements.

To each input tone corresponds a well localized maximal deflection of the basilar membrane along its length. In Figure 4.4, the peak locations appear at 0, 10, 20 and

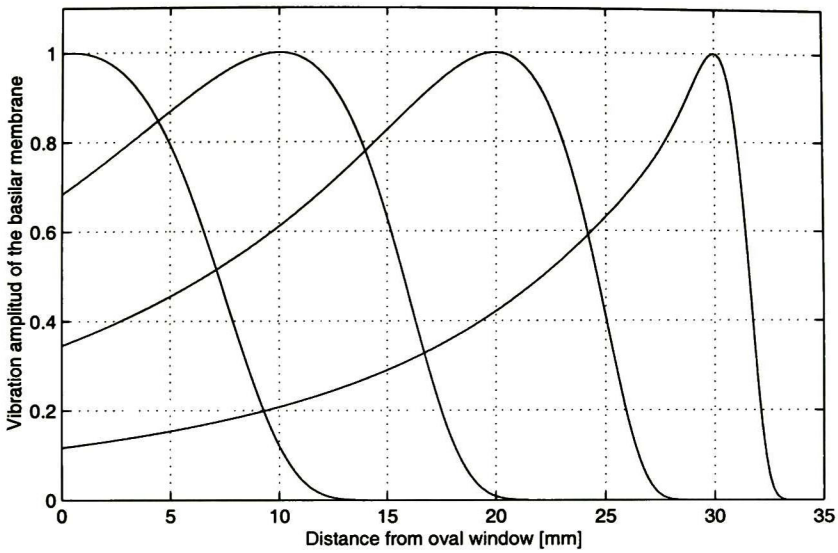


Figure 4.4: *Envelope of the traveling wave appearing on the basilar membrane for* $f = 19000$, 4900, 990 *and* 190 *Hz (from left to right).*

30 mm from the oval window for the four decreasing frequency values. Thus, as already mentioned, the basilar membrane behaves as a nonuniform transmission line performing a coarse spectrum analysis of the sounds fed into the ear.

By varying f in Equation 4.1 instead of x, we obtain the tuning curves presented in Figure 4.5. The membrane locations have been chosen such that the resonance frequencies are $f = 19000$, 4900, 990 and 190 Hz, as previously, leading to $x = 4.675$, 12.9, 22.1 and 29.9 mm. With the exception of the last location value ($x = 29.9$ mm), it is easy to see that the resonance locations are situated behind the points where the maximal basilar membrane displacements occur, i.e., the peak of the traveling wave does not exactly correspond to the resonant site of the cochlea.

Multi-Tone Behavior

In a real listening context the ear is submitted to complex sound excitations like, music or speech, rather than to single tone experimental situation described in the previous subsection. The basilar membrane envelope behavior for multiple tone combinations is given by the *nominal limit* of motion of the membrane [83]. This value is

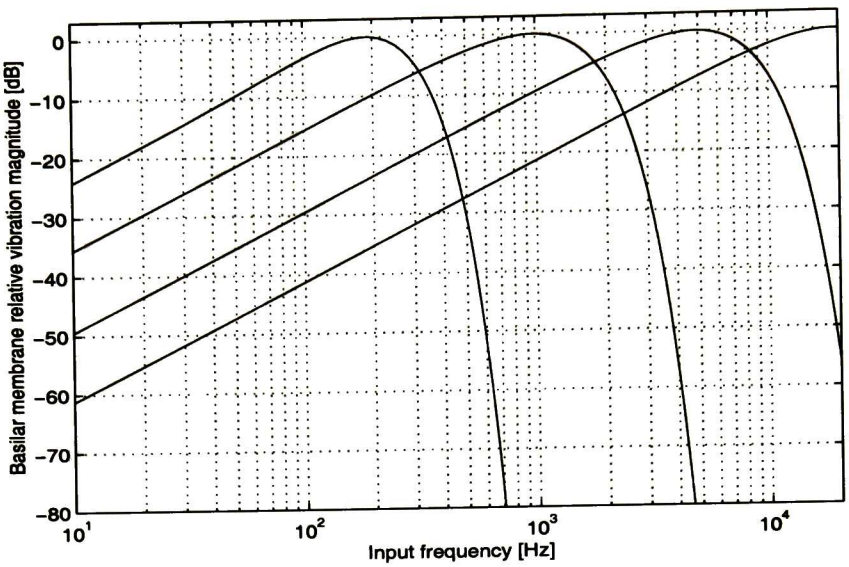


Figure 4.5: *Envelope of the traveling wave appearing on the basilar membrane versus input frequency for x =4.675, 12.9, 22.1 and 29.9 mm (from left to right).*

calculated by taking the square root of the sum of the squared envelopes due to each excitation tone. The addition rule of the of individual components is based on energy considerations [10, 83]. The basilar membrane presents a mechanical deflection which is an image of the sound power (as verified by transmission line models). Figure 4.6 shows the idealized multi-tone response of the basilar membrane to the combination of the four tones previously employed at f = 19900, 4900, 990 and 190 Hz. It can still be seen that the peaks due to the individual components still remain well localized along the membrane.

4.3 Masking

Masking refers to a process where one sound is rendered inaudible because of the presence of another sound. Masking plays a very important role in our life, for example, for a conversation on a quiet street little speech power is necessary for the speakers to understand each other. However, if a loud car passes by, the conversation is severely disturbed: by keeping the speech power constant, the listener can no longer hear the speaker. One way of overcoming this phenomenon of masking is by raising the speaker's

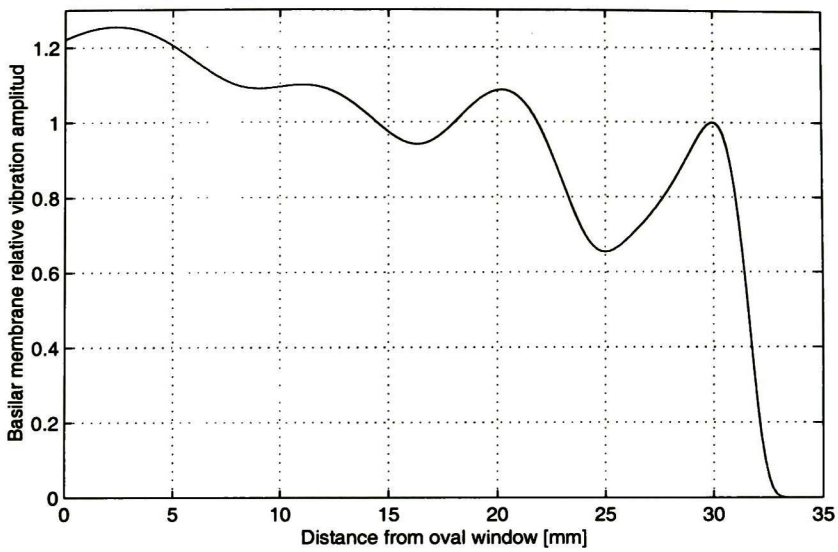


Figure 4.6: *Envelope of the traveling wave appearing on the basilar membrane for a combination of* $f = 19000$, 4900, 990 *and* 190 *Hz tones.*

voice to produce more speech power and consequently greater loudness. Similar effects take place in most pieces of music. One instrument may be masked by another if one of them produces high levels while the other remains faint [60, 83].

To measure the effect of masking quantitatively, the *masking threshold* is usually determined. The masking threshold is the sound pressure level of a test sound (normally a sinusoidal test tone), necessary to be just audible in the presence of other sounds. The notion of masking threshold is very important in the context of signal coding, since such threshold determines how much quantization noise can be introduced in the coding process without producing audible disturbances. The added noise is said to be masked if it lies below the masking threshold. Some thresholds are static and are strictly related to the physiological resolution of the auditory system, such is the case of the *absolute threshold of hearing.* On the other hand, dynamic thresholds take into account that fact that the presence of some sounds can raise the level or threshold at which other simultaneous sounds are normally perceived. In the psycoacustical literature, we usually refer to the former sounds as *maskers*, while the later ones are the *maskees.* Such thresholds always lie above the absolute threshold of hearing. The masking sounds may be wide-band noises, narrow-band noises, pure tones, harmonic series or even

combinations of them [40, 60, 83].

4.3.1 Absolute Threshold of Hearing

The absolute threshold of hearing, also known as *threshold in quiet* or *absolute threshold*, characterizes the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment [60]. This threshold is frequency dependent and is typically expressed in terms of dB SPL, as can be seen in Figure 4.7. It is measured in quiet by presenting pure tones to subjects either through earphones or free field conditions. Further information on how to measure this limit is provided in [83]. The free field conditions are more realistic, since it takes binaural perception and body diffraction (head, pinna and ear canal) into account. The reproducibility of the threshold in quiet for a single subject is high an lies normally within ±3 dB. As can be seen in Figure 4.7, the maximal acoustical sensitivity is reached between 2 and 5 kHz, which roughly corresponds to the mechanical resonance frequency of the ear canal. Hearing deteriorates with age, producing a raise of the absolute hearing threshold, mainly at frequencies above 1 kHz. The threshold presented in Figure 4.7 corresponds to a sensitive young listener (20 years approximately), provided that he has not been exposed to sound levels that produce a hearing loss [83].

The absolute threshold of hearing is well approximated [60] by the non linear function

$$T(f) = 3.64\left(\frac{f}{1000}\right)^{0.8} - 6.5e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 10^{-3}\left(\frac{f}{1000}\right)^4 \quad \text{[dB]} \tag{4.2}$$

where f is in Hertz. In what follows, Equation 4.2 will be used to perform operations where a model of the hearing threshold is required. A real implementation of this threshold is a very delicate task, since it requires perfect knowledge of the play-back sound level. This requires tracing signal processing systems to sound pressure levels at the listener and fixing electroacoustical transducers and level controls [10].

4.3.2 Masking of Pure Tones by Broad-Band White Noise

The analytical properties of the auditory system can be investigated by measuring its ability to detect pure tones in the presence of wide-band white noise. This can be done by generating a white noise signal at a given level L_M and then present pure tones to listening subjects [83]. The pressure level L_m at which tones are just perceived

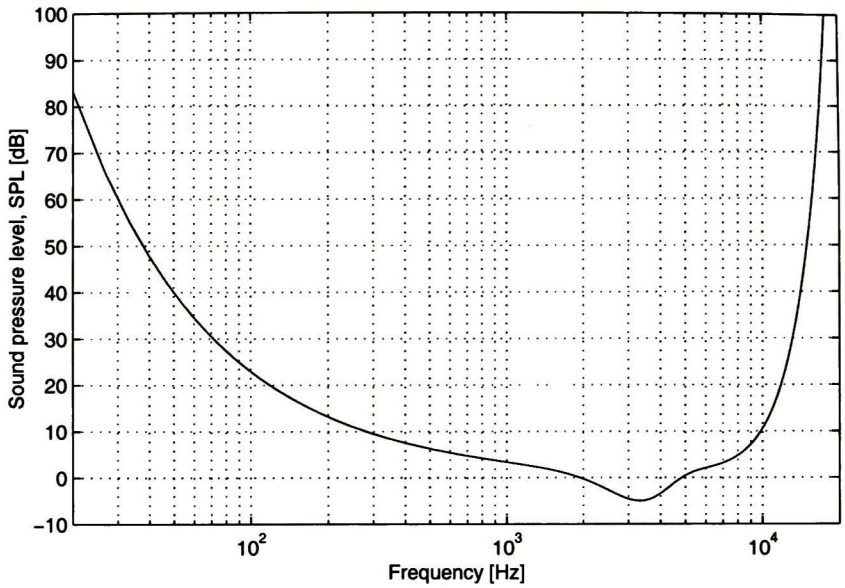


Figure 4.7: *Idealized absolute threshold of hearing.*

defines the masking threshold under white noise conditions. Simplified curves obtained through this process are shown in Figure 4.8 for different *background noise* levels. Such curves represent an approximation to the psychoacoustical results presented in [83] and are computed here as the maximum between the absolute threshold of hearing and the expression

$$L_m(f) = 10\log(1 + \frac{f}{f_0}) + L_M + 17 \quad [\text{dB}] \tag{4.3}$$

where $f_0 = 500$ Hz. We can highlight several points after observing the thresholds of Figure 4.8

- The curves are parallel and equidistant to each other.
- They meet the threshold in quiet at very low and very high frequencies.
- They uniformely rise by 10 dB as the noise level is increased by 10 dB.
- They are approximately flat up to 500 Hz.
- They increase by approximately 10dB par decade beyond 500 Hz.

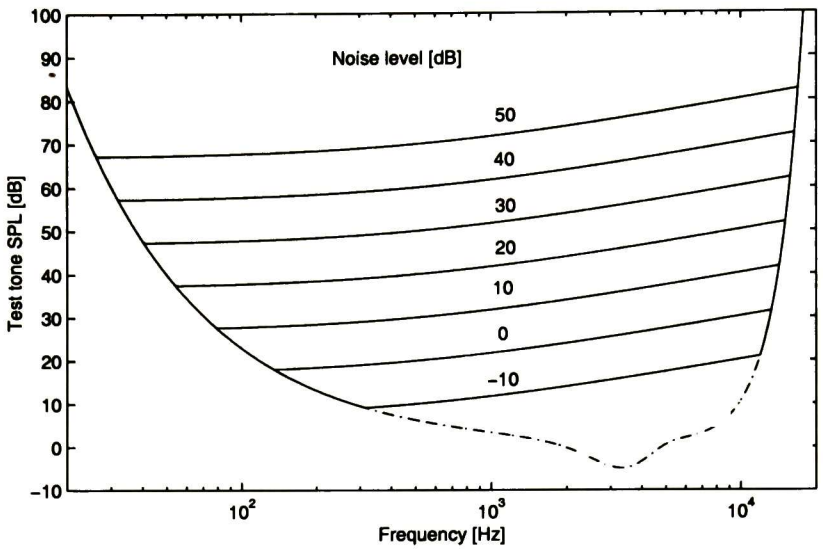


Figure 4.8: *Detection of pure sine waves in wide-band white noise. The dashed curve represents the absolute threshold of hearing.*

Masking of a tone by broad-band noise is normally accomplished by a set of frequency components in a band surrounding the test tone [10]. According to the curves depicted in Figure 4.8 it can be deduced that the width of such bands must be almost constant up to 500 Hz and then, since masking becomes stronger, increase by a certain factor beyond that frequency. This bands are called *critical bands* [60, 39, 83]. For a given test tone, the only noise spectral components contributing to the masking of the tone are those within the critical band centered around it. Increasing the bandwidth of the noise beyond the critical bandwidth does not change the threshold at which the centered tone is detected.

