Spectral Coding and Post-Processing of High Quality Audio

DISSERTAÇÃO APRESENTADA
À FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO
PARA OBTENÇÃO DO GRAU DE DOUTOR
EM ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES

por

Aníbal João de Sousa Ferreira

Porto
Maio de 1998

This Dissertation has been formally discussed and approved on November 6, 1998.

Thesis Committee

PRESIDENT:	Prof. Pedro G. Oliveira, *Faculdade de Engenharia, Porto, Portugal*
FIRST ARGUER:	Prof. Nikil Jayant, *Georgia Institute of Technology, USA*
SECOND ARGUER:	Prof. Francisco O. Restivo, *Faculdade de Engenharia, Porto, Portugal*
THESIS ADVISOR:	Prof. Mário J. Leitão, *Faculdade de Engenharia, Porto, Portugal*
THESIS CO-ADVISOR:	Prof. José M. Tribolet, *Instituto Superior Técnico, Lisboa, Portugal*
	Prof. Artur P. Alves, *Faculdade de Engenharia, Porto, Portugal*

Resumo

A representação digital de um *medium contínuo* como por exemplo, áudio, permite o projecto de sistemas de comunicação obedecendo a requisitos específicos de qualidade, flexibilidade e controlo.

Os enormes e permanentes avanços tecnológicos que se têm verificado, particularmente nas duas últimas décadas, conduziram a uma crescente disponibilidade de capacidade de processamento a preços reduzidos, e estimularam uma vasta actividade de investigação na área da codificação eficiente de sinais áudio.

Apesar de inicialmente circunscrita à codificação de sinais de voz, esta actividade ampliou-se, no final dos anos setenta, para a codificação de sinais de áudio de alta qualidade. O aparecimento do Disco Compacto, em 1985, definiu uma referência consensual para o áudio de alta qualidade, e imprimiu um *momentum* entusiástico àquela actividade.

O interesse na área foi ainda ampliado pelos ganhos de codificação, absolutamente inauditos, entretanto conseguidos com um novo paradigma de codificação fortemente enraizado nas características de um sistema fascinante: o sistema auditivo humano. O paradigma ficou conhecido por Codificação Perceptiva de sinais áudio.

Tratando-se de um componente estrutural do conceito multimédia, o áudio tem também vindo a ser reconhecido pelo seu papel valorizador das aplicações multimédia, nomeadamente por acentuar o seu carácter interactivo, e por permitir funcionalidades não-convencionais, tais como alteração da escala dos tempos e alteração do "*pitch*". Mas o seu potencial para aplicações novas ou não-convencionais é enorme, bastando considerar as possibilidades oferecidas pelos dispositivos de armazenagem digital de baixo custo, e as redes digitais de comunicação permitindo modos de transferência síncronos ou assíncronos, canais de débito constante ou variável, de alta ou de baixa capacidade de transmissão.

Em consequência, tem-se verificado uma grande pressão internacional para a definição de normas vocacionadas para a representação de informação multimédia codificada, e de áudio em

particular, de modo a promover uma vasta compatibilidade entre equipamentos e aplicações. Estes esforços têm sido sobretudo desenvolvidos no âmbito da Organização Internacional de Normalização, e muito particularmente pelo sub-grupo ISO/IEC JTC1/SC29/WG11, também conhecido por Grupo de Especialistas em Imagens Móveis (MPEG). Este grupo já produziu duas normas importantes: o MPEG-1 em 1992, e o MPEG-2 em 1994. Presentemente, encontra-se em curso um novo período de actividades de normalização, o MPEG-4, que visa aprovar no final de 1998 uma nova norma, estruturada no conceito de funcionalidades, para a comunicação, acesso e manipulação de informação audio-visual. Existem já também planos para uma nova fase de normalização, o MPEG-7, cujo objectivo é produzir uma norma para a descrição de informação multimédia, facilitando o seu acesso, indexação e recuperação.

O trabalho descrito nesta Dissertação foi desenvolvido em paralelo com - e por vezes contribuindo para - as actividades de normalização MPEG-4 Áudio. Apresentam-se os conceitos e as técnicas mais relevantes na codificação de sinais de voz e áudio, investigam-se alguns tópicos mais críticos associados à sua análise, modificação e síntese, caracteriza-se o estado-da-arte em codificação de áudio de alta qualidade e descreve-se um novo Codificador Espectral Áudio (ASC). Este codificador foi projectado para beneficiar de uma grande eficiência de codificação, de uma grande flexibilidade funcional, e de múltiplas funcionalidades. A investigação aqui relatada enquadra três ideis fundamentais que constituem os fundamentos da Tese e, consequentemente, dos resultados e contribuições mais importantes apresentados nesta Dissertação.

- O desempenho de um codificador de áudio de alta qualidade traduz um problema multidimensional envolvendo, em particular, a qualidade subjectiva do sinal áudio codificado, o débito binário da representação codificada, o atraso de codificação e a complexidade algorítmica do sistema codificador. Um desempenho óptimo é conseguido maximizando o primeiro parâmetro e minimizando os restantes três. Dado que estes parâmetros não são independentes, o problema do desempenho óptimo equaciona-se pela pesquisa de um extremo global cuja investigação exige um ambiente interactivo de simulação e avaliação especialmente concebido.

- Os codificadores perceptivos usam tipicamente uma escala de frequências não-linear para analisar e quantificar sinais áudio. Porém, os sinais harmónicos representam uma classe extremamente importante de sinais áudio cujo perfil é revelado de forma clara numa escala de frequências linear. Por outro lado, os sinais harmónicos são percebidos como um efeito combinado e não como uma soma de efeitos individuais, o que revela que os sinais harmónicos possuem propriedades peculiares de mascaramento ainda não satisfatoriamente compreendidas. Nem tão pouco são explicitamente exploradas pelos actuais codificadores perceptivos. Uma possível solução para melhorar a modulação do ruído de quantificação no caso de sinais harmónicos, consiste em monitorizar a estrutura harmónica relevante do sinal e estabelecer, em concordância, um perfil *global* de quantificação harmónica.

- Para além da codificação, outras funcionalidades adicionais como a igualização, a geração sintética de ruído e sua modulação, e expansão da escala de tempos, podem ser eficientemente implementadas, tirando partido de algumas propriedades resultantes de uma reformulação adequada do banco de filtros de ASC. Deste modo, evita-se o recurso a módulos de pós-processamento independentes.

Abstract

The digital representation of a *continuous medium* like audio, bears the possibility to design communication systems with specific requirements of quality, flexibility and control.

The permanent and impressive technological advances that have been achieved, particularly in the last two decades, have led to an increasing availability of low-cost computing power, and stimulated a vast research activity in the area of efficient coding of audio signals.

This activity was initially focussed on the coding of speech signals but expanded, by the late seventies, to the coding of high quality audio signals. It gained an enthusiastic *momentum* after the appearance of the Compact Disc in 1985, which has established a reference of general consensus for high quality audio signals.

The interest in this area has been further stimulated by the substantial coding gains allowed by a new coding paradigm deeply rooted on the characteristics of a marvelous perception mechanism: the human auditory system. It became known as Perceptual Coding of audio.

Being a structural component of the multimedia concept, audio has been recently recognized for its importance in the design of multimedia applications, allowing for interactivity and permitting new or unconventional functionalities such as pitch or speed change. The potential for new or improved applications is enormous considering the possibilities offered by low-cost digital storage devices and digital communication networks allowing synchronous or asynchronous transfer modes, constant or variable bit rate channels, very high or very low data rate channels.

As a consequence, pressure has been placed on the definition of international standards for the efficient representation of coded multimedia information, and audio in particular, in order to promote wide compatibility between equipments and applications. Such efforts have been mainly developed within the International Standardization Organization, particularly by the ISO/IEC JTC1/SC29/WG11 group, also known as the Moving Pictures Experts Group (MPEG). This group already produced two important standards: the MPEG-1 in 1992, and the MPEG-2 in 1994. Currently, a new period of activities, the MPEG-4, is in progress and aims to approve

a new standard by the end of 1998, which is structured on the concept of functionalities for the communication, access and manipulation of audio-visual information. Plans already exist for a new phase, the MPEG-7, targeted at standardizing a suitable description of multimedia information allowing for its easy access, indexing and retrieval.

The work described in this Dissertation has been developed in parallel with - and sometimes contributing to - the MPEG-Audio standardization activities. The main concepts and techniques in speech and audio coding are presented, a few critical topics concerning the analysis, the modification and the synthesis of audio signals are investigated, the state-of-the-art in high quality audio coding is characterized, and a new Audio Spectral Coder (ASC) designed for coding efficiency, flexibility and multiple functionalities is described. Three main ideas drove this research which represent the grounds of the Thesis and consequently, the most important contributions and results presented in this Dissertation.

- The optimization of the performance of a high quality audio coder is a multidimensional problem involving, in particular, the subjective audio quality, the bit rate of the audio coded representation, the coding delay and the algorithmic complexity. Optimum performance is achieved by maximizing the first parameter and minimizing the remaining three. Since these parameters are not independent, the problem of optimal performance corresponds to the search for a global extreme whose investigation demands a specially designed interactive simulation and evaluation environment.

- Perceptual Audio Coders use typically a non-linear frequency scale to analyze and quantize audio signals. However, harmonic signals represent an important class of audio signals whose structure is clearly revealed on a linear frequency scale. Furthermore, harmonic signals are perceived as a combined effect rather than the sum of individual effects, which suggests they have peculiar masking properties not yet completely understood. Not even are they explicitly exploited by actual perceptual audio coders. A possible solution to improve the modulation of the quantization noise in the case of harmonic signals, is to monitor their relevant harmonic structure, and to set up a *global* harmonic quantization profile accordingly.

- Besides coding, additional functionalities such as digital equalization, generation and modulation of synthesized noise, and time scale expansion, can be efficiently implemented by taking advantage of the properties of a suitable formulation of the ASC analysis/synthesis filter bank. With this approach, independent post-processing modules may be spared.

Résumé

La représentation numérique d'un *medium* continu comme par exemple, audio, permet le projet de systèmes de communication obéissant à des exigences précises de qualité, flexibilité et contrôle.

Les progrès technologiques permanents produits en particulier pendant les vingt dernières années, on conduit à une disponibilité croissante de la capacité de traitement et ceci à des prix de plus en plus bas, stimulant la recherche dans le domaine du codage efficace des signaux audio. Bien que cette activité a trouvé ses racines dans le cadre du codage des signaux de voix, elle s'est aussitôt étendue au domaine du codage des signaux audio de très haute qualité. L'apparition du *Compact Disc* en 1995 a établi une référence de qualité consensuelle pour l'audio de haute qualité et a aussi imprimé un *momentum* très fort à l'activité de recherche.

L' intérêt du domaine a été encore renforcé par les gains extraordinaires permis par un nouveau paradigme de codage fondé sur le système auditif humain, appelé Codage Perceptif de signaux audio.

S'agissant d'un élément structurel du concept *multimedia*, l'audio est devenu progressivement valorisé soit parce qu'il accentue le caractère interactif des applications multimedia, soit parce qu'il permet de nouvelles fonctionnalités telles que la modification du "*pitch*" ou de la vitesse de reproduction. Bien d'autres possibilités et avantages sont vraisemblables, il suffit de considérer le potentiel permis par les dispositifs numériques de stockage d' information, et par la vocation des divers genres de réseaux de communication qui existent aujourd'hui.