There are approximately 25 critical bands over the entire hearing frequency range. A list of these bands is presented in Table 4.1 [10, 39, 44, 60, 83]. It must be noted that this table may naturaly arise from the frequency mapping that performs the basilar membrane. A difference of one critical band is also called a *Bark* [4]. Thus, the Bark scale is a synonym of critical subband decomposition. Additionally, the Bark scale is directly related to places along the basilar membrane. To each 1.4 mm segment in the basilar

[4]Named after the German scientist A. G. von Barkhausen who introduced the concept of loudness level in the 1920s.

membrane corresponds an increment of 1 Bark, i.e., the critical band decomposition is a *linear scale* on the basilar membrane. As mentioned, the *critical bandwidth* (CBW)

Subband number	Lower edge [Hz]	Center [Hz]	Upper edge [Hz]	Bandwidth [Hz]
0	0	50	100	100
1	100	150	200	100
2	200	250	300	100
3	300	350	400	100
4	400	450	510	110
5	510	570	630	120
6	630	700	770	140
7	770	840	920	150
8	920	1000	1080	160
9	1080	1170	1270	190
10	1270	1370	1480	210
11	1480	1600	1720	240
12	1720	1850	2000	280
13	2000	2150	2320	320
14	2320	2500	2700	380
15	2700	2900	3150	450
16	3150	3400	3700	550
17	3700	4000	4400	700
18	4400	4800	5300	900
19	5300	5800	6400	1100
20	6400	7000	7700	1300
21	7700	8500	9500	1800
22	9500	10500	12000	2500
23	12000	13500	15500	3500
24	15500	19500		

Table 4.1: *Idealized critical band distribution.*

up to about 500 Hz is relatively constant and has a value of 100Hz. Then, it increases at about 20% of the central frequency. In [83] the author provides a good analytical approximation given by the expression

$$CBW(f) = 25 + 75\left(1 + 1.4 \cdot 10^{-6} f^2\right)^{0.69} \quad \text{[Hz]} \tag{4.4}$$

where f stands form the center frequency in Hz. This curve, as well as the bandwidhts given in Table 4.1, are plotted in Figure 4.9 versus center frequency values. As can be

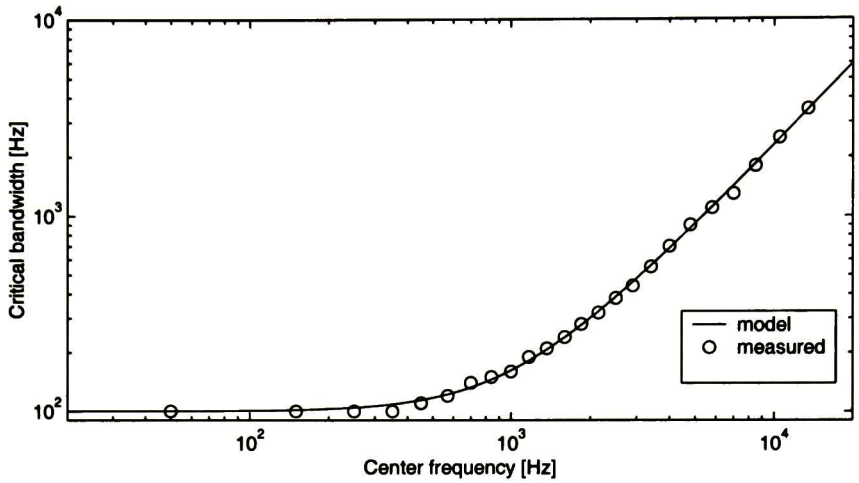


Figure 4.9: *Critical bandwidth vs. center frequency.*

seen, the bandwidth progression is approximately linear up to about 500 Hz, and the approximately logarithmic beyond 500 Hz. The second part corresponds to constant relative bandwidth, and therefore constant-Q filters.

Also provided in [83], an approximation of the relationship between frequency and critical band numbers, given by the *critical band rate* (CBR) expression

$$CBR(f) = 13 \arctan\left(7.6 \cdot 10^{-4} f\right) + 3.5 \arctan\left[\left(1.33 \cdot 10^{-4} f\right)^2\right] \quad \text{[Bark]} \qquad (4.5)$$

where f stands for the center frequency in Hz. This relation is shown in Figure 4.10, superimposed on the center frequency values of Table 4.1. The first center frequency of the table is located at 0.5 Bark and next evolves at 1 Bark intervals.

4.3.3 Masking of Pure Tones by a Narrow-Band White Noise

Now that we have determined the importance of the Bark scale in the perceptual domain, we will find what happens if the masker signal is a narrow-band white noise with a bandwidth no longer than the critical bandwidth. We must determine how the corresponding masking thresholds spreads over the frequency domain. The results of the corresponding psychoacoustical measures have been presented in [83] and one example for a masker centered at 1000 Hz is shown in Figure 4.11. In the figure, L_{CB} stands for the level of the noise within one critical band.

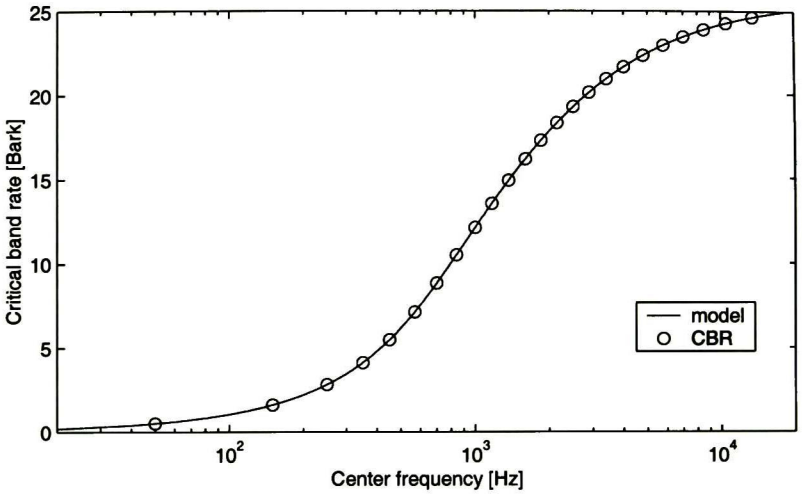


Figure 4.10: *Critical band rate vs. center frequency.*

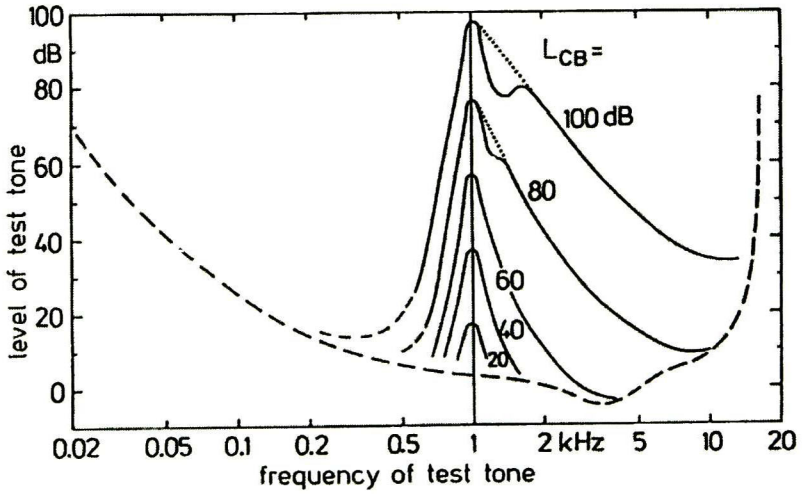


Figure 4.11: *Masking of pure tones by narrow-band white noise. The masker noise is centered at 1000 Hz.*

The masking thresholds produced by narrow-band noise maskers have the following characteristics

- The curves are triangle-shaped.
- They meet the absolute threshold of hearing at very low and very high frequencies.
- They shift by the same amount as the noise level.
- They lie a few dB's below the noise level.
- They are asymmetric, showing a steep rise from lower frequencies up to the location of the masker and a slower decrease from this point up to the higher frequencies.
- Their upward slope becomes shallower as the level of the masker increases (L_{CB} = 80 and 100 dB), it is a nonlinear level dependent mechanism.
- They show additional peaks at higher masker levels, which are due to audible difference noises arising from the interaction between the test tone and the narrow-band noise.

Another important property of the masking patterns is that their global shape depends on the central frequency of the masker. This can be seen in the top of Figure 4.12, where the masking thresholds corresponding to 4 maskers, localized at 150, 1000, 4000 and 11000 Hz, are presented. The curves correspond to maskers at $L_{CB} = 60$ dB. This frequency dependence disappears when the frequency axis is converted to the Bark scale, as can be seen in the bottom of Figure 4.12. Thus, masking patterns, centered at different frequencies are, very similar when represented over a Bark scale. The triangled-shaped curve that approximates the masking pattern is called the *spreading function* and has been proposed by [26], where the following mathematical expression, for the Bark scale, was suggested:

$$B(z) = a + \frac{l+u}{2}(z - z_i + c) - \frac{l-u}{2}\left[t + (z - z_i + c)^2\right]^{1/2} \tag{4.6}$$

where the critical band rate z is in the range 0-34 Barks, with fractional values allowed. Critical band z_i corresponds to the center frequency of the narrow-band masker. The meanings of the remaining parameters in $B(z)$ are the following:

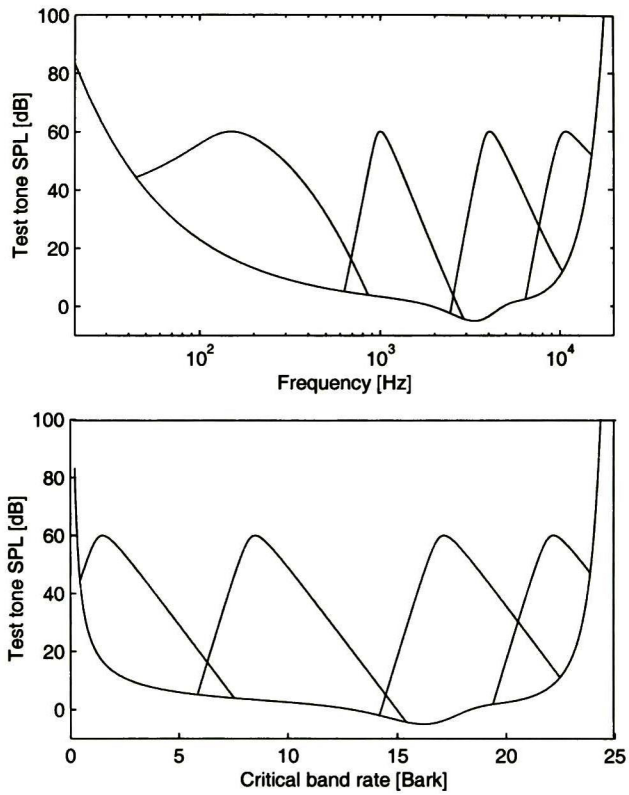


Figure 4.12: *Masking threshold vs. frequency (top) and critical band rate (bottom) for 150, 1000, 4000 and 12000 Hz center frequencies (from left to right).*

- l is the lower critical band rate slope.
- u is the higher critical band rate slope.
- t is a damping factor that defines the flatness of the peak.
- a and c determine the peak level and position of the pattern.

Some typical values for these parameters are provided in [10], which are presented in Table 4.2. Pattern II is intended to underestimate masking towards higher frequencies. Pattern III was proposed in an attempt to build 1-Bark wide auditory filters, after noting that Pattern I was 1.4-Bark wide at an attenuation of -3 dB. As we will see later, the ISO/MPEG standard [37] uses picewise linear masking patterns, which also depend on the lever of the masker.

Source	a [dB]	l [dB/Bark]	u [dB/Bark]	c [Bark]	t [Bark^2]
I	15.81	25	-10	0.474	1
II	13.94	27	-24	0.03	0.3
III	7.00	25	-10	0.215	0.196

Table 4.2: *Typical values of the spreading function parameters.*

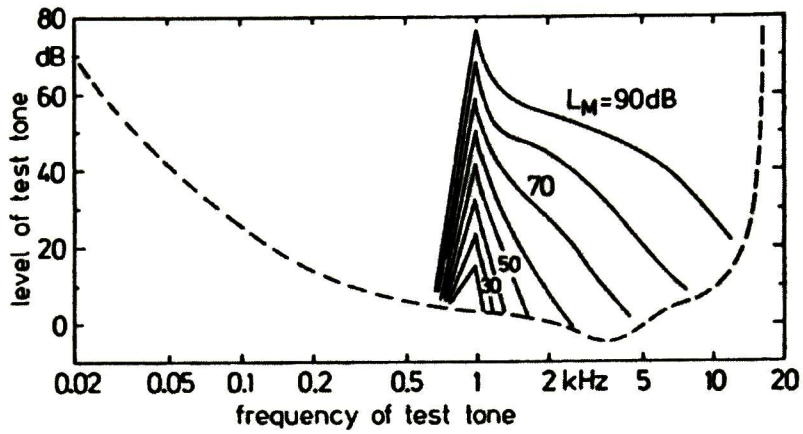


Figure 4.13: *Masking of pure tones by tones.*

4.3.4 Masking of Pure Tones by Tones

The behavior of masking patterns associated with pure tones that mask other pure tones is far more complex than the spreading function previously described. They are very difficult to measure due to the strong perception of beats, as well as of difference tones, by the test subjects [83]. This mainly happens for test tones close to the masker or its harmonics. Some differences in the slopes of the upper and lower skirts can be observed in Figure 4.13, reprinted from [83]. Lower slopes are steeper for the tone masker than for the noise masker; on the other side, higher slopes are shallower for the tone masker than for the noise masker. [10] presents a masking model derived from the data presented in [83]. Masking patterns for pure tone maskers roughly show a triangular shape; just like the band noise maskers, however their maxima is located at a lower level.

Narrow-band masking thresholds appear at a level of a_{tmn} dB below the noise level $L_C B$. For pure tones that at level L_T, which are masking other pure tones, this happens at about

$$a_{tmt}(z) = -0.275z - 6.025 \qquad \text{[dB]} \tag{4.7}$$

below the level L_T. Here, *tmt* stands for *tone-masking-tone*. The slopes of the spreading function also show a very different behavior with varying the masker level. Over the Bark scale, at lower masker levels, masking spreads more towards lower frequencies than towards higher frequencies [10]. As can be seen in Figure 4.13, at a masker level of 40 dB the masking patterns are approximately symmetrical. At higher levels the spreading of masking behaves almost like masking due to narrow band maskers. If we adopt a conservative model which neglects the lower frequency spread of masking at lower masker levels, the masking threshold can be estimated by calculating the maximum between the threshold in quiet and

$$T(z, z_m) = B(z) + L_T + a_{tmt}(z) \qquad \text{[dB]} \tag{4.8}$$

where $B(z)$ is the triangular function given by Equation 4.6. Pure tones rarely exist in nature. The most common "pure" tones sounds are composed of a fundamental tone and its harmonics, such is the case for musical intruments. Thus, the masking curve spreads over the frequency range where the tones are located. A narrow band noise in a critical band can be approximated by a small number (five or more) of equal-amplitude pure tones, with randomly distributed frequencies over the noise band [10].

4.3.5 Masking of Narrow-Band White Noise

From a subband coding standpoint, the maskee signal is the quantization noise introduced by the system in different subbands and which must be masked. In the two previous subsections we have discussed only cases where the maskee is a tonal signal, two more masking cases arise from the possible noise/tone combinations

- Tone-masking-noise.
- Noise-masking-noise.

This two situations have received little attention in psychoacoustic research, partly due to the difficulty of making such measurements [10].

After numerous informal listening tests, [10] proposes that the value of the masking index for tone masking noise is

$$a_{tmn} = -9.0 + a_{tmt}(z) \qquad \text{[dB]} \tag{4.9}$$

which is a 9 dB shifted version of the tone-masking-noise. According to [10] this is a non conservative approach justified by the observation that natural sounds possess a more noise-like structure towards the higher frequencies. Thus, the tone-masking-noise masking threshold induced in a critical band z by a pure tone, located at z_m Bark, is approximated by the expression

$$T(z, z_m) = B(z) + L_T + a_{tmn}(z) \qquad \text{[dB]} \tag{4.10}$$

The lack of literature concerning the noise-masking-noise case implies that one has to rely on the results for tone maskers and try to find an empiric masking model for the noise-masking-noise case. A common phenomenon in all masking situations is the upward spreading of masking, i.e., masking is stronger towards higher frequencies. In that case, it is possible to maintain the global shape of the threshold given by $B(z)$ in Equation 4.6. In [10] it is proposed that the masking index should lie somewhere below the tone-masking-tone case and above the tone-masking-noise case; this comes from observations that noise is better masker than a tone and that a tone is a better maskee than noise. We must point out that in the noise-masking-noise situation both masker and maskee have the same bandwidth. This remark suggests that the value of the masking index should not depend on the Bark rate. In the present work we assume a constant value of

$$a_{nmn} = -9.0 \qquad \text{[dB]} \tag{4.11}$$

For comparison, [43] proposes a value of a_{nmn}=-5.5 dB. Finally, the noise-masking-noise masking threshold induced in a critical band z by a pure tone, located at z_m Bark, can be approximated by the maximum between the threshold in quiet and the expression

$$T(z, z_m) = B(z) + L_T + a_{nmn}(z) \qquad \text{[dB]} \tag{4.12}$$

So far, we have assumed that the masker and the maskee are stationay signals. From a stationay standpoint, the stimuli can be considered stationay if their duration is at least of 200 ms. At longer durations the masking threshold previously reviewed are stable.