En conséquence, une pression internationale très forte s'est faite sentir autour de la necessité des normes internationales concernant la représentation numérique d'information multimedia codée, notamment audio, de façon à assurer une très vaste compatibilité entre équipements et applications. Ces efforts ont été surtout développés au sein de L'Organisation Internationale de Normalisation, et en particulier par le groupe ISO/IEC JTC1/SC29/WG11, qui est aussi connu sous le nom de MPEG. Ce Groupe a déjà fait approuver deux normes importantes: le MPEG-1 en 1992, et le MPEG-2 en 1994. À présent, une nouvelle période d'activités de normalisation,

le MPEG-4, se trouve en cours. Son objectif est le développement et l'acceptation vers la fin de 1998, d'une norme structurée sur le concept de fonctionnalités, et concernant l'accès, la communication et la manipulation d'information multimedia.

De nouveaux plans existent déjà pour une nouvelle période d'activités, le MPEG-7, dont le but est la spécification normalisée de la description d'information multimedia, ce qui facilitera son accès, son indexation et sa sélection.

Le travail décrit dans cette Dissertation a été développé en parallèle avec les activités de normalisation, et parfois a même apporté une contribution à ces activités. Les concepts et les techniques plus importants en codage des signaux de voix et audio y sont présentés, certains aspects critiques concernant l'analyse, la modification et la synthèse des signaux audio font l'objet de notre recherche, l'état-de-l'art en matière de codage d'audio de haute qualité est caractérisé, et un nouveau Codeur Spectral d' Audio (ASC) est décrit. Ce codeur a été conçu pour bénéficier d'une grande efficacité de codage, d'une grande flexibilité et de plusieurs fonctionnalités. Trois idées fondamentales encadrent la recherche présentée dans cette Dissertation et représentent aussi la Thèse et par conséquent les résultats et les contributions plus importants présentés dans cette Dissertation.

- L'optimisation de la performance d'un codeur d'audio de haute qualité est un problème à plusieurs dimensions, engageant en particulier la qualité subjective du signal audio codé, le délai de codage et la complexité du système codeur. L'état optimal est atteint si le premier de ces paramètres est maximisé et si les trois derniers paramètres sont minimisés. Étant donné que tous ces paramètres ne sont pas indépendants, le problème de la performance optimale se formule par la quête d'un extremum global ce qui demande un environnement interactif de simulation et d'évaluation spécialement préparé.

- Typiquement, les codeurs perceptifs d' audio utilisent une échelle non-linéaire de fréquences pour analyser et quantifier les signaux audio. Cependant, les signaux harmoniques représentent une classe très importante de signaux audio dont la structure est clairement visible sur une échelle linéaire de fréquences. De plus, les signaux harmoniques sont plutôt perçus comme un seul effet combiné que comme la somme de plusieurs effets individuels, ce qui montre qu'ils possèdent des propriétés de masquage pas encore complètement éclaircies. Ceci explique que ces propriétés ne sont pas convenablement exploitées par les codeurs perceptifs actuels. De façon à améliorer la modulation du bruit de quantification dans le cas des signaux harmoniques, on peut par exemple surveiller la structure harmonique du signal audio, et établir en accord avec cette structure, un profile harmonique et *global* de quantification.

- Si le banc de filtres d'analyse et de synthèse de ASC est convenablement reformulé, certaines propriétés résultent permettant l'implémentation efficace d'autres fonctionnalités additionnelles au delà du codage, comme par exemple, l'égalisation numérique, la génération et modulation contrôlée de bruit synthétique, et l'extension de l'échelle du temps. Ceci permet une économie de modules de traitement indépendants.

Agradecimentos

O caminho trilhado para chegar até esta Dissertação de doutoramento foi longo mas gratificante. Implicou trabalho árduo, beneficiou de circunstâncias excepcionais e oportunidades felizes, soçobrou nalguns pequenos reveses, mas avançou, confiante, estimulado também por alguns resultados que mereceram reconhecimento junto da Comunidade Científica internacional.

Contudo, calcorreá-lo não foi seguramente o resultado de um empenho solitário. Muitas pessoas e instituições emprestaram-lhe um importante cunho e contribuíram para que ele se reflicta, hoje, numa realidade consequente. Devo-lhes, por isso, um emocionado agradecimento, sendo certo que permanecerei sempre em dívida.

Ao meu orientador de doutoramento e de sempre, Prof. Mário Jorge Leitão, e ao meu co-orientador de doutoramento, Prof. José Tribolet, pelo apoio incondicional e pelo estímulo e confiança manifestados nas fases mais importantes da minha investigação.

Ao Dr. Nikil Jayant por, em 1990/1991 e depois 1993, me ter permitido partilhar a excelência e o entusiasmo contagiante do grupo que então dirigia de Processamento de Sinal dos Laboratórios Bell da AT&T, Murray Hill, em Nova Jérsia, Estados Unidos da América. Esta vivência e os muitos contactos de que beneficiei, nomeadamente com o Dr. Schuyler Quackenbush, condicionaram definitivamente a minha forma de encarar e fazer investigação. Durante os dois períodos da minha estada nos Laboratórios Bell da AT&T, colaborei mais intensamente com o Sr. James Johnston, do que resultou o desenvolvimento, em co-autoria, da primeira versão do codificador *perceptivo* estereofónico da AT&T: PAC. Por ter partilhado comigo, sem reservas, o seu saber e perspicácia, acumulei um importante capital de conhecimento e domínio das técnicas de codificação, mas também uma dívida de difícil reparo.

A alguns investigadores distintos e membros do MPEG-4, nomeadamente o Prof.

Fernando Pereira, Dr. Karlheinz Brandenburg e Dr. Bernd Edler, que estimularam e enriqueceram a minha participação nas actividades de normalização internacional MPEG-Audio.

Ao INESC que me proporcionou excelência no ambiente, nos meios e condições de trabalho, sem os quais esta Dissertação nunca ouviria o som do mundo; além do encorajamento sempre reafirmado pelos seus directores. Devo um reconhecimento especial ao Prof. Pimenta Alves pela visão, pelo interesse e pelo permanente apoio; e ao Prof. Pedro Guedes que nunca declinou um pedido de revisão dos meus artigos, por vezes com carácter urgente, e cuja análise me comunicava sempre a autoridade e a confiança de uma certificação.

À Junta Nacional de Investigação Tecnológica pela Bolsa de doutoramento CIENCIA/BD/2069/92 que me foi concedida.

Aos meus colegas do Departamento de Electrónica e Telecomunicações da Universidade de Aveiro, do meu Departamento de Engenharia Electrotécnica e Computadores da Faculdade de Engenharia da Universidade do Porto e ao André Puga, pelo contributo de inúmeras discussões entusiasmadas e esclarecedoras. Discussões que, não raro, se sublimaram em trabalho conjunto e resultados publicados. A alguns colegas do meu departamento, em particular, devo a impagável dispensa de serviço docente para, durante cerca de um mês e meio de puro silêncio, poder concluir a escrita desta Dissertação. Agradeço também, em particular, à Prof. Isabel Martins Trancoso, ao Prof. Luís Côrte-Real, ao Prof. Francisco Restivo, ao José Vieira e João Rodrigues, os muitos comentários a partes do texto desta Dissertação.

Aos meus pais e irmão que viveram a ânsia de uma aurora em que sempre acreditaram, apesar de parecer teimar em não brilhar.

À *Betsy* e à *Sofia*, que acreditaram sempre no meu entusiasmo quase fundamentalista e o souberam pacificar pela face mais colorida. Porque o meu sorriso não é senão um corolário da sua existência, esta Dissertação é-lhes dedicada.

Acknowledgements

The road to this Dissertation has been long but rewarding. It has implied hard work, has benefited from exceptional circumstances and fortunate opportunities, has hesitated against minor difficulties, but eventually made its way receiving sometimes a stimulating recognition from the international Scientific Community.

However, that journey has certainly not been the result of a solitary commitment. Many individuals and institutions have offered their support along the way, making it possible, visible and consequent. Certain that I will be forever in debt to them, I want to express my deepest gratitude.

From my advisor, Prof. Mário Jorge Leitão, and my co-advisor, Prof. José Tribolet, I have always received an unconditional support. They have provided me with the most important opportunities during my career as a doctoral candidate.

I had a most enriching experience both scientifically and personally, at AT&T Bell Laboratories, Murray Hill, New Jersey, United States of America, in 1990/1991 and in 1993. I owe to Dr. Nikil Jayant the privilege of sharing the excellence and the enthusiasm of the Signal Processing Group he headed at that time. The felicitous guidance of Dr. Jayant and the numerous enlightening contacts I had with distinguished researchers at AT&T, notably with Dr. Schuyler Quackenbush, simply shaped definitively my attitude regarding research. In both the two periods of my stay at AT&T, I worked closely with Mr. James Johnston. This fruitful collaboration led ultimately to the co-authoring of the first version of the AT&T Perceptual Audio Coder. I am specially grateful to him for his kindness and for sharing with me his knowledge and wisdom.

Several distinguished researchers and members of the MPEG-4, including Prof. Fernando Pereira, Dr. Karlheinz Brandenburg and Dr. Bernd Edler, have stimulated and enriched my participation within the MPEG-4 Audio standardization activities. For their encouragement and trust, I am thankful.

My colleagues at the Department of Electronics and Telecommunications of the University of Aveiro, at the Department of Electrical and Computer Engineering of the Faculty of Enginnering of the University of Porto, and André Puga, were always willing to maintain fruitful discussions, which sometimes brought about spontaneous form of collaboration and joint work. I am very specially in debt to a few colleagues of my Department for giving me about one and a half month of pure silence making it possible the complete writing of this Dissertation. I am particularly thankful to Prof. Isabel Martins Trancoso, Prof. Luís Côrte-Real, Prof. Francisco Restivo, José Vieira and João Rodrigues for carefully proofreading parts of the draft of this Dissertation.

I have been affiliated with INESC since 1988. Ever since this private Research and Development Institute provided me with the best conditions and means to pursue my research, and with the encouragement of its board of directors. In particular, I am thankful to Prof. Pimenta Alves for his vision, interest and permanent support; and to Prof. Pedro Guedes for his willingness to review some of my papers. His reviews had always had on me the effect of a reassuring certification.

I am also grateful to JNICT, the Portuguese Research Council, for sponsoring the first three years of my research, under Grant No. CIENCIA/BD/2069/92.

My parents and my brother deserve a word of special appreciation for believing in a sunrise whose delay was never taken for reluctance.

To *Betsy* and *Sofia*, my beloved ones, who understood my enthusiasm and made it more colorful, I dedicate this Dissertation. My smile is just a corollary of their existence.