4.3.6 Nonsimultaneous Masking

As shown in Figure 4.14, masking phenomena extend in time beyond the window of simultaneous stimuli presentation. It means that, for a masker of finite duration, non-simultaneous (some times denoted as *temporal*) masking occurs both prior to masking onset and after masker removal. The skirts of both regions are presented in Figure 4.14. Essentially, absolute audibility thresholds for masked sounds are artificially increased prior to, during and following the occurrence of a masking signal. Pre-masking alone plays a relatively secondary role, because the effect lasts only 20 ms and therefore is usually ignored. Post-masking has an effect that will extend anywhere from 50 up to 200 ms [39, 60, 83]. Post-masking has a more important effect than pre-masking, since it has a longer duration. A deeper coverage of temporal masking can be found in [83].

Temporal masking reaches a maximum for signals close in frequency, particularly within the same critical band. The influence of forward masking increases with the bandwidth of the masker [83]. The full effect of temporal masking is very related to the duration of the masker, as well of the maskee. Maximum masking effect is produced by a masker lasting about 200 ms. Below that value, the masking threshold shows faster decay slopes and hence a shorter duration. The results provided in [83] suggest that temporal masking is a highly nonlinear effect, therefore some precaution is necessary when taking it into consideration.

In the present work, in order to reduce the complexity of the proposed coder, we have decided not to take into account the effect of temporal masking.

4.4 Perceptual Entropy

In [42, 43] Johnston combines notions of psychoacoustic masking with signal quantization principles to define the *perceptual entropy*, a measure of perceptually relevant information contained in an audio signal. Expressed in bits per sample, the perceptual entropy represents a theoretical limit on the compressibility of a particular signal. The perceptual entropy works based on the principle that the masking threshold of a signal indirectly indicates the amount of quantization noise that may be applied in the frequency domain, i.e., the quantization, according to the masking model, that may be done without corrupting the signal such that it can be distinguished from the original. The part of the signal that can be changed without making the signal distinguishable

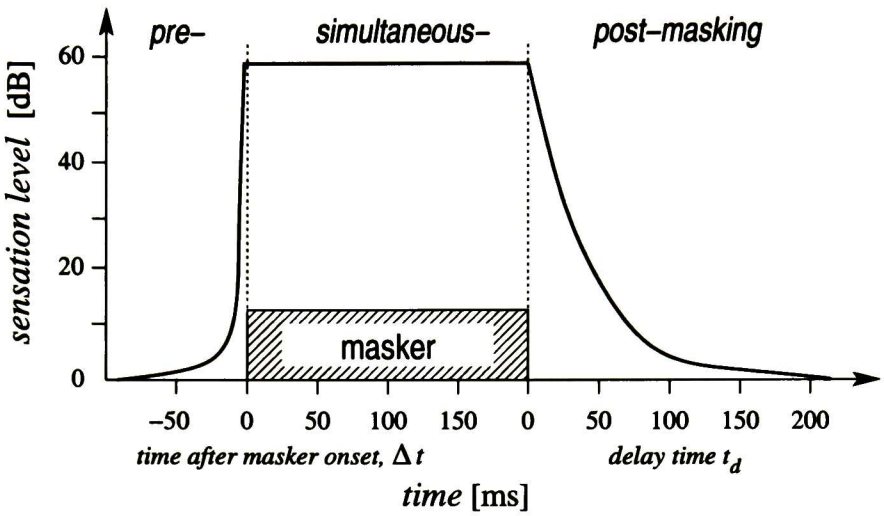


Figure 4.14: *Behaviour of temporal masking. Pre-masking occurs prior to masker onset and lasts only a few milliseconds; post-masking may persist up to 200 ms after masker removal.*

is therefore perceptually redundant, and the part that must be reproduced represents real information that can be quantized and measured [42].

The process to estimate the perceptual entropy is accomplished as follows.

1. The signal is windowed using a Hann window, $w(n) = \frac{1}{2}\left(1 - \cos\frac{2\pi n}{M-1}\right)$, where M stands for the frame length in samples and n for the sample number. Then, a 2048-point fast Fourier transform (FFT) is performed.

2. Real and imaginary transform components are then used to calculate the power spectral density (PSD) components

$$P(\omega) = (\mathrm{Re}(\omega))^2 + (\mathrm{Im}(\omega))^2 \tag{4.13}$$

then a discrete Bark spectrum is formed by summing the energy in each critical band (see Table 4.1)

$$P_i = \sum_{\omega=bl_i}^{bh_i} P(\omega) \tag{4.14}$$

where the summation limits are the critical band boundaries. The range of index i is sample rate dependent. For the present work $i \in \{0,\ 20\}$.

3. The spreading function $B(z)$, presented in Equation 4.6, is convolved with the discrete Bark spectrum

$$C_i = P_i \star B_i \tag{4.15}$$

to account for the spread of masking.

4. An estimation of the tone-like or noise-like structure of C_i is obtained using the spectral flatness measure (SFM)

$$SFM = \frac{\mu_g}{\mu_a} \tag{4.16}$$

where μ_g and μ_a, respectively, stand for the geometric and arithmetic means[5] of the PSD components for each band. The SFM has the property that it is bounded by zero and one. A coefficient of tonality α is next derived from the SFM on a dB scale

$$\alpha = \min\left(\frac{SFM_{dB}}{SFM_{max}}, 1\right) \tag{4.17}$$

where $SFM_{max} = -60$ dB. An SFM close of zero dB indicates that the signal is completely noise-like, on the contrary if the SFM is close to SFM_{max} indicates that the signal is entirely tone-like.

5. Then, the index α is used to geometrically weigh the two threshold offsets[6]

$$O_i = \alpha(14.5 + i) + (1 - \alpha)5.5 \qquad \text{[dB]} \tag{4.18}$$

6. A set of JND estimates in the frequency power domain are then formed by substracting the offsets from the bark spectral components

$$T_i = 10^{\log_{10}(C_i) - \frac{O_i}{10}} \tag{4.19}$$

7. Due to the fact that the spread spectrum functions do not have a normalized gain, and that the gains in the critical bands around zero and the sampling rate will be different, the spread spectrum is renormalized by $\frac{1}{\text{the DC Gain}}$ for each critical band.

---

[5] $\mu_a = \frac{1}{M}\sum_{j=1}^{M} k_j$ and $\mu_g = \left(\prod_{j=1}^{M} k_j\right)^{1/M}$

[6] Notice that we have re-printed the values given in [42], thus they do not match the masking values previously discussed.

i	index of critical band;
bl_i and bh_i	upper and lower bounds of band i;
k_i	number of transform components in band i;
T_i	masking threshold in band i;
$nint$	rounding to the nearest integer.

Table 4.3: *Meaning of the parameters presented in Equation 4.20.*

8. Then, the resulting Bark threshold is compared to the absolute threshold of hearing, to make sure that they do not demand a level of noise below the absolut limits of hearing. The absolute threshold of hearing is set such that a signal at 4 kHz, with a peak magnitude of ± 1 least significant bit in a 16 bit integer is at the absolute threshold of hearing.

9. By applying uniform quantization principles to the signal and its associated set of JND estimates, it is possible to estimate a lower bound on the number of bits required to achieve transparent coding biven by the expression [42, 43, 60]

$$PE = \sum_{i=0}^{20} \sum_{\omega=bl_i}^{bh_i} \log_2 \left[2 \left| nint \left(\frac{\mathrm{Re}(\omega)}{\sqrt{6T_i/k_i}} \right) \right| + 1 \right] + \log_2 \left[2 \left| nint \left(\frac{\mathrm{Im}(\omega)}{\sqrt{6T_i/k_i}} \right) \right| + 1 \right] \quad \text{[bits/sample]} \tag{4.20}$$

where the meaning of each parameter in Equation 4.20 is presented in Table 4.3

The perceptual entropy measured is obtained by constructing a perceptual entropy histogram over many frames and then choosing a worst case value as the actual measurement.

4.5 Masking Threshold Implementation

The psychoacoustic model, used to calculate the masking threshold, plays a central role in the compression technique in the present work. Some of the most important masking models existing were reviewed [6, 12, 13, 27, 37, 43, 49, 63, 64, 71, 74, 75]. In [49] the authors evaluate several different psychoacoustic models and they report that ISO/IEC MPEG Standard [37] provides near transparent audio quality. Moreover, in [6, 64, 74] the authors employ a modified adaptation of the ISO/IEC MPEG Standard

[37] with high audio quality results. For these reasons we decided to implement in our present work the ISO/IEC MPEG Standard masking model II [37]. It is important to point out that all the psychoacoustic models reviewed are linear, in which case the addition of masker components often results in a much lower overall threshold than determined experimentally. If the reader is interested in more sophisticated masking models, [4] provides a non linear psychoacoustic model, which better approximates the HAS processing.

The main operations involved in the calculation of the masking threshold are the following (see Appendix E for an insight of the MPEG psychoacoustic model II)

1. Blocking and windowing the signal. A Hann window is used.
2. Calculation of the Bark spectrum energy of the current signal block, over a discretized bark scale.
3. Estimation of the tonality of the current block.
4. Calculation of the simultaneous masking threshold, taking tonality into account.
5. Comparison of the resulting masking threshold with the absolute threshold of hearing, the maximum between both is the masking threshold.

Point 1 and Point 2 are exactly the same as previously seen in section 4.4, except that the length of the FFT is 1024 (See Appendix E). In Point 3, the tonality is estimated for each spectral line of the FFT based on a measure of the predictability of the signal [37]. In Point 4 the discrete Bark spectrum is convolved with the spreading function and adjusted according to the tonality index, then the result is normalized as in section 4.4. Finally, Point 5 makes sure that the calculated threshold does not demand a level of noise below the absolute limits of hearing.

In the following Figures we present some examples of the results provided by the psychoacoustic model. Figure 4.15 presents the masking pattern for a narrow-band white noise signal. We must point out that power spectrum of the noise signal is not completely flat due to the nature of the pseudo-random number generator used to generate the signal. The noise signal is located at the $14^{th.}$ critical band and its average noise power is 65 dB. As can be seen in Figure 4.15, this pattern agrees with the previous results, the lower slope is steeper than the upper slope. Moreover, the upper slope extends to the higher frequencies.

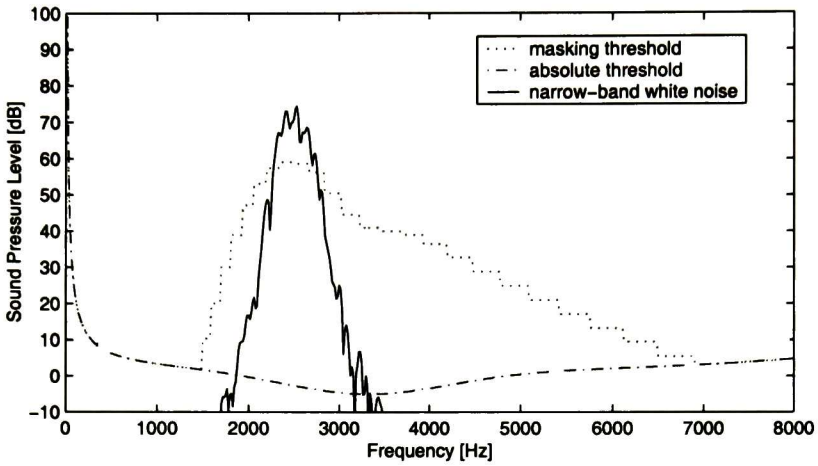


Figure 4.15: *Masking pattern (dotted line) for a narrow-band white noise signal (solid line) located in the* $14^{th.}$ *critical band (2320–2700 Hz). The average power of the noise signal is 65 dB.*

Figure 4.16 presents the tone-masking-noise case where the previously mentioned noise signal completely masks a tone signal which has a carrier frequency of 4000 Hz. The tone signal is rendered inaudible by the noise signal, therefore it can be completely removed from acoustic signal without perceiving any difference or loss.

In Figure 4.17 the tone-masking-tone case is presented. The power spectrum of an acoustic signal formed by three tones which have a carrier frequency of 1500, 3500 and 4750 hz (from left to right) and a power of 70, 85 and 30 dB respectively are depicted. As expected, the masking pattern (dotted line) has a lower level compared to the noise-masking-tone case. Moreover, the lower slopes are steeper than in the previous case, as well as the upper slopes are shallower. In this case the upper tone is still audible, but most of its power has been masked by the other two tones, therefore only a few bits are required to quantize it.

Figure 4.18 presents a voiced frame's power spectrum of female speech. As can be seen, almost half of the signal is rendered inaudible by the masking threshold (dotted line) due to masking. In the lower frequencies voiced speech has a tone-like nature, opposite to the noise-like nature of the higher frequencies.

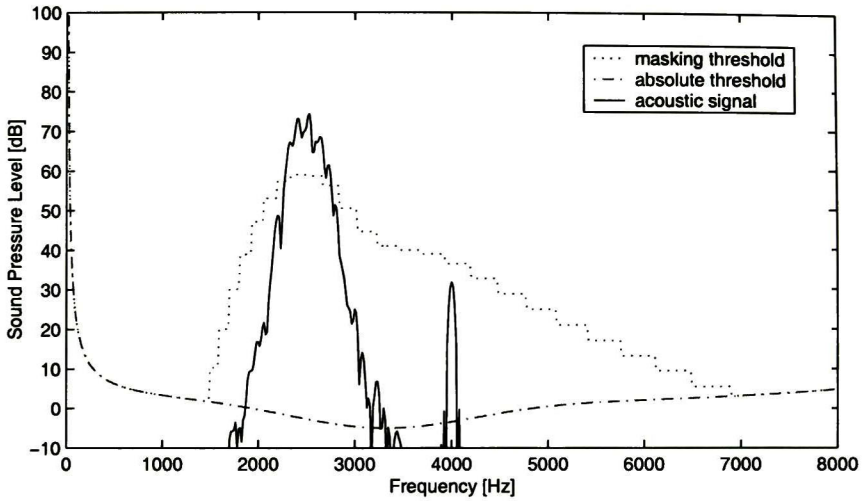


Figure 4.16: *Masking of a tone by a narrow-band noise. The noise signal is located in the $14^{th.}$ critical band (2320–2700 Hz), its average power is 65 dB. The tone signal has a carrier frequency of 4000 Hz and a power of 32 dB. The masking pattern (dotted line) generated by the noise signal completely masks the tone signal.*

Figure 4.19 presents a very similar case, but now the power spectrum corresponds to a voiced frame of male speech. As in the previous case, almost half of the signal is rendered inaudible due to masking.

Finally, Figure 4.20 presents the power spectrum of an audio (music) signal. As mentioned in Chapter 2, audio signals have a much more complex nature than speech due to the variety of signal sources. In spite of that, a considerable part of the signal is rendered partially/totally inaudible.

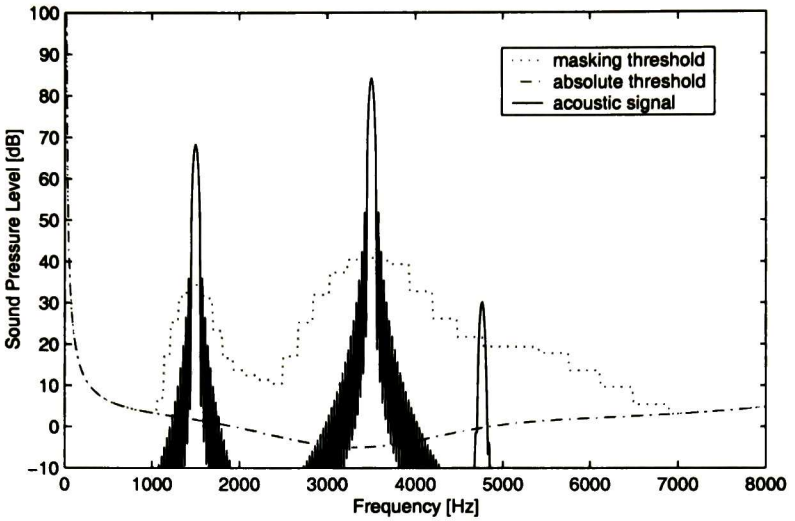


Figure 4.17: *Masking of a tone by another tone. The acoustic signal is formed by three tones of frequency 1500, 3500 and 4750 Hz (from left to right) with a power of 70, 85 and 30 dB respectively. In this case the upper tone is barely perceptible.*

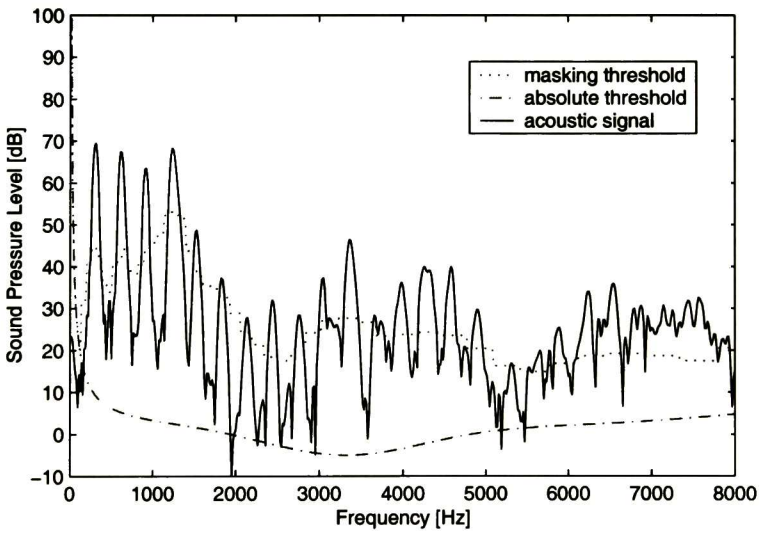


Figure 4.18: *Power spectrum of a voiced frame of female speech. The masking threshold (dotted line) shows that almost half of the signal is rendered inaudible due to masking.*

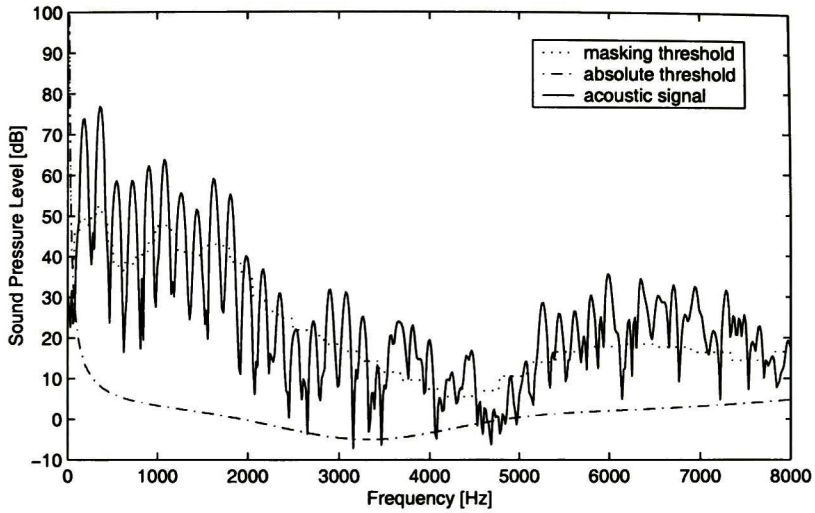


Figure 4.19: *Power spectrum of a voiced frame of female speech. The masking threshold (dotted line) shows that almost half of the signal is rendered inaudible due to masking.*

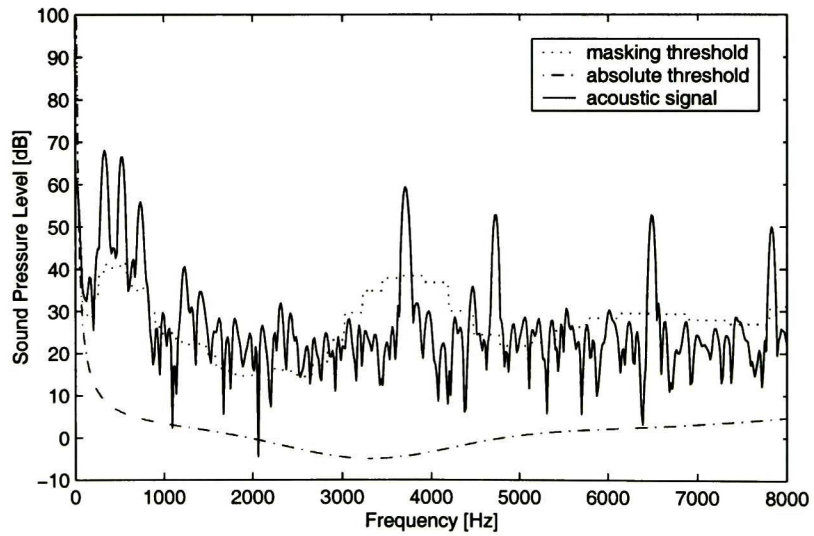


Figure 4.20: *Power spectrum of an audio (music) signal. The spectral nature of audio signals is more complex than the nature of speech, but still a considerable part of the signal is rendered partially or totally inaudible due to masking.*

Chapter 5

Perceptual Coding of Wide-Band Speech Audio

In this chapter we describe the different functions of the proposed perceptual coding system. The encoder is presented in Section 5.1. A description for each of its components is also provided. The decoder is presented in Section 5.2, where also a description of its components is provided.

5.1 Description of the Encoder

We have previously presented, in Chapter 1, a block diagram of the entire coding system proposed. Figure 5.1 presents the building blocks of the encoder. The input signal is segmented into analysis frames of 256 samples. In the time domain, the analysis frame has a length of 16 ms with 1 ms (16 samples) of overlapping between adjacent frames in order to avoid boundary effects. Due to the nature of the WT it is preferable to use an analysis frame of dyadic length. We consider that if the analysis frame has a length of 128 samples (8 ms) it is short and does not allow us to take full advantage of the quasi-stationarity property of audio signals, in addition it generates more side information to encode. On the other side, if the analysis frame has a length of 512 samples (32 ms) it is too long and may cause pre-echo problems with audio signals containing sharp attacks or abrupt transients (such as those produced by the drums, the castanets, the triangle). Therefore, an analysis frame of 256 samples provides a good compromise between stationarity, side information and pre-echo problems. Then, every frame is analyzed with a non-uniform tree-structured decomposition using the WPT.

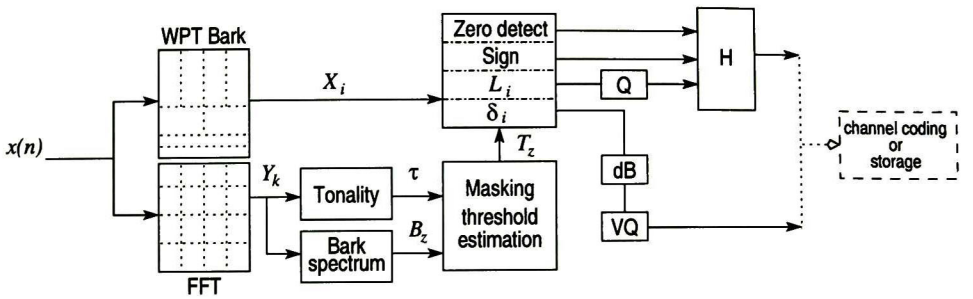


Figure 5.1: *Block diagram of the encoder.*

The tree-structured decomposition is presented in Figure 3.19, and it approximates the time and frequency resolutions of the HAS. Simultaneously, an FFT of 1024 points is also performed over the same analysis block, as depicted in Figure 5.1.

The WPT block provides the set of transform coefficients X_i, where $1 \leq i \leq 256$. The goal of this first transformation is to remove the statistical redundancy of the audio signal. The output of the FFT block are the transform coefficients Y_k, where $1 \leq k \leq 513$. The goal of this second transformation is to provide the elements to calculate the tonality index (τ) and the Bark scale spectrum (B_z, where $1 \leq z \leq 49$), as presented in Chapter 4. Once these two parameters have been computed, they are employed in the calculation of the masking threshold (see Appendix E), which serves to shape the quantization noise so it remains hidden from a perceptual stand point. This phenomenon is illustrated in Figure 5.2. In the figure we consider the case of a masking sound at the center of a critical band. For the band under consideration, the *minimum masking threshold* denotes the spreading function in-band minimum. If we assume that the masker sound is quantized using an m-bit uniform scalar quantizer, noise might be introduced at the level m. In Figure 5.2, signal-to-mask ratio (SMR) and noise-to-mask ratio (NMR) denote the log distances from the minimum masking threshold to the masker and noise level, respectively. Ideally, any distorsion lying below the minimum masking threshold will be imperceptible for the HAS. Thus, we can equate the SNR to the SMR and keep m as small as possible, this is also known as the JND level.

Next, the quantization of each coefficient X_i is achieved through a constrained bit allocation procedure. In each frame, the following information is extracted from the

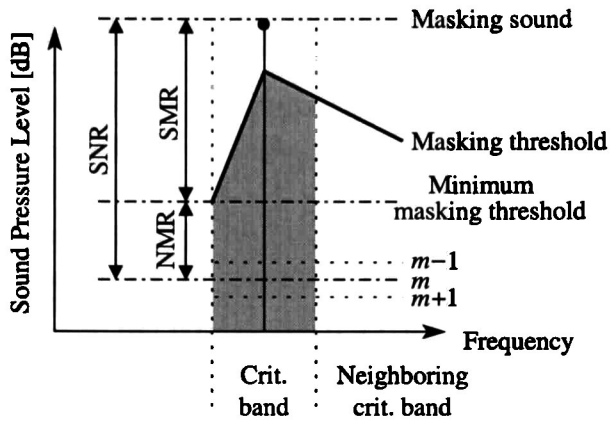


Figure 5.2: *Schematic representation of simultaneous masking. Ideally, we can equate the SNR to the signal-to-mask ratio (SMR) and keep m as small as possible.*

256 WPT coefficients:

- Zero values.
- Sign.
- Number of bits. Represented by the number of quantization levels.
- Quantization step δ_z, in the critical band.

These parameters must be encoded and transmitted to the decoder, as we will show next.

5.1.1 Optimal Bit Allocation of Transform Coefficients

As previously mentioned, the bit allocation procedure must be done in such a way that the introduced distortion (quantization noise) remains imperceptible. The masking threshold (T_i) completely satisfies this constraint, it suffices to restrict the quantization noise to lie at or below the masking threshold level to achieve transparent encoding of the audio signal. Without imposing the additional constraint of a constant bit rate, there are two possible ways of meeting this requirement. The first is based on the NMR of each noise source in the transform domain [10]. This method uses Lagrange multipliers to minimize the NMR. The second method, which we decided to employ in

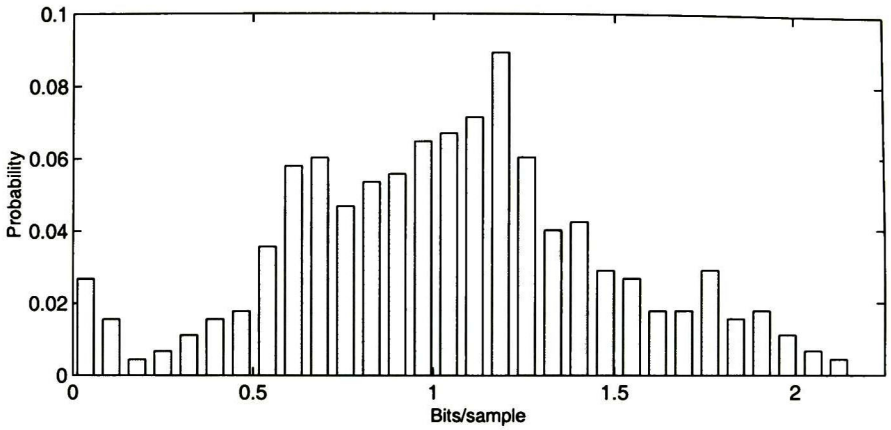


Figure 5.3: *Histogram presenting the Perceptual Entropy values of several wide-band speech audio sources. This result suggests a lower bound of about 2.1 bits/sample for transparent audio coding.*

this work, is much simpler. If all the coefficients are uniformely quantized, then, the quantization step of coefficient X_i is given by the expression

$$\delta_i = \sqrt{12\,\sigma_{q_i}^2} \tag{5.1}$$

where we make the assumption that $\sigma_{q_i}^2 = T_i$, i.e., the variance of the quantizer is set equal to the masking threshold. Thus, the number of bits allocated to X_i can be found by computing the number of quantization levels using the expression

$$N_i(n) = \left\lfloor \frac{|X_i|}{\delta_i} + \frac{1}{2} \right\rfloor \tag{5.2}$$

previously developed in Chapter 2. We have introduced the concept of perceptual entropy (PE) as a good estimation of the lower bound for transparent coding audio signals. We have computed the PE over 1 minute of several wide-band speech audio sources using the model proposed by [42, 43]. The result is shown in Figure 5.3, a lower limit around 2.1 bits/sample is suggested by the PE to achieve transparent audio coding. This means that if the input signal was originally quantized using 16 bits/sample PCM, the theoretical upper bound for the compression ratio is 86%.

A histogram presenting the number of bits required by the transform coefficients, computed over 1 minute of several wide-band speech audio sources, is shown in Figure

5.4. It can be seen that 100% of the coefficients require 9-bits or less to be encoded. Moreover, 99.4% of the coefficients require a maximum of 6-bits. This last value is highlighted by a vertical line in the figure. It is important to point out that approximately 67% of the coefficients require a maximum of 1-bit to be encoded. In order to compress the signal even further, we decided to clip the number of levels to 7-bits. This value was determined by performing informal listening tests, a lower clipping value would produce perceptible disturbances in some of the reconstructed audio signals (such as the bagpipe or harpsichord).