Contents

Resumo i

Abstract v

Résumé vii

Agradecimentos xi

Acknowledgements xiii

Abbreviations xxxv

1 Introduction 1

1.1 Context and Motivation 1

1.1.1 The Main Challenge 1

1.1.2 Source and Sink Coding 2

1.1.3 Perceptual Audio Coding 8

1.1.3.1 The Roots of the Paradigm 9

1.1.3.2 The Implementation of the Paradigm 12

1.1.4 The MPEG Standardization Activities 14

1.1.5 Additional Challenges 16

1.2 The Thesis in the Dissertation 16

1.3 Organization of the Dissertation 17

2 Designing for Maximum Performance 21

2.1 Introduction 21
2.2 The Coding Performance 22
2.3 A Visual Approach to Simulation 24
2.3.1 The Rationale 24
2.3.2 Interactive Evaluation Environment 25
2.3.2.1 A Tool for the Analysis of FIR Filters 26
2.3.2.2 Audio Time and Spectral Analysis Tool 27
2.3.3 Interactive Simulation Environment 29
2.4 Complexity Concerns 31
2.5 Summary 32

3 The Human Auditory System 33

3.1 Introduction 33
3.2 The Peripheral Structure 34
3.2.1 The Outer Ear 35
3.2.2 The Middle Ear 35
3.2.3 The Inner Ear 36
3.2.4 The Transduction Process 37
3.3 Monaural Hearing 39
3.3.1 The Hearing Area and Loudness 39
3.3.2 Critical Bandwidths and Band-Rate 41
3.3.3 Excitation and Masking 44
3.3.4 Models of Masking 46
3.3.4.1 MPEG-1 Audio Psychoacoustic Model 1 48
3.3.4.2 MPEG-1 Audio Psychoacoustic Model 2 49
3.3.4.3 A Simplified Model of Simultaneous Masking 50
3.3.5 Phase Locking and Time Resolution 52
3.3.6 The Perception of Pitch and Timbre 54

3.4 Binaural Hearing 58
3.4.1 Sound Differentiation with the two Ears 58
3.4.2 Physical and Physiological Implications 59
3.4.3 Properties of Binaural Hearing 61
3.4.3.1 The Precedence Effect 61
3.4.3.2 The Cocktail Party Effect 61
3.4.3.3 Binaural Masking Level Differences 62
3.5 A Final Comment on the Accuracy of a Psychoacoustic Model 62
3.6 Summary 63

4 Audio Decomposition Using Multirate Filter Banks 65

4.1 Introduction 65
4.2 Desired Properties of the Multirate Filter Bank 66
4.3 Theory of Multirate Filter Banks 67
4.3.1 Uniform Filter Banks 67
4.3.1.1 The General Structure of Uniform Filter Banks 67
4.3.1.2 The Relation Between Input and Output 68
4.3.1.3 Input to Output Relation Using the Polyphase Decomposition . 70
4.3.1.4 Cosine Modulated FIR Filter Banks 73
4.3.1.5 The Design of the Prototype Filter 76
4.3.1.6 Objective Coding Gain 77
4.3.1.7 Perceptual Coding Gain 80
4.3.2 Non-Uniform Filter Banks 81
4.3.2.1 The General Structure of Non-Uniform Filter Banks 81
4.3.2.2 Objective and Perceptual Coding Gain 82
4.3.3 Multiresolution Analysis/Synthesis Systems 83
4.3.4 The Choice of the Multiresolution Scheme 84
4.4 Coherence Detection Using an Uniform Filter Bank 87

4.4.1 Common Tonality Classification Methods 88
4.4.2 Tonality Detection Using the TDAC Filter Bank 90
4.4.3 A Few Results 94
4.5 Spectral Modification Effects and the Design of the Prototype Filter 97
4.5.1 A Model of Spectral Modification 98
4.5.2 The TDAC Filter Bank 99
4.5.3 Transform Windows and Requirements 101
4.5.4 Linear Filtering and Window Design 102
4.5.5 Convolution and Time Aliasing Using the Odd-DFT Filter Bank 105
4.5.6 Optimization Criteria and Results 108
4.5.7 Convolution and Time Aliasing Using the TDAC Filter Bank 115
4.6 Summary 117

5 Coding Techniques and the State-of-The-Art 119

5.1 Introduction 119
5.2 Quality Measures 120
5.3 Speech Coding 121
5.3.1 Speech: Properties, Quality and Bit-Rate Targets 121
5.3.2 Coding Schemes 123
5.3.2.1 LPC Analysis and Pitch Prediction 123
5.3.2.2 Waveform Coders 125
5.3.2.3 Vocoders 126
5.3.2.4 Hybrid Coders 127
5.4 Wideband Speech Coding 136
5.5 High Quality Audio Coding 139
5.5.1 Companding 139
5.5.1.1 Instantaneous Companding 139
5.5.1.2 Near-Instantaneous Companding 140

5.5.2 Perceptual Audio Coders . . . 141
5.5.2.1 The MPEG-1 Audio Standard . . . 143
5.5.2.2 The MPEG-2 Audio Standards . . . 146
5.5.2.3 Proprietary Audio Coding Schemes . . . 150
5.6 Summary . . . 155

6 ASC - Audio Spectral Coder **157**
6.1 Introduction . . . 157
6.2 The Encoder . . . 159
6.2.1 General Structure . . . 159
6.2.2 Time Resolution Coding Mode . . . 160
6.2.3 Time-Frequency Transformation . . . 162
6.2.4 Harmonic Analysis . . . 164
6.2.4.1 Rationale . . . 164
6.2.4.2 The Algorithm . . . 165
6.2.5 The Quantization Noise Profile . . . 167
6.2.5.1 Quantization . . . 168
6.2.5.2 Coding of Quantization Step-Sizes . . . 169
6.2.5.3 Harmonic Quantization and Coding . . . 171
6.2.5.4 Rate/Distortion Control . . . 172
6.2.6 Huffman Coding . . . 173
6.2.7 The Bit Stream Structure . . . 175
6.3 The Decoder . . . 177
6.3.1 The Decoder Structure . . . 177
6.3.2 Synthesis of Missing Harmonics . . . 178
6.4 A Few Results . . . 178
6.4.1 ASC Evaluation at High Bit Rates . . . 178
6.4.2 ASC Evaluation at Low Bit Rates . . . 179

6.5 ASC in MPEG-4 185

6.5.1 Introduction 185

6.5.2 Pre-Screening Tests 186

6.5.3 Ranking Criteria 187

6.5.4 MPEG-4 Results and Ranking of Proposals 188

6.6 Summary 191

7 ASC - Additional Functionalities 193

7.1 Introduction 193

7.2 Equalization 195

7.2.1 Introduction 195

7.2.2 Circular Convolution and Linear Convolution 197

7.2.3 A 20-Band Graphic Equalizer 200

7.2.4 Subjective Tests 203

7.3 Noise Synthesis and Modulation 204

7.3.1 Introduction 204

7.3.2 Analysis of Impaired Audio Material 205

7.3.3 Synthesis of Impaired Audio Material 209

7.4 Time Scale Expansion 211

7.4.1 Introduction 211

7.4.2 A Review of Known Techniques 213

7.4.2.1 Time Domain Techniques 213

7.4.2.2 Frequency Domain Techniques 213

7.4.2.3 Time Scaling Using Speech Models 214

7.4.2.4 Relation to the New Technique 216

7.4.3 The New Technique 217

7.4.3.1 The Central Idea 217

7.4.3.2 The Research Context and Related Constraints 218

7.4.3.3 Properties of the Odd-DFT Based Analysis/Synthesis System . 219

7.4.3.4 Constraints on the Frequency Expansion 224

7.4.3.5 Estimating the Fractional Frequency $\Delta\ell$ 224

7.4.4 Spectral Expansion of Narrow and Wideband Signal Components 226

7.4.5 A Few Results 229

7.5 Summary 235

8 Conclusion **237**

8.1 The Dissertation: Results and Contributions 237

8.2 The Future: Research Topics and New Areas 238

List of Figures

1.1 Illustration of a simplified communication chain where it is emphasized that the bit rate compression implemented at the encoder is followed by expansion at the decoder. An ideal channel is assumed. 2

1.2 Illustration of the relation between the coding gains allowed by source coding (redundancy reduction) and by sink coding (irrelevancy reduction). 8

1.3 Simplified structure of a Perceptual Coder. The opportunities to reduce the bit rate are illustrated by the lower diagram. It shows that irrelevancy reduction represents the most significant contribution to the compression gain. 13

2.1 A possible approach to the design of a new coding system. The interactive simulation and evaluation environment reduces the time needed to design, to debug and to optimize the coding system. 26

2.2 Illustration of an interactive tool (FILTRES) allowing the simultaneous analysis of multiple FIR filters. This tool classifies each filter and displays its impulse, frequency and group delay responses. 27

2.3 Illustration of ATSAT, an audio editor allowing extended spectral analysis features and listening capabilities. This example shows two coded versions of a high quality harmonic audio signal (see text). 28

2.4 Illustration of the interactive simulation environment designed to develop and optimize ASC. Each window represents the main audio data and system variables related to the coding and decoding of an audio frame (see text). 29

3.1 A sketch of the peripheral region of the human auditory system (adapted from [Beranek90]). 35

3.2 A sketch of the internal structure of the cochlea (adapted from [Silver.92]).The arrows illustrate the movement of fluids as a consequence of an inward movement of the stapes footplate at the oval window. 36

3.3 A representation inspired by the results of Fletcher and Munson [Beranek90] of equal loudness contours for pure tones. The lower contour corresponds to the threshold of audibility. Zwicker [Zwicker90] argues this contour corresponds to *3 Phon* (and not to *0 Phon*) and *0 Sone* at 1 *kHz*, whereas Moore [Moore89] suggests *2 Phon* and *0.003 Sone*. 39

3.4 Organization of the natural auditory filters. On the left: relation between the linear frequency scale (Hertz) and the critical band rate (Bark). On the right: bandwidth in Hertz of each critical band. 44

3.5 Masking experiments revealing the shape of the natural auditory filters for three different center frequencies: 200 *Hz*, 2 *kHz*, and 5 *kHz*. On the left, the solid lines represent the intensity of a masker, as a function of frequency, just masking a tonal maskee of fixed intensity and frequency (filled circles). On the right, the solid lines represent the intensity of a maskee just masked by a fixed frequency and intensity masker (filled circles). The dashed line represents the Absolute Threshold of Hearing. 45

3.6 Excitation patterns. On the left, solid lines represent the individual excitation patterns generated by tones (filled circles) at 200 *Hz* (2 *Bark*), 2 *kHz* (13 *Bark*), and 5 *kHz* (about 18.5 *Bark*). On the right, solid lines represent the variation of the excitation pattern with the intensity of a 2 *kHz* masker. The dashed lines correspond to the Absolute Threshold of Hearing. 46

3.7 Masking indexes for tone maskers (upper model) and for noise maskers (lower model), as derived in real audio coding conditions. 51

3.8 The sound of castanets shown on top, was time windowed to the lowest segment (bottom) that maintains sound recognizability. Despite a change in loudness, this 3.7 *ms* window could even be shifted to the right as much as 7 *ms* into the signal, that recognizability was not affected, provided that the borders of the time window occur in zero crossing regions. 54

3.9 Harmonic analysis of different sound items. Each bar represents the relative occurrence of short audio segments where a specific number of tones harmonically related were detected. The upper left results characterize a strongly non-stationary signal. The upper right, the lower left and lower right results characterize signals with an increasing predominance of a well defined harmonic structure. 57

4.1 An uniform filter bank with N channels. Each channel comprises one analysis filter ($H(z)$) and one synthesis filter ($F(z)$). These are necessary to avoid aliasing after decimation by M, and to avoid spectral images after interpolation by M. The filter bank is maximally decimated if M=N. 68