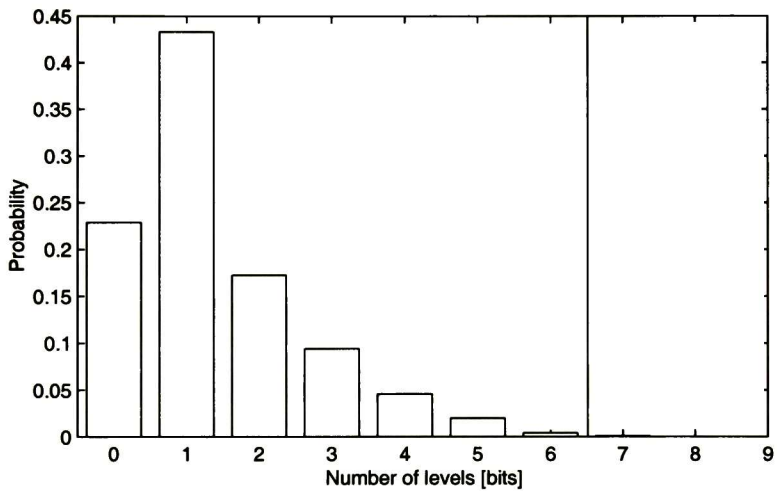


Figure 5.4: *Histogram presenting the number of bits required to encode several wide-band speech audio sources. The upper bound, represented by the vertical line, indicates that 99.4% of the coefficients can be encoded using a maximum of 6-bits. Additionally, 92.3% of the coefficients can be encoded using a maximum of 3-bits.*

Since the number of levels of the quantized coefficients is non-uniformly distributed, as shown in Figure 5.4, the use of a lossless entropy coding method should allow for a reduction on the average bit rate. We decided to use a Huffman encoder as proposed in [29, 41, 70]. This coder is represented by the H block in Figure 5.1. The Huffman table was estimated by passsing several different audio sources through the proposed encoder.

Band	Number of coefficients	Coefficients δ_i	Number of bits used to quantize
0	4	1–4	8
1	4	5–8	8
2	4	9–12	8
3	4	13–16	8
4	4	17–20	8
5	4	21–24	8
6	4	25–28	8
7	4	29–32	8
8	8	33–40	9
9	8	41–48	9
10	8	49-56	9
11	8	57–64	9
12	8	65–72	9
13	8	73–80	9
14	16	81–96	10
15	16	97–112	10
16	16	113–128	10
17	32	129–160	12
18	32	161–192	12
19	32	193–224	12
20	32	225–256	12

Table 5.1: *Vector quantization arrangement of the masking threshold coefficients* δ_i.

5.1.2 Masking Threshold Quantization

The transmitted/stored masking threshold or quantization step is obtained by vector quantizing the data provided by the masking model (T_z). As mentioned, there are 21 critical bands within the frequency range 50 Hz–7000 Hz, which are represented by 256 coefficients (X_i) in the WPT domain. Due to the strong relationship between the masking threshold coefficients (δ_i) within each critical band, we decided to vector quantize them as single block, the clustering arrangement can be seen in Table 5.1. Prior to quantization, the vectors are converted to dB and normalized by the maximum masking threshold value achievable (90.3 dB). The number of bits assigned to each vector were determined by performing subjective tests.

5.2 Description of the Decoder

The decoder structure, presented in Figure 5.5, is much simpler than the encoder's. The received data frame is split into two parts: the WPT coefficients and the masking

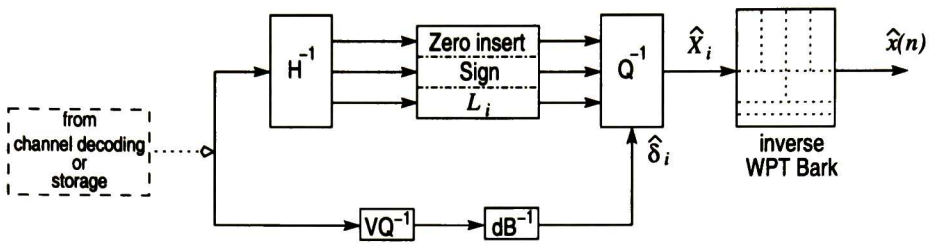


Figure 5.5: *Block diagram of the decoder.*

threshold coefficients. The Huffman table used to encode the quantized WPT coefficients is available at the decoder, so the quantized coefficients are easily retrieved. Additionally, the inverse quantization of the masking threshold coefficients is a simple lookup operation in the table generated by the VQ. Then, the WPT coefficients are inverse quantized using the masking threshold information. Finally, the coefficients are fed into a non-uniform tree, identical to the one used in the encoder, and the inverse WPT is performed to obtain the decoded signal $\hat{x}(n)$.

5.3 Coder Results

The proposed coder achieved an average bit rate of 56.1 kbit/s at a near transparent audio quality. Subjective informal listening tests, with several subjects, showed that for most audio sources the coded signal was indistinguishable from the original and for certain sources small differences were perceived. Due to the results provided in [13, 56] we assigned a MOS value of 3.3 to the source signal sampled at 16 kHz and quantized using 16-bit linear PCM. Thus, due to the almost indistinguishable differences produced by the developed coder we suggest a MOS value slightly above 3.0. This suggests that our coder has a performance close to that of the standard G.722 [35]. Table 5.2 provides the bit rate for ten different audio sources encoded using the proposed system. The objective and subjective quality assessment of our coder is fully addressed in the next chapter.

Bit rate could be further reduced by adjusting the masking threshold, but this occurs at the expense of a poorer reproduced audio quality. Informal tests show that coded signals at around 35 kbit/s were perfectly identifiable, but some annoying disturbances were perceived.

Instrument/Style	Code	Encoded bit rate
Female Voice (German)	Female	57.1 kbit/s
Male voice (English)	Male	50.4 kbit/s
Solo violin	Violin	63.6 kbit/s
Solo bass guitar	Bass	49.4 kbit/s
Solo trumpet	Trumpet	57.8 kbit/s
Solo clarinet	Clarinet	54.6 kbit/s
Accordion and other instruments	Accordion	58.9 kbit/s
Solo tambourine	Tambourine	53.2 kbit/s
Solo triangle	Triangle	58.6 kbit/s
Highlight from Carmen	Carmen	57.2 kbit/s

Table 5.2: *List of source material and their coded bit rate. The source material has been band limited to the range 50–7,000 Hz, sampled at 16 kHz and quantized using 16-bit linear PCM.*

Chapter 6

Coder Performance Evaluation

In this chapter we present the performance evaluation for the proposed coding system. Objective measures are applied to the encoded audio signal. In Addition, subjective measures (i.e., listening tests) are used to evaluate the coding performance of the system.

This chapter is divided into three sections. In the first part, perceptual objective measures are addressed as a mean of quantifying the similarity between two patterns. Secondly, subjective measurements are introduced as a mean to detect differences between stimuli presented to human subjects. Finally, the proposed system evaluation results are presented.

6.1 Perceptual Objective Measures

The quality of an audio coder can be determined either objectively or subjectively. Traditionally, subjective tests are difficult to reproduce, expensive and time consuming. On the other side, with objective methods there are in general problems with the relationship between the measurements and the perceived quality [5, 10]. To be useful, an objective audio distortion measure should be

- Acoustically significant. Small and large values should correspond to good and bad subjective quality, respectively.
- Mathematically explicit. It can be analyzed and implemented.
- Defined in a well chosen parametric space. It can be adapted to the system under analysis.

The distortion measure, also known as distance measure, can be computed either over time frames or in the frequency domain. The distance between the second order properties of two processes is called *spectral distance measure.* Spectral distortion measures have been introduced with the intention of simulating the frequency analysis that the HAS naturally performs. In addition, they also offer the advantage of being little sensitive to the phase differences and waveform misalignments, which usually are not perceptually relevant during signal comparison.

Let $x(n)$ be a WSS process and $r(n)$ its autocorrelation function. Then, the power spectral density (PSD) $S_x(e^{i\omega})$ of $x(n)$ is given by the expression

$$S_x(e^{i\omega}) = \sum_{k=-\infty}^{+\infty} r(k)e^{-ik\omega} \tag{6.1}$$

Usually, in the case of audio signals $x(n)$ can be modeled as an AR process, described by the p-th order model $A_p(e^{i\omega})$ with system gain σ_p^2. Then its PSD can be written as

$$S_x(e^{i\omega}) \approx \frac{\sigma_p^2}{|A_p(e^{i\omega})|^2} \tag{6.2}$$

This approximation becomes an equality for an infinite number of poles ($p \to +\infty$). However, finit order models give already an accurate representation of the perceptually relevant spectral peaks.

As we have seen, when processing non-stationary stochastic signals such as audio, most of the times they are decomposed into frames. In this case, the short-time autocorrelation of an N-sample frame $x(n)$ for $n = 0, 1, \cdots, N-1$ is given by the expression

$$r_N(i) = \sum_{k=0}^{N-i-1} x(k)x(k+|i|) \tag{6.3}$$

for $i = 0, \cdots, N-1$. Then, we calculate the short-time power spectral density $S_x(e^{i\omega})$ of such sequence. The frame length should correspond to a temporal duration where the signal can be considered quasi-stationary. As mentioned in Chapter 2, for audio signals this duration is in the range from 2 to 50 msec.

Two of the most common objective measures for perceptual coders are the segmental signal-to-noise ratio (SNR_{SEG}) and the Itakura-Saito distortion (d_{IS}) [5, 13, 71] In the next subsections we provide a brief explanation of these measurements.

6.1.1 Segmental Signal-to-Noise Ratio

This measure is the well known ratio of the signal variance to the noise variance for each signal frame. The goal of this measure is to assign equal weights to loud and faint segments of nonstationary signals. The SNR_{SEG} can be computed using the expression

$$SNR_{SEG} = \frac{1}{M}\sum_{i=0}^{M-1} 10\log_{10}\frac{\sum_{n=0}^{N-1} x^2(n+Ni)}{\sum_{n=0}^{N-1}[x(n+Ni)-\hat{x}(n+Ni)]^2} \quad \text{[dB]} \tag{6.4}$$

where M stands for the number of processed frames along the signal and N is the number of samples per frame. In some cases, problems can arise if low energy (i.e., silence) frames are included, since large negative SNR values bias the overall measure. A threshold can be used to exclude any frames that contain unusually low SNR values.

The SNR_{SEG} measure has been used by [13, 71] for the objective evaluation of a wavelet-based perceptual audio/speech coder. In [13] values above 20 dB were measured for speech signals and in [71] values between 21 and 28 dB were measured for hi-fi audio. Both coding systems claim to achieve transparent audio quality.

6.1.2 Itakura-Saito Distortion

The HAS is little sensitive to phase distortion in acoustic signals, thus, most of the audio coding systems focus on the spectral magnitude. A consequence of this fact is that the coder output waveform can be very different from the original signal, nonetheless no difference is perceived by the listener. Therefore, measures like the SNR_{SEG}, based on the sample difference between two waveforms, do not provide significant information on the performance of the coder if the signals are off-phase or misaligned.

One of the ways of overcoming this restriction is by means of the Itakura-Saito algorithm [38, 67], one of most succesful distortion measures. Also known as a likelihood ratio distance measure, the Itakura-Saito distortion measure performs a comparison between the spectral envelopes of signals and is more influenced by a mismatch in formant location than in spectral valleys. It is mostly used in the context of LPC, but its generalization can be expressed as

$$d_{IS_{total}} = \frac{1}{M}\sum_{k=0}^{M-1}\left[\frac{1}{N}\sum_{j=0}^{N-1}\left(\frac{S_{x,k}(j)}{S_{\hat{x},k}(j)} - \log_e\frac{S_{x,k}(j)}{S_{\hat{x},k}(j)} - 1\right)\right] \tag{6.5}$$

For a signal frame, this measure can also be expressed as the polynomial [3]

$$d_{IS} = \left(\frac{G}{\hat{G}}\right)^2 \frac{\hat{\mathbf{a}}^\mathsf{T}\mathbf{R}\hat{\mathbf{a}}}{\mathbf{a}^\mathsf{T}\mathbf{R}\mathbf{a}} - 2\log_e\left(\frac{G}{\hat{G}}\right) - 1 \tag{6.6}$$

where $\hat{\mathbf{a}} = [1,\ \hat{a}_1,\ \hat{a}_2, \ldots,\ \hat{a}_p]^\mathsf{T}$ $\mathbf{a} = [1,\ a_1,\ a_2, \ldots,\ a_p]^\mathsf{T}$ are the linear prediction coefficients of the coded and source signal respectively and $\mathbf{R}$ is the autocorrelation matrix of $x(n)$. When the gains are assumed to be equal, as in our case, the Itakura-Saito measure is simply

$$d_{IS} = \frac{\hat{\mathbf{a}}^\mathsf{T}\mathbf{R}\hat{\mathbf{a}}}{\mathbf{a}^\mathsf{T}\mathbf{R}\mathbf{a}} - 1 \tag{6.7}$$

The Itakura-Saito measure is not symmetric, for a symmetric representation the reader is referred to [3].

6.2 Subjective Measures

The design of reliable methods for the study of sensory process is covered by the field of *psychoacoustics.* It generates a description of the HAS by designing acoustical experiments containing stimuli that human subjects have to detect. Psychoacoustical tests can be mainly divided into two groups [10]: *explicit* and *implicit.* In the first case, one attempts to measure quality of sound sequences by a subjective judgment, traditionally, the subject scores or chooses the sequence from a proposed score/stimuli set. This is the method employed in the thesis. In the second case, one estimates parameters of auditory system models. This second method is out of the scope of our work.

6.2.1 Mean Opinion Score

The Mean Opinion Score (MOS) is a popular assessment method largely used to measure the quality of speech/audio processing systems [41, 58]. The 5-point scale generally used for algorithm assessment is represented in Table 6.1. The advantage of MOS values is that different impairment factors can be assessed simultaneously and that even small impairments can be graded. On the negative side, MOS procedures often require trained subjects and appropriate facilities. Additionally, experience has shown that MOS values can vary with time and from listener panel to listener panel, it is very difficult to duplicate test results at a different test site. Though this subjective

MOS	Quality Scale	Impairment Scale
5	Excellent	Imperceptible
4	Good	Perceptible, but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Unsatisfactory	Very annoying

Table 6.1: *Five-point MOS scale.*

measure has the advantage of being simple, from an implementation stand point, the results are, however, not easy to compare.

6.3 Evaluation of the Proposed System

The objective evaluation of the proposed coding system is based on the segmental signal-to-noise ratio and the Itakura-Saito spectral distortion. As already mentioned, these assessments have been mainly chosen for their acceptable correlation with subjective quality measures. The results for the proposed coder are presented in Table 6.2. The SNR_{SEG} values obtained are within the range of those given in [71] and above those presented by [13]. In addition, the average SNR_{SEG} of the proposed system is 25.6 dB, slightly below the 27.4 dB of the G.722 [1]

The subjective listening tests were performed as in [71]. This test involved a group of 20 subjects that were requested to volunteer for subjective evaluation of the proposed coding system. The group consisted in eighteen males and two females. Most of the individuals in this group came from the Telecommunications and Control research groups at the *Cinvestav-IPN Unidad Guadalajara.* Their ages ranged from 22 to 34 years. Some of them had previous experience in subjective evaluation of signal compression schemes. The subjective test was carried as follows. The listeners were presented with a total of 10 audio pairs. Each pair was composed of a single source material and was formed of 7 seconds of audio stimulus, 5 seconds of silence, and finally 7 seconds of audio stimulus. They were then asked to identify the stimulus which they found to be better in overall quality for each audio pair. A "*not sure*" response was permitted.

The responses of the listeners were averaged for each audio source. In Table 6.3 are summarized the results of the transparency tests. Particularly, column 2 shows the probability that the original music sample was preferred over the encoded version

[1] This value has been reprinted from [13].