4.2 Polyphase representation of a filter bank with N channels. The filter bank is maximally decimated if M=N. 71

4.3 Polyphase representation of a filter bank with N channels where the *noble identities* were used to shift the filtering operations to the side of lowest possible computing rate. The filter bank is maximally decimated if M=N. 72

4.4 Modulation of a prototype filter (top) resulting in *even channel stacking* (left) or in *odd channel stacking* (right). 74

4.5 Cosine modulated filter bank with M channels. The system is critically sampled and achieves PR if the polyphase components satisfy the power complementary requirement. It is usual in practice to take N=2M. 75

4.6 An non-uniform filter bank with N channels. M_k represents the maximal decimation/interpolation factor in channel k, $0 \leq k \leq N-1$ since the bandwidth of that (real) subband is π/M_k. 82

4.7 Uniform and non-uniform time/frequency division of a real signal. On the left, the bandwidth $[0, \pi]$ of the real signal is divided into 16 uniform subbands. On the center, a dynamic switching between two uniform filter banks is illustrated: the uniform filter bank represented on the left, and another uniform filter bank with four times finer time resolution (four times coarser frequency resolution). This configurations allows two time/frequency scales. On the right, a fixed non-uniform filter bank divides the signal bandwidth into five bands organized in four octaves. This configuration allows four time/frequency scales. 84

4.8 Tonality estimation for a synthesized signal comprising four pure tones. The time representation (on the left) denotes the 50% overlap between adjacent segments of data. Tonality is detected only for three segments where the windowed signal is clearly stationary. The reminiscent tonality indexes for the last two segments are due to exponential smoothing. 95

4.9 Tonality detection for a speech signal (female vocal). A transition from a voiced region to an unvoiced region is clearly reflected on the estimated tonality index. . 96

4.10 Examples of windows that comply with the TDAC perfect reconstruction constraint ($N = 32$). 102

4.11 The dark sub-segment y_r results from a linear convolution with a zero-phase filter, or from the OA operation between transformed segments y_{m-1} and y_m. 103

4.12 A window with minimized stop-band leakage (solid line) relative to the reference window (dashed line) when $N = 32$. Top: impulse response, bottom: frequency response. 109

4.13 An optimal window (solid line) allowing gains in stop-band leakage and time aliasing relative to the reference window (dashed line) when $N = 32$. Top: impulse response, bottom: frequency response. 110

4.14 A modeled approximation (solid line) of the phase deviation (dashed line) of an optimal window from the phase of the reference window ($N = 32$). 111

4.15 An approximation given by a closed form expression (solid line) to the ideal optimal window (dashed line) when $N = 32$. 112

4.16 An optimal window (solid line in upper figures and in lower right figure) having reduced time aliasing and stop-band leakage relative to the reference window (dotted line in upper figures and solid line in lower left figure) when $N = 1024$. . 113

4.17 An optimal window (solid line in upper figures and in lower right figure) having less time aliasing but more stop-band leakage than the reference window (dotted line in upper figures and solid line in lower left figure) when $N = 1024$. The higher stop-band leakage is primarily due to the higher first lobe in the stop-band. 114

5.1 Simplified structure of the ITU G.726 ADPCM standard. 126

5.2 Simplified structure of a CELP coder. 128

5.3 Simplified block diagram of the Zelinsky and Noll Adaptive Transform Coder. . . 131

5.4 Simplified block diagram of the Tribolet and Crochiere speech specific ATC. . . . 132

5.5 Flexible shaping in frequency of the quantization noise between two limits: white noise and constant SNR in frequency. 132

5.6 Simplified schematic representation of a Multiband Excited encoder (on the left) and decoder (on the right). 135

5.7 Simplified structure of the ITU G.722 SB-ADPCM standard. 137

5.8 Near-Instantaneous companding characteristic of the ITU-T J.41 standard. . . . 140

5.9 Basic section of the MPEG-1 Audio encoder that is common to all layers. 144

5.10 Complementary section of the MPEG-1 Audio encoder for Layer I or Layer II. . 144

5.11 Complementary section of the MPEG-1 Audio encoder for Layer III. The diagram represents the finest spectral resolution of the hybrid filter bank. However, the resolution is adaptively controlled by means of a window switching mechanism (WS) that modifies the segmentation on each subband (see text). 145

5.12 Adaptive segmentation within the Layer III hybrid filter bank: monotonic segmentation (on the left) using 36-point "long" windows, dynamic segmentation (on the right) using "long" and 12-point "short" windows, as well as two special "start" and "stop" windows necessary to comply with transition requirements. . 145

5.13 Structure of the MPEG-2 Advanced Audio Coder (AAC). 148

6.1 A simplified diagram of the ASC encoder. 159

6.2 Possible window switching modes allowing an improvement of the coding resolution in time by factors 2, 4, and 8 (from top to bottom). 161

6.3 Spectral envelope of the LPC predictor used to detect signal transients. 162

6.4 Difference between the sampled frequencies of the DFT and those of the Odd-DFT. In this Figure, N=8 and reflects one period (2π) of the digital frequency. . 164

6.5 The TDAC spectrum is upper-bounded by the Odd-DFT spectrum. In specific cases, this relation might be misleading (see text). 167

6.6 ASC decimates the quantization step-sizes as illustrated. The coding of spectral coefficients assumes the interpolated QNP as illustrated on the right-hand side. . 169

6.7 Possible solutions for the interpolation of quantization step-sizes (see text). . . . 170

6.8 Modeling of the QNP of 4 short windows according to the start of the signal non-stationarity (dark areas): a) single QNP model for all short blocs, b), c), d) two QNPs for the short windows before and after the start of the non-stationarity. 171

6.9 General structure of the bit stream generated by ASC. The unit identified as Variable Length Header displays all possible fields in that header although some of them are mutually exclusive (see text). 175