Code	SNR_{SEG} [dB]	d_{IS}
Female	28.5	0.22
Male	29.4	0.21
Violin	27.9	0.22
Bass	25.1	0.28
Trumpet	26.1	0.27
Clarinet	22.3	0.39
Accordion	26.3	0.25
Tambourine	23.9	0.34
Triangle	19.5	0.42
Carmen	27.4	0.22

Table 6.2: *Objective measures for the proposed coding system.*

Code	Average probability of original signal preferred over the proposed encoder signal	Comments
Female	0.46	Transparent
Male	0.61	Nearly transparent
Violin	0.46	Transparent
Bass	0.71	Original preferred
Trumpet	0.53	Transparent
Clarinet	0.63	Nearly transparent
Accordion	0.63	Nearly transparent
Tambourine	0.57	Transparent
Triangle	0.43	Transparent
Carmen	0.50	Transparent

Table 6.3: *Subjective listening test result: transparency test.*

using the proposed system. Coder quality was considered to be transparent if the probability in the particular row of Table 6.3 is close to 0.5 (i.e., equal probability of choosing the original signal or the encoded). It is necessary to point out that the trial size for these tests is relatively small (because of the number of evaluated subjects and audio stimuli presented), therefore, the quantified average probabilities have only limited confidence levels. Nevertheless, these ciphers demonstrate that the proposed coder provided a transparent or nearly transparent quality for all audio coding sources but for one. The quality of the bass guitar signal encoded with the proposed coder was found low compared to the other audio pieces. The bass guitar source signal contains segments of silence before an abrupt transient, and the disturbances presented in the coded signal seem like pre-echo problems in the low frequencies. The proposed coder

needs to be further optimized for such signals.

Chapter 7

Conclusions

This chapter is divided into two parts. In the first one, we summarize the main results and achievements of the work discussed in this dissertation. In the second part, we provide some possible directions for future research and improvements for the proposed coding system.

7.1 Conclusions

During the present work, we have formulated a time-frequency auditory decomposition based on the wavelet packet transform, and a masking model based on the MPEG Psycoachoustic Model II. The merger of both systems contributed to the development of a conceptually simple and efficient coding system.

With the proposed coding system, we were able to encode wide-band speech audio signals at an average bit rate of 56.1 kbit/s. The overall quality of the encoded signal was found subjectively transparent or nearly transparent. Throughout the development of this work, it appeared clearly that auditory modeling is a key issue for the improvement of current audio processing systems. Thus, the use of signal analysis tools that mimic the resolution capabilities of the ear allow for a better definition and implementation of masking models.

Though a wavelet packet transform tree approximating the human auditory system and the adaptation of the MPEG pyscoacoustic model are not novel, their combination for wide-band speech audio coding is. Additionally, by the time this thesis was being written, no information regarding vector quantization techniques for the MPEG psychoacoustic model were known.

The wavelet packet transform can be seen as a natural choice for the approximation of time frequency mappings aproaching those performed by the human auditory system. However, it possesses a major drawback associated with its limited temporal and spectral localizations. That is, the basis functions or filters, do not possess a sufficiently fast decay in time and present important secondary lobes in frequency that spread far apart from the main lobe. Furthermore, such an overlapping contribution is not identical for all the filters. This situation can be easily seen in Figure 3.22. This phenomenon is mainly caused by the fact that the filters are only optimized at the first stage, and then simply iterated to build the desired tree structure.

Ideally, time and frequency overlapping should only be introduced from a perceptual point of view and resulting masking patterns. If some amount of overlapping is already created by the employed time-frequency mapping, it is indispensable to know by which amount a given band affects each of its neighbors when quantization noise is introduced. In this way, all the spurious contributions could be partially or totally removed from the affected subbands.

Though there is a large amount of statistical and perceptual redundancy removed by the coder, the bit rate achieved is still considerably high. Consecutive samples in one channel of the transform always posses some amount of correlation. This could be removed by using prediction techniques. During the steady parts of the audio signal, the masking threshold changes relatively slow, thus, techniques such as *matrix quantization* could be employed by grouping together a sequence of successive frame vectors and encode them as a single matrix.

A major drawback of the proposed coding system is its computational burden. The system has to perform a wavelet packet decomposition in parallel with an FFT, these operations require a great amount of computational power. An efficient implementation of wavelet packet decomposition could be employed.

7.2 Possible Extensions and Future Research

In what follows we give several issues that should be addressed to continue this research. Several improvements in the proposed method are possible, both in terms of reducing its computational complexity and its bit rate requirements.

So far, in the proposed coder we have not taken into consideration the auditory system property of temporal masking, thus we should consider incorporating it. Several

authors have explored this technique claiming promising results. In addition, adapting the MPEG psychoacoustic model to use the wavelet transform instead of the FFT would considerably reduce the computational burden, because only one transform would be required. Upgrading the wavelet packet transform to a fast wavelet packet transform using a lattice implementation structure would reduce ever further the computational requirements. An optimization of the wavelet basis (under the auditory system constraint) to match the audio data clearly results in a significant reduction in the bit rate requirement.

The side information currently amounts for about 35% of the bit rate requirements, thus, some methods to reduce it could be employed.

Time-frequency mappings with higher frequency and time selectivity than that achieved by the present transform could be used. The following filter/transform cases could be considered:

- Orthonormal filters optimized at each stage of the tree structure.
- Mixed nonuniform/uniform decompositions.
- Eliminate the filter orthonormality constraint, which restricts frequency selectively.

Several methodologies have been proposed to reduce the pre-echoes that tend to affect block-based coding schemes. We should consider incorporating the *window switching* technique to the proposed coder. This consists in changing the analysis block length from long duration (e.g., 15 to 20 msec) during stationary segments to short duration (e.g., 4 msec) when transients are detected.

The superposition of threshold components used in the psychoacoustic model proposed by MPEG is based on linear addition. This results in a conservative masking threshold estimation. Novel non-linear superposition techniques that better approximate the real threshold have been proposed, thus we could incorporate them to the proposed coder.

Because the proposed coder extracts a great amount of redundancy, it is likely to be sensitive to channel errors. For some applications, where hardwired lines or digital storage media are used, the sensitivity to errors is not likely to be a problem. In channels where the probability of error is relatively large (e.g., wireless), some protection of the information must be done in order to prevent errors. This last situation suggests

the future use of techniques such as *joint source-channel coding* and *multiresolution transmission* in the proposed coding system. So far, most of the effort in joint source-channel coding in the audio domain has been devoted to toll quality. Thus, it would be of interest to develop a joint source-channel coding scheme for wide-band speech audio.

Appendix A

High Resolution Hypothesis

In this appendix we analyze the high resolution hypothesis of a scalar quantizer. If the number of quantization levels, L, is large and we know the PDF, $p_s(x)$, of the signal being quantized. Then, we can obtain an expression for the optimal partition and an expression for the quantization noise power, only in terms of $p_s(x)$. The high resolution hypothesis states that we can suppose the PDF constant inside the interval $[t^{i-1}, t^i]$ and its representative value can be taken from the middle of the interval.

$$p_s(x) \approx p_s(\hat{s}^i) \quad \text{for} \quad x \in [t^{i-1}, t^i] \tag{A.1}$$

$$\hat{s}^i \approx \frac{[t^{i-1} + t^i]}{2} \tag{A.2}$$

We define

$$\Delta(i) = t^i - t^{i-1} \tag{A.3}$$

the length of the interval is $[t^{i-1}, t^i]$ and the probality that $s(n)$ belongs to this interval is given by

$$P_r(i) = P_r\{S \in [t^{i-1}, t^i]\} = p_s(\hat{s}^i)\Delta(i) \tag{A.4}$$

The variance of the quantization noise is given by the expression

$$\sigma_q^2 = \sum_{i=1}^{L} p_s(\hat{s}^i) \int_{t^{i-1}}^{t^i} (x - \hat{s}^i)^2 dx \tag{A.5}$$

we can write the integral in the right side of Equation A.5 as

$$\int_{t^{i-1}}^{t^i} (x - \hat{s}^i)^2 \, dx = \int_{-\Delta(i)/2}^{+\Delta(i)/2} x^2 \, dx = \frac{\Delta^3(i)}{12} \tag{A.6}$$

then, we obtain

$$\sigma_q^2 = \frac{1}{12} \sum_{i=1}^{L} p_s(\hat{s}^i) \Delta^3(i) \tag{A.7}$$

with the help of Equation A.4, we can write the last expression as

$$\sigma_q^2 = \sum_{i=1}^{L} P_r(i) \frac{\Delta^2(i)}{12} = \frac{1}{12} \mathrm{E}\{\Delta^2\} \tag{A.8}$$

This expression lets us know that the variance of the quantization noise *only depends* of the interval length $\Delta(i)$. We must find $\{\Delta(1), \cdots, \Delta(L)\}$ that let us minimize σ_q^2. We will write part of the Equation A.7 in a different way

$$\alpha^3(i) = p_s(i) \Delta^3(i) \tag{A.9}$$

Now we can write

$$\sum_{i=1}^{L} \alpha(i) = \sum_{i=1}^{L} \left(p_s(\hat{x}^i)\right)^{\frac{1}{3}} \Delta(i) \approx \int_{-\infty}^{+\infty} \left(p_s(x)\right)^{\frac{1}{3}} \, dx = \text{constant} \tag{A.10}$$

this integral does not depend anymore of $\Delta(i)$. Now, the problem is to minimize the sum of L positive numbers in such a way that their sum is a constant. The easiest way to overcome this problem is to make the L numbers identical

$$\alpha(1) = \quad \cdots \quad = \alpha(L)$$

this implies that

$$\alpha^3(1) = \cdots = \alpha^3(L)$$
$$p_s(\hat{s}^1) \Delta^3(1) = \cdots = p_s(\hat{s}^L) \Delta^3(L)$$

This means that an interval will be smaller than the probability that $s(n)$ belongs to such interval. All the intervals will have the same contribution in the quantization noise variance. For the next part we use a simplification, $L = 2^b - 1 \approx 2^b$, which is

valid due to the high resolution hypothesis. Now, we can write the quantization noise variance as

$$\sigma_q^2 = \frac{L}{12}\alpha^3 \tag{A.11}$$

where α stands for

$$\alpha = \int_{-\infty}^{+\infty} (p_s(x))^{\frac{1}{3}}\, dx \tag{A.12}$$

Finally, we obtain

$$\sigma_q^2 = \frac{1}{12}\left(\int_{-\infty}^{+\infty} (p_s(x))^{\frac{1}{3}}\, dx\right)^3 2^{-2b} \tag{A.13}$$

This is not a rigorous mathematical demonstration, but will be very helpful for our purposes. Now, let us suppose the signal $s(n)$ has a uniform PDF, $p_s(x) = \frac{1}{2A}$, $-A \leq x \leq A$, with mean $\mu_s = 0$ and variance $\sigma_s^2 = A^2/3$. In this case, using Equation A.13 we obtain

$$\sigma_q^2 = \frac{1}{12}\left(\int_{-A}^{+A} \left(\frac{1}{2A}\right)^{\frac{1}{3}} dx\right)^3 2^{-2b} \tag{A.14}$$

$$\sigma_q^2 = \frac{A^2}{3L^2} = \sigma_s^2 A^{-2b} \tag{A.15}$$

Now, let us suppose that we have a Gaussian source, centered ($\mu_s = 0$), with variance σ_s^2, for the which

$$p_s(x) = \frac{1}{\sqrt{2\pi\sigma_s^2}}\, e^{\frac{x^2}{2\sigma_s^2}} \tag{A.16}$$

then we have

$$\int_{-\infty}^{+\infty} (p_s(x))^{\frac{1}{3}}\, dx = \int_{-\infty}^{+\infty} \frac{1}{(2\pi\sigma_s^2)^{1/6}}\, e^{\frac{x^2}{6\sigma_s^2}}\, dx \tag{A.17}$$

$$\int_{-\infty}^{+\infty} (p_s(x))^{\frac{1}{3}}\, dx = \left(2\pi\sigma_s^2\right)^{\frac{1}{3}} \sqrt{3} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi 3\sigma_s^2}}\, e^{\frac{x^2}{6\sigma_s^2}}\, dx \tag{A.18}$$

$$\int_{-\infty}^{+\infty} (p_s(x))^{\frac{1}{3}}\, dx = \left(2\pi\sigma_s^2\right)^{\frac{1}{3}} \sqrt{3} \tag{A.19}$$

Now, we can conclude that

$$\sigma_q^2 = \frac{1}{12} 2\pi \sigma_s^2 3^{3/2} 2^{-2b} \tag{A.20}$$

$$\sigma_q^2 = h \sigma_s^2 2^{-2b} \tag{A.21}$$

where $h = \frac{\sqrt{3}}{2} \pi$. In the two previous examples, we have seen that the quantization noise variance is always proportional to the signal variance and the number of quantization levels,

$$\sigma_q^2 = \kappa \sigma_s^2 2^{-2b} \tag{A.22}$$

where the proportionality constant κ depends on the PDF of $s(n)$. As we have previously seen for a uniform PDF $\kappa = 1$ and for a Gaussian PDF $\kappa = \frac{\sqrt{3}}{2} \pi$. These two proportionality constants will be very useful for our purposes. The calculation of more proportionality constants is out of the scope of this thesis, but the reader is referred to [41, 29] for further information.

Appendix B

Bit Allocation

To solve the bit allocation problem of a transform coder, we assume that the transform coefficients have a Gaussian PDF, thus all the constants h_i will be identical and they will not suffer any modification during the linear transformation. The goal of optimal bit allocation is to minimize the distortion D in the expression

$$D = \frac{h}{M} \sum_{i=0}^{M-1} \sigma_{Y_i}^2 2^{-2b_i} \qquad \text{where } h = \frac{\sqrt{3}}{2}\pi \tag{B.1}$$

under the constraint

$$\sum_{i=0}^{M-1} b_i \leq bM. \tag{B.2}$$

We know the arithmetic/geometric mean inequality states that for any positive real numbers a_i, $i = 0,\ 1,\ \cdots,\ M-1$,

$$\frac{1}{M} \sum_{i=0}^{M-1} a_i \leq \left(\prod_{i=0}^{M-1} a_i \right)^{\frac{1}{M}} \tag{B.3}$$

if we make $a_i = \sigma_{Y_i}^2 2^{-2b_i}$, the last expression becomes

$$\frac{1}{M} \sum_{i=0}^{M-1} \sigma_{Y_i}^2 2^{-2b_i} \leq \left(\prod_{i=0}^{M-1} \sigma_{Y_i}^2 2^{-2b_i} \right)^{\frac{1}{M}} = \left(\prod_{i=0}^{M-1} \sigma_{Y_i}^2 \right) 2^{-2\sum_{i=0}^{M-1} \frac{b_i}{M}}$$

$$\frac{1}{M} \sum_{i=0}^{M-1} \sigma_{Y_i}^2 2^{-2b_i} \leq \alpha^2 2^{-2b} \tag{B.4}$$

where $\alpha^2 = \left(\prod_{i=0}^{M-1} \sigma_{Y_i}^2\right)^{\frac{1}{M}}$ The optimal value is reached when all the terms involved in the sum are equal, i.e., all the transform coefficients have the same variance, in such case the inequality becomes an equality. Thus, for any i we have

$$\sigma_{Y_i}^2 2^{-2b_i} = \alpha^2 2^{-2b} \tag{B.5}$$

Then,

$$\begin{aligned} 2^{b_i} &= 2^b \sqrt{\frac{\sigma_{Y_i}^2}{\alpha^2}} \\ b_i &= b + \frac{1}{2}\log_2\left(\frac{\sigma_{Y_i}^2}{\alpha^2}\right) \end{aligned} \tag{B.6}$$

Appendix C

Continous Wavelet Transform

Although the definition of the Continuous Wavelet Transform (CWT) was briefly introduced in Chapter 2, we repeat it here for completeness. A slightly different proof can also be found in [16, 51, 77, 82]. The CWT of a function $g \in \mathbf{L}^2(\mathbb{R})$ at time u and scale s is defined as

$$Wg(u,s) = \langle g, \psi_{u,s} \rangle = \int_{-\infty}^{+\infty} g(t) \frac{1}{\sqrt{|s|}} \psi^* \left(\frac{t-u}{s} \right) dt \tag{C.1}$$

Theorem. (CALDERÓN-GROSSMAN-MORLET). *Let $\psi \in \mathbf{L}^2(\mathbb{R})$ be a real function such that*

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

Any $g \in \mathbf{L}^2(\mathbb{R})$ satisfies

$$g(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Wg(u,s) \frac{1}{\sqrt{|s|}} \psi \left(\frac{t-u}{s} \right) du \frac{ds}{s^2} \tag{C.3}$$

PROOF. Let

$$G(\omega) = \int_{-\infty}^{+\infty} g(t) e^{-j\omega t} dt \tag{C.4}$$

and

$$\Psi(\omega) = \int_{-\infty}^{+\infty} \psi(t) e^{-j\omega t} dt \tag{C.5}$$

The Fourier Transform of $\psi_{u,s}(t) = \frac{1}{\sqrt{|s|}} \psi \left(\frac{t-u}{s} \right)$ is given by the expression

$$F \left\{ \psi \left(\frac{t-u}{s} \right) \right\} = \int_{-\infty}^{+\infty} \psi \left(\frac{t-u}{s} \right) e^{-j\omega t} dt \tag{C.6}$$

if we make $t' = t/s$

$$\begin{aligned} &= \int_{-\infty}^{+\infty} \psi\left(t' - \frac{u}{s}\right) e^{-j\omega s(t'-\frac{u}{s})} e^{-j\omega u}\, s\, dt' \\ &= s\Psi(s\omega)e^{-j\omega u} \end{aligned} \tag{C.7}$$

Another way of writing this expression is

$$\psi\left(\frac{t-u}{s}\right) = \frac{s}{2\pi}\int_{-\infty}^{+\infty} \Psi(s\omega)e^{-j\omega u}e^{j\omega t}d\omega. \tag{C.8}$$

Inserting the right side of Equation C.4 into -C.1 gives

$$Wg(u,s) = \frac{1}{\sqrt{|s|}}\int_{-\infty}^{+\infty}\left(\frac{1}{2\pi}\int_{-\infty}^{+\infty} G(\omega)e^{j\omega t}d\omega\right)\psi^*\left(\frac{t-u}{s}\right)dt \tag{C.9}$$

arranging the order of the integrals gives

$$Wg(u,s) = \frac{1}{2\pi\sqrt{|s|}}\int_{-\infty}^{+\infty} G(\omega)\int_{-\infty}^{+\infty}\psi^*\left(\frac{t-u}{s}\right)e^{j\omega t}dt\, d\omega \tag{C.10}$$

and using Equation C.7 in the last expression results in

$$Wg(u,s) = \frac{\sqrt{|s|}}{2\pi}\int_{-\infty}^{+\infty} G(\omega)\,\Psi^*(s\omega)e^{j\omega u}d\omega \tag{C.11}$$

Now, substituting Equations C.8, -C.11 into Equation C.3

$$g(t) = \frac{s}{4\pi^2 C_\psi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} G(\omega)\Psi^*(s\omega)e^{j\omega u}d\omega\int_{-\infty}^{+\infty}\Psi(s\Omega)e^{-j\Omega u}e^{j\Omega t}d\Omega\,\frac{ds}{s^2}du \tag{C.12}$$

and

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{ju(\omega-\Omega)}du = \delta(\omega-\Omega) \tag{C.13}$$

Therefore $\omega = \Omega$, otherwise the result would be zero. Arranging the order of the integrals we obtain

$$g(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} G(\omega)e^{j\Omega t}d\omega\cdot\frac{1}{C_\psi}\int_{-\infty}^{+\infty}\frac{1}{s}\Psi^*(s\omega)\Psi(s\omega)ds \tag{C.14}$$

The first part of the right side is easily identified as the inverse Fourier transform of $G(\omega)$. In the second part, if we make $\alpha = s\omega$

$$g(t) = g(t)\cdot\frac{1}{C_\psi}\int_{-\infty}^{+\infty}\frac{|\Psi(\alpha)|^2}{\alpha}d\alpha \tag{C.15}$$

Finally, using the condition imposed by Equation C.2 the proof is complete.

Appendix D

Noble Identities

Sample rate converters and and filters can be exchanged in a multirate system if they satisfy the properties depicted in Figure D.1, where $G(z)$ is a rational transfer function. These equivalences are known as *noble identities* [77, 80, 79].

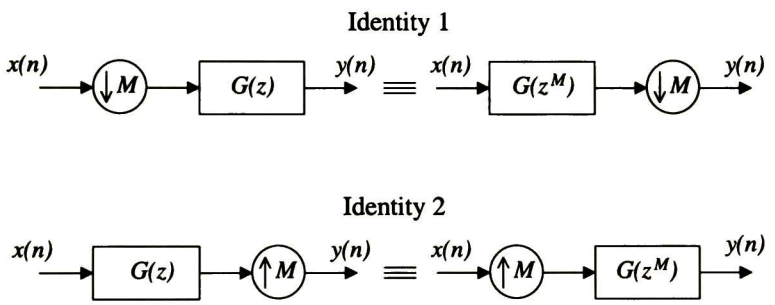


Figure D.1: *Noble identities in multirate systems, for subsamplers (top) and downsamplers (bottom).*

Appendix E

ISO/IEC MPEG Psychoacoustic Model

The MPEG psychoacoustic models are independent psychoacoustic algorithms that can be adjusted and adapted to different requirements. This appendix has been taken and adapted from [37] and presents the general Psychoacoustic Model II. This psychacoustic model has been modified to fit the particular necessities of the proposed coder.

The following are the necessary steps for calculation of the masking threshold used in the coder.

1. *Reconstruct 512 samples of the input signal.* The window size required to calculate the masking threshold is 512 samples. This window (s_i, where $1 \leq i \leq 512$) is formed by concatenating the 256 samples of the current data block to the 256 samples of previous data block. The first masking threshold is calculated by duplicating the first 256 samples block. This process can also be seen as a sliding window s_i of 512 samples with an overlapping of 50%.

2. *Calculate the complex spectrum of the input signal.* First, s_i is windowed by a 512 point Hann window, i.e., $sw_i = s_i \cdot \left(0.5 - 0.5\cos\left(\frac{2\pi(i-0.5)}{512}\right)\right)$.

   Second, a standard 1024-point FFT of sw_i is then calculated. sw_i is centered and then zero padded.

   Third, the polar representation of the transform is then calculated. r_ω and f_ω represent the magnitude and phase components of the transformed sw_i respec-

tively, where $1 \leq \omega \leq 513$. An index of 1 for ω corresponds to the DC term and 513 corresponds to Nyquist frequency (8 kHz.).

3. *Calculate a predicted r and f.* A predicted magnitude $\hat{r}_\omega$, and phase $\hat{f}_\omega$ are calculated from the preceeding two threshold calculation blocks r and f:

$$\hat{r}_\omega = 2r_\omega(t-1) - r_\omega(t-2) \tag{E.1}$$

$$\hat{f}_\omega = 2f_\omega(t-1) - f_\omega(t-2) \tag{E.2}$$

where t represents the current block number, $t-1$ indexes the previous block data, and $t-2$ indexes the data block before that.

4. *Calculate the unpredictability measure c_ω.* The unpredictability measure is defined as:

$$c_\omega = \frac{\left[\left(r_\omega \cos(f_\omega) - \hat{r}_\omega \cos(\hat{f}_\omega)\right)^2 + \left(r_\omega \sin(f_\omega) - \hat{r}_\omega \sin(\hat{f}_\omega)\right)^2\right]^{\frac{1}{2}}}{r_\omega + \text{abs}(\hat{r}_\omega)} \tag{E.3}$$

5. *Calculate the energy and unpredictability in the threshold calculation partitions.* The energy in each partition, e_b, is:

$$e_b = \sum_{\omega=\omega\text{low}_b}^{\omega\text{high}_b} r_\omega^2 \tag{E.4}$$

and the weighted unpredictability, c_b, is

$$e_b = \sum_{\omega=\omega\text{low}_b}^{\omega\text{high}_b} r_\omega^2 c_\omega \tag{E.5}$$

The threshold calculation partitions provide a resolution of approximately either one FFT line or $\frac{1}{3}$ critical band, which ever is wider. At low frequencies, a single line of the FFT will constitute a calculation partition. At high frequencies, many lines will be combined into one calculation partition. A set of partition values is provided in Table E.1 for a sampling rate of 16 kHz. The elements of this table will be used in the threshold calculation process. The elements in the table are:

- The index of the calculation partition, b.

- The lowest frequency line in the partition, ωlow_b.
- The highest frequency line in the partition ωhigh_b.
- The median Bark value of the partition, $bval_b$.
- A lower limit for the SNR in the partition that controls unmasking effects, $minval_b$.
- The value for tone masking noise (in dB) for the partition, TMN_b.

6. *Convolve the partitioned energy and unpredictability with the spreading function.*

$$ecb_b = \sum_{bb=1}^{bmax} e_{bb} \cdot \text{sprdngf}(bval_{bb}, bval_b) \tag{E.6}$$

$$ct_b = \sum_{bb=1}^{bmax} c_{bb} \cdot \text{sprdngf}(bval_{bb}, bval_b) \tag{E.7}$$

In this appendix several points are refered to spreading function. It is calculated by the following method:

$$tmpx = 1.05(j - i) \tag{E.8}$$

where i is the Bark value of the signal being spread, j is the Bark value of the band being spread into, and $tmpx$ is a temporary variable.

$$x = 8 \text{ minimum } \left((tmpx - 0.5)^2 - 2(tmpx - 0.5), 0\right) \tag{E.9}$$

where x is a temporary variable, and minimum (a, b) is a function returning the more negative of a or b.

$$tmpy = 15.811389 + 7.5(tmpx + 0.474) - 17.5\left(1.0 + (tmpx + 0.474)^2\right)^{\frac{1}{2}} \tag{E.10}$$

where $tmpy$ is another temporary variable.

$$\text{if } (tmpy < -100) \text{ then } \{\text{sprdngf(i,j)=0}\} \text{ else } \left\{\text{sprdngf}(i,j) = 10^{\frac{x+tmpy}{10}}\right\} \tag{E.11}$$

Because ct_b is weighted by the signal energy, it must be renormalized to cb_b.

$$cb_b = \frac{ct_b}{ecb_b} \tag{E.12}$$

At the same time, due to the non-normalized nature of the spreading function, ecb_b should be renormalized and the normalized energy en_b calculated

$$en_b = ecb_b \cdot rnorm_b \tag{E.13}$$

where, the normalization coefficient, $rnorm_b$, is:

$$rnorm_b = \frac{1}{\sum_{bb=0}^{bmax} \text{sprdngf}(bval_b b, bval_b)} \tag{E.14}$$

7. *Convert cb_b to tb_b, the tonality index.*

$$tb_b = -0.299 - 0.43 \log_e(cb_b) \tag{E.15}$$

Each tb_b is limited to the range of $0 < tb_b < 1$.

8. *Calculate the required SNR in each partition.* $NMT_b = 5.5$dB for all b. NMT_b is the value for noise masking tone (in dB) for the partition. The required signal to noise ratio, SNR_b, is:

$$SNR_b = \text{maximum } (minval_b,\ tb_b \cdot TMN_b + (1 - tb_b) \cdot NMT_b) \tag{E.16}$$

where maximum(a, b) is a function returning the least negative of a or b.

9. *Calculate the power ratio.* The power ratio, b_{cb}, is:

$$bc_b = 10^{\frac{-SNR_b}{10}} \tag{E.17}$$

10. *Calculation of the actual energy threshold nb_b.*

$$nb_b = en_b\, bc_b \tag{E.18}$$

11. *Spread the threshold energy over FFT lines, yielding nb_ω.*

$$nb_\omega = \frac{nb_b}{\omega\text{high}_b - \omega\text{low}_b + 1} \tag{E.19}$$

12. *Include absolute threshold.* This yields the final energy threshold of audibility, thr_ω, given the expression

$$thr_\omega = \text{maximum}(nb_\omega, absthr_\omega) \tag{E.20}$$

The dB values of *absthr* are calculated using Equation 4.2. This threshold must be fixed relative to the level of a sine wave of ± 1 lsb @ 4000 Hz in the FFT used for threshold calculation. The dB values must be converted into the energy domain after considering the FFT normalization actually used.

Index b	ω_{low} bins	ω_{high} bins	*bval* [Bark]	*minval* [dB]	TMN [dB]
1	1	1	0	0	24.5
2	2	5	0.4631	0	24.5
3	6	9	1.0789	20	24.5
4	10	13	1.6904	20	24.5
5	14	17	2.2950	20	24.5
6	18	21	2.8903	20	24.5
7	22	25	3.4741	20	24.5
8	26	29	4.0446	20	24.5
9	30	33	4.6001	20	24.5
10	34	37	5.1394	20	24.5
11	38	41	5.6614	20	24.5
12	42	45	6.1655	17	24.5
13	46	49	6.6513	17	24.5
14	50	53	7.1186	15	24.5
15	54	58	7.6221	15	24.5
16	59	63	8.1545	13	24.5
17	64	68	8.6590	10	24.5
18	69	73	9.1366	7	24.5
19	74	78	9.5884	7	24.5
20	79	84	10.0572	4.4	24.5572
21	85	90	10.5369	4.4	25.0369
22	91	96	10.9858	4.4	25.4858
23	97	102	11.4061	4.5	25.9061
24	103	109	11.8319	4.5	26.3319
25	110	116	12.2591	4.5	26.7591
26	117	124	12.6841	4.5	27.1841
27	125	132	13.1041	4.5	27.6041
28	133	141	13.5167	4.5	28.0167
29	142	150	13.9205	4.5	28.4205
30	151	160	14.3142	4.5	28.8142
31	161	171	14.7156	4.5	29.2156
32	172	182	15.1035	4.5	29.6035
33	183	194	15.4784	4.5	29.9784
34	195	207	15.8558	4.5	30.3558
35	208	221	16.2334	4.5	30.7334
36	222	236	16.6101	4.5	31.1101
37	237	252	16.9852	4.5	31.4852
38	253	269	17.3583	4.5	31.8583
39	270	287	17.7294	4.5	32.2294
40	288	306	18.0987	4.5	32.5987
41	307	326	18.4661	4.5	32.9661
42	327	347	18.8313	4.5	33.3313
43	348	369	19.1938	4.5	33.6938
44	370	392	19.5525	4.5	34.0525
45	393	416	19.9061	4.5	34.4061
46	417	442	20.2596	4.5	34.7596
47	443	470	20.6160	4.5	35.1160
48	471	500	20.9706	4.5	35.4706
49	501	513	21.2152	4.5	35.7152

Table E.1: *Calculation partition table. This table is valid only at a sampling rate of 16 kHz.*

Bibliography

[1] Johnson I. AGBINYA. "Discrete wavelet transform techniques in speech processing" *IEEE Proceedings TENCON'96*, pages 514–519, 1996.

[2] Joseph ARROWOOD, Tami RANDOLPH, and Mark SMITH. "Filter Bank Design. In Vijay MADISETI and William DOUGLAS, editors, *The Digital Signal processing Handbook*, chapter 36, pages 36.1–36.17. CRC Press, Boca Raton, Florida, 1998.

[3] Nadim BATRI. Robust Spectral Parameter Coding in Speech Processing. Master's thesis, Department of Electrical Engineering, McGill University, Montreal, Canada, May 1998.

[4] Frank BAUMGARTE, Charalampos FEREKIDIS, and Hendrik FUCHS. "A Nonlinear Psychoacoustic Model Applied to the ISO MPEG Layer 3 Coder". Technical report, University of Hannover, Germany, 1998.

[5] John G. BEERENDS and Jan A. STEMERDINK. A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation. *Journal Audio Engineering Society*, 40(12):963–978, December 1992.

[6] Mark BLACK and Mehmet ZEYTINOĞLU. "Computationally Efficient Wavelet Packet Coding of Wide-Band Stereo Audio Signals". In *IEEE Proceedings ICASSP'95*, pages 3075 – 3078, Detroit, Mi., 1995.

[7] Simon BOLAND and Mohamed DERICHE. "New results in low bit rate audio coding using a combined harmoni-wavelet representation". In *Proceedings of the IEEE ICASSP'97*, pages 351–354, 1997.

[8] K. BRANDENBURG, J. HERRÉ, and J.D. JOHNSTON. "ASPEC: Adaptive Spectral Perceptual Entropy Coding of High Quality Music Signals", preprint 3011 of the 90th. AES Convention 1991. March, 1991.

[9] Karlheinz BRANDENBURG. "MP3 and AAC explained", Audio Engineering Society (AES) 17^{th} Conference on High Quality Audio Coding 1991.

[10] Benito CARNERO. *Perceptual coding and enhancement of wide-band speech.* PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, 1997.

[11] Benito CARNERO and Andrzej DRYGAJLO. "Perceptual coding of speech using a fast wavelet packet transform algorithm". In *Proceedings EUSIPCO'96*, pages 1661 – 1664, Trieste, Italy, 1996.

[12] Benito CARNERO and Andrzej DRYGAJLO. "Perceptual speech coding using time and frequency masking constraints" In *IEEE Proceedings ICASSP'97*, pages 1363 – 1367, 1997.

[13] Benito CARNERO and Andrzej DRYGAJLO. "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms" *IEEE Transactions On Signal Processing*, 47(6):1622 – 1635, June 1999.

[14] Wai-Yip CHAN and Allen GERSHO. "High Fidelity Audio Transform Coding with Vector Quantization" In *Proceedings IEEE ICASSP'90*, pages 1109–1112, 1990.

[15] Wai-Yip CHAN and Allen GERSHO. "Constrained-storage vector quantization in high fidelity audio transform coding". In *Proceedings IEEE ICASSP'91*, pages 3597–3600, 1991.

[16] Y. T. CHAN. *Wavelet Basics.* Kluwer Academic Publishers, Norwell, Massachusetts, 1995.

[17] Albert COHEN and Jelena KOVAČEVIĆ. "Wavelets: The Mathematical Background" *Proceedings of the IEEE*, 84(4):514–522, April 1996.

[18] Ronald COIFMAN, Yves MEYER, Steven QUAKE, and Victor M. WICKERHAUSER. "Signal Processing and Compression with Wave Packets" Numerical Algorithms Research Group, Departament of Mathematics, Yale University. New Haven, Connecticut.

[19] Ronald COIFMAN, Yves MEYER, Steven QUAKE, and Victor M. WICKERHAUSER. "Acoustic Signal Compression with Wave Packets". In *Wavelet Workshop*, Marseille, France, October 1992.

[20] A. CROISIER, D. ESTEBAN, and C. GALAND. "Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques" In *Int. Conf. on Inform. Sciences and Systems*, pages 443–446, Patras, Greece, August 1976.

[21] Ingrid DAUBECHIES. "Orthonormal bases of compactly supported wavelets" *Communications on Pure and Applied Mathematics*, 41:909–996, November 1988.

[22] Ingrid DAUBECHIES. "The Wavelet Transform, Time-Frequency Localization and Signal Analysis " *IEEE Transactions on Information Theory*, 36(5):961–1005, September 1990.

[23] Ingrid DAUBECHIES. "Where do wavelets come from?– A personal point of view" *Proceedings of the IEEE*, 84(4):510–513, April 1996.

[24] Y.F. DEHERY, M. LEVER, and P. URCUM. "A MUSICAM source codec for digital audio broadcasting and storage" In *Proceedings IEEE ICASSP'91*, pages 3605–3609, San Francisco, CA., 1991.

[25] S. DEUTSCH and A. DEUTSCH. *Understanding the Nervous System: An Engineering Perspective.* IEEE Press, Piscataway, N.J., 1993.

[26] W. A. DEUTSCH, A. NOLL, and G. ECKEL. "The perception of audio signals reduced by overmasking of the most prominent spectral amplitudes (peaks)" *preprint 3331 of the 92nd. AES Convention*, 1992.

[27] Hervé DIA, Gang FENG, and Yannick MAHIEUX. "Codage par transformée de la parole à bande élargie (0 à 7 khz)" *Quatorzième Colloque Gretsi*, pages 471 – 474, September 1993.

[28] Allen GERSHO. "Advances in speech and audio compression". *Proceedings of the IEEE*, 82(6):900 – 918, June 1994.

[29] Allen GERSHO and Robert M. GRAY. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, 1992.

[30] Randy GOLDBERG and Lance RIEK. *A practical handbook of speech coders.* CRC Press, Boca Raton, Florida, 2000.

[31] Joseph HALL. "Auditory Psychoacoustics for Coding Applications" In Vijay MADISETI and William DOUGLAS, editors, *The digital signal processing handbook*, chapter 36, pages 39.1–39.25. CRC Press, Boca Raton, Florida, 1998.

[32] Khaled N. HAMDY, Ali MURTAZA, and Ahmed H. TEWFIK. "Low bit rate high quality audio coding with combined harmonic and wavelet representations In *IEEE Proceedings ICASSP'96*, pages 1045 – 1048, 1996.

[33] Cormac HERLEY. "Wavelets and Filter Banks" In Vijay MADISETI and William DOUGLAS, editors, *The digital signal processing handbook*, chapter 36, pages 35.1–35.14. CRC Press, Boca Raton, Florida, 1998.

[34] Nikolaj HESS-NIELSEN and Victor M. WICKERHAUSER. "Wavelets and Time Frequeny Analysis" *Proceedings of the IEEE*, 84(4):523–540, April 1996.

[35] International Telecommunication Union (ITU). *7 kHz Audio-Coding within 64 kbit/s. ITU-T Reommendation G.722.* 1988.

[36] International Telecommunication Union (ITU). *Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP). ITU-T Recommendation G.729.* 1996.

[37] International Standard ISO/IEC IS11172-3. *Information technology Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s- Part 3: Audio.* Switzerland, 1993.

[38] F. ITAKURA and S. SAITO. "Analysis-synthesis telephone based on the maximum likelihood method" In *Proceedings of the 6th. International Congress on Acoustics*, pages C17–C20, Japan, 1968.

[39] Nikil JAYANT, James D. JOHNSTON, and Robert SAFRANEK. "Signal compression based on models of human perception" *Proceedings of the IEEE*, 81(10):1385 – 1421, October 1993.

[40] Nikil S. JAYANT. "High quality coding of telephone speech and wideband audio" In Sadaoki FURUI and M. Mohan SONDHI, editors, *Advances in Speech Signal Processing*, chapter 3, pages 85–108. Marcel Dekker, New York, NY, 1992.

[41] Nikil S. JAYANT and Peter NOLL. *Digital Coding of Waveforms: Principles and Applications to Speech and Video.* Prentice Hall, Englewood Cliffs, NJ, 1984.

[42] James D. JOHNSTON. "Estimation of perceptual entropy using noise masking criteria". *IEEE Proceedings ICASSP'88*, pages 2524 – 2527, 1988.

[43] James D. JOHNSTON. "Transform coding of audio signals using perceptual noise criteria" *IEEE Journal On Selected Areas In Communications*, 6(2):314 – 323, February 1988.

[44] James D. JOHNSTON and Karlheinz BRANDENBURG. "Wideband coding: Perceptual considerations for speech and music" In Sadaoki FURUI and M. Mohan SONDHI, editors, *Advances in Speech Signal Processing*, chapter 4, pages 109–140. Marcel Dekker, New York, NY, 1992.

[45] R. KASTANTIN, D. ŞTEFANOIU, G. FENG, N. MARTIN, and M. MRAYATI. "Optimal wavelets for high quality speech coding". *Proceedings EUSIPCO'94*, pages 399 – 402, 1994.

[46] H. P. KRAMER and M.V. MATHEWS. "A linear coding for transmitting a set of correlated signals" *IRE Transactions on Information Theory*, IT-23:41–46, September 1956.

[47] Alejandro LEÑERO-BERACOECHEA. Programación y simulación del estándar MELP para codificación de voz a 2400 bps. Master's thesis, CINVESTAV-IPN Unidad Guadalajara, 1998.

[48] Jean M. LE ROUX. *Ondelettes et paquets d'ondelettes pour le traitement de la parole.* PhD thesis, Université Pierre et Marie Curie (Paris VI), Paris, France, December 1994.

[49] M. LYNCH, E. AMBIKAIRAJAH, and A. DAVIS. "Comparison of auditory masking models for speech coding". In *Proceedings Eurospeech'97*, pages 1495 - 1498, Rhodes, Greece, 1997.

[50] Yannick MAHIEUX and Jean Pierre PETIT. "High-quality audio transform coding at 64 kbps" *IEEE Transactions on Communications*, 42(11):3010–3019, November 1994.

[51] Stéphane MALLAT. *A Wavelet Tour Of Signal Processing.* Academic Press, second edition, 1999.

[52] Stéphane G. MALLAT. "A theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis, Machine Intelligence*, 11(7):674–693, July 1989.

[53] M. MASON, S. BOLAND, S. SRIDHARAN, and M. DERICHE. "Combined Coding of Audio and Speech Signals Using LPC and the Discrete Wavelet Transform" In *IEEE Proceedings TENCON'97*, pages 747–750, 1997.

[54] Robert J. MCAULAY and Thomas F. QUATIERI. "Speech analysis/synthesis based on a sinusoidal representation" *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, August 1986.

[55] Alan V. MCCREE and T. P. BARNWELL III. "A mixed excitation LPC vocoder model for low bit rate speech coding" *IEEE Transactions on speech and audio processing*, 3(4):242–250, July 1995.

[56] Paul MERMELSTEIN. "G.722, A new CCITT coding standard for digital transmission of wideband audio signals" *IEEE Communications Magazine*, pages 8–15, January 1988.

[57] Nicolas MOREAU. *Techniques de Compression des Signaux.* Masson, Paris, France, 1995.

[58] Peter NOLL. "Wideband speech and audio coding". *IEEE Communications Magazine*, pages 34–44, November 1993.

[59] Peter NOLL. "Digital audio coding for visual communications" *Proceedings of the IEEE*, 83(6):925 - 943, June 1995.

[60] Ted PAINTER and Andreas SPANIAS. "A review of algorithms for perceptual coding of digital audio signals" *Proceedings of the IEEE*, 88(4):449 - 513, April 2000.

[61] Davis PAN. "A tutorial on MPEG/audio compression". *IEEE Multimedia Magazine*, pages 60–74, Summer 1995.

[62] José de Jesús PÉREZ SEVILLA. "Compresión de la señal electrocardiográfica basada en el análisis de ondeletas e implementación en un procesador de señales" Master's thesis, CINVESTAV-IPN Unidad Guadalajara, Guadalajara, México, Septiembre 1997.

[63] Marcos PERREAU-GUIMARAES, Madeleine BONNET, and Nicolas MOREAU. "Allocation binaire et déconvolution psychoacoustique de complexité réduit dans un codeur audio de haute qualité". In *Dix-septième colloque GRETSI*, pages 881 – 884, Vannes, France, September 1999.

[64] Pierrick PHILIPPE, François MOREAU DE SAINT-MARTIN, and Michel LEVER. "Wavelet packet filterbanks for low time delay audio coding" *IEEE transactions on speech and audio processing*, 7(3):310 – 322, May 1999.

[65] Pierrick PHILIPPE, François MOREAU DE SAINT-MARTIN, Michel LEVER, and Joël SOUMAGNE. "Optimal wavelet packets for low delay audio coding" In *IEEE Proceedings ICASSP'96*, pages 550 – 553, 1996.

[66] John G. PROAKIS and Dimitris G. MANOLAKIS. *Digital Signal Processing.* Prentice Hall, Englewood Cliffs, NJ, third edition, 1996.

[67] Lawrence RABINER and Biing-Hwang JUANG. *Fundamentals of speech recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993.

[68] Kannan RAMCHANDRAN, Martin VETTERLI, and Cormac HERLEY. "Wavelets, subband coding, and best bases" *Proceedings of the IEEE*, 84(4):541 – 560, April 1996.

[69] Olivier RIOUL and Martin VETTERLI. "wavelets and signal processing" *IEEE SP Magazine*, pages 14–38, October 1991.

[70] Khalid SAYOOD. *Introduction to data compression.* Morgan Kaufmann Publishers, 1996.

[71] Deepen SINHA and Ahmed H. TEWFIK. "Low bit rate transparent audio compression using adapted wavelets". *IEEE Transactions On Signal Processing*, 41(12):3463 – 3479, December 1993.

[72] A. K. SOMAN and P.P. VAIDYANATHAN. "On Orthonormal Wavelets and Paraunitary Filter Bank" *IEEE Transactions on Signal Processing*, 41(3):1170–1183, March 1993.

[73] Andreas S. SPANIAS. "Speech coding: A tutorial review" *Proceedings of the IEEE*, 82(10):1541–1582, October 1994.

[74] Pramila SRINIVASAN and Leah H. JAMIESON. "High quality audio compression using an adaptative wavelet packet decomposition and psychoacoustic modeling" *IEEE Transactions on Signal Processing*, 46(4):100–108, April 1998.

[75] Dan ŞTEFĂNOIU, Kastantin RADWAN, and Gang FENG. "Speech coding based on the discrete-time wavelet transform and human auditory system properties" In *Proceedings Eurospeech'95*, pages 661 – 664, Madrid, Spain, 1995.

[76] Oswald STEWARD. *Functional Neuroscience.* Springer-Verlag, New York, NY, 2000.

[77] Gilbert STRANG and Truong NGUYEN. *Wavelets and filter banks.* Wellesley-Cambridge Press, 1997.

[78] Ahmed TEWFIK. "Digital watermarking". *IEEE Signal Processing Magazine*, pages 17–18, September 2000.

[79] P. P. VAIDYANATHAN. "Multirate digital filters, filter banks polyphase networks and applications: a tutorial" *Proceedings of the IEEE*, 78(1):56 – 93, January 1990.

[80] P.P. VAIDYANATHAN. *Multirate systems and filter banks.* Prentice Hall Signal Processing Series, Englewood Cliffs, NJ, 1993.

[81] Martin VETTERLI and Cormac HERLEY. "Wavelets and Filter Banks: Theory and Design" *IEEE Transactions os Signal Processing*, 40(49):2207 – 2232, September 1992.

[82] Martin VETTERLI and Jelena KOVAČEVIĆ. *Wavelets and Subband Coding.* Prentice Hall, Englewood Cliffs, NJ, 1995.

[83] E. ZWICKER and H. FASTL. *Psychoacoustics: Facts and Models.* Springer-Verlag, Berlin, 1990.

CENTRO DE INVESTIGACION Y DE ESTUDIOS AVANZADOS DEL IPN
UNIDAD GUADALAJARA

El Jurado designado por la Unidad Guadalajara del Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, aprobó la tesis: "Codificación Perceptual de Audio Banda Extendida Usando Paquetes de Ondeletas" del Sr. Miguel Ángel Alonso Arévalo, el día 24 de Agosto de 2001.

Dr. Manuel Edgardo Guzmán Rentería
Investigador Cinvestav 3A
CINVESTAV DEL IPN
Guadalajara

Dr. Arturo Veloz Guerrero
Investigador Cinvestav 3A
CINVESTAV DEL IPN
Guadalajara

Dr. Deni Librado Torres Román
Investigador Cinvestav 3A
CINVESTAV DEL IPN
Guadalajara

Dr. Oscar Yáñez Suárez
Profesor Titular C
Universidad Autónoma Metropolitana - Iztapalapa
México D.F.