6.10 A simplified diagram of the ASC decoder. 177

6.11 ASC coding bit rates for the transparent coding of a "mix" of critical audio signals sampled at 44.1 kHz, as a function of the segment length (which is two times the frame length). These results were derived using a version of ASC significantly less optimized than the current version. 179

6.12 Coding of a harmonic signal by ASC. Upper left: time segment. Upper right: power density of the Odd-DFT spectrum. Lower left: spectral density in a Bark related frequency scale of the signal power (dotted line), of the estimated Just Noticeable Distortion (dashed line), and of the final Quantization Noise Profile (solid line) that complies with bit rate constraints. Lower right: spectral density of the Quantization Noise Profile in the linear frequency scale after taking into account the relevant harmonic structure of the signal. 180

6.13 Coding (using ASC) of a 232 ms time segment of "pitch pipe" with and without the harmonic analysis and coding technique. Upper left: 1/10 of the total time segment. Upper right: spectrum of the original signal. Lower left: spectrum when coding without the technique (-HC). Lower right: spectrum when coding with the technique (+HC). 181

6.14 First attack of "castanets" coded at $32Kbit/s$ by ASC (upper right), MPEG Layer 3 (lower left), and MPEG Layer 2 (lower right). 183

6.15 Spectrum of "pitch pipe" (upper left) after being coded at $32Kbit/s$ by ASC (upper right), by MPEG Layer 3 (lower left), and by MPEG Layer 2 (lower right). 184

6.16 MPEG-4 Audio test results for the functionality of compression at 16 $Kbit/s$. The test signals whose first two initials are "MN" as well "g722", are reference or anchor signals (see text). 188

6.17 MPEG-4 Audio test results for the functionality of compression at 24 $Kbit/s$. The test signals whose first two initials are "MN" as well "mp24", are reference or anchor signals (see text). The symbols numbered 1 – 5 denote specific test items. 189

6.18 MPEG-4 Audio test results for the functionality of compression at 40 $Kbit/s$. The test signal identified as "mp40" is an anchor signal. The symbols numbered 1 – 5 denote specific test items (see text). 190

7.1 Reformulation of the ASC analysis/synthesis framework in order to accommodate pre/post-processing. 194

7.2 Reformulation of the conventional OA/OS methods in order to simplify the design of the equalization filter. 196

7.3 Simplified scheme of the overlap-add (OA) and overlap-save (OS) methods when $N_{data} = N/2$. These methods are the dual of each other in the sense that the analysis and synthesis windows, as well as their role, are reversed in the two cases. As illustrated, an hybrid OA/OS method is obtained by simply using suitable analysis/synthesis windows. 197

7.4 Illustration of the analysis/synthesis time windows used in a 50% OA/OS method when the equalization filter is zero-phase. Bottom: a hybrid OA/OS method is possible where both windows are symmetric and identical. 199

7.5 Frequency response of the 20-band digital audio equalizer. 202

7.6 Example of a desired equalization curve. 203

7.7 Example of a smoothed equalization curve. 203

7.8 Frequency response of the approximated equalization filter. 203

7.9 Impulse response of the approximated equalization filter. 203

7.10 Generation of two equalized signals for subjective evaluation. One version is obtained by linear convolution in the time domain using the ideal equalization filter. The other version is obtained by filtering in the frequency domain. 204

7.11 Time and spectral representation of two examples of background noise. a) the upper example was taken from the sound-track of a 1944 movie and b) the lower example was taken from the sound-track of a telecine documentary about the Beatles. 206

7.12 Simplified spectral model of the background noise found in the "Bluebeard" sound-track. 207

7.13 Example of scratches: isolated (upper left), close but separated (upper right and lower left), and merged (lower right). 208

7.14 Time and spectral representation of a tonal signal contaminated by impulsive noise. 208

7.15 Adaptation of the ASC analysis/synthesis filter bank in order to synthesize impaired audio material from unimpaired audio material. 209

7.16 Examples of synthesized impairments: a) colored noise plus scratches (upper figure), b) impulsive noise added to a pure tone (lower figure). 210

7.17 Illustrative representation of the new frequency domain approach to time-scale expansion. From top to bottom: (A) original signal, (B) spectral modification, (C) same as B with sampling frequency $\frac{2}{K}F_N$ instead of $2F_N$, (D) same as C after interpolation by K. 217

7.18 Phase offset when $\Delta\ell$ varies from 0.05 (top) to 0.95 (bottom). N is 256 and ℓ is 12. 222

7.19 Spectral modifications when a tone is modulated in time by a real exponential. In all cases N is 256 and $\Delta\ell = 0.5$. From top to bottom: $\ell = 12$ and $\alpha = 0.0$, $\ell = 12$ and $\alpha = 2.0$, $\ell = 24$ and $\alpha = 2.0$. 223

7.20 Passband and stopband approximations (solid line) to the frequency response of the analysis/synthesis window (dotted line). 226

7.21 Estimated fractional frequency (solid line) as a function of the exact synthesized fractional frequency (dotted line) for 200 tones. $N = 1024$ and $\ell = 12$. 227

7.22 Frequency expansion of a four tone harmonic signal by a factor of two. Top: original signal. Middle: synthesized (frequency expanded) signal. Bottom: processed signal. For clarity only 1/4 of the time segment is displayed. The sampling frequency is 48 *kHz*, $N = 1024$ and $(\ell + \Delta\ell) = 12.3$. 230

7.23 Frequency expansion of a time modulated tone by a factor of two. Top: original signal. Middle: synthesized (frequency expanded) signal. Bottom: processed signal. The sampling frequency is 48 *kHz*, $N = 1024$ and $(\ell + \Delta\ell) = 12.3$. 231

7.24 Frequency expansion of a tonal pulse by a factor of two. Top: original signal. Middle: synthesized (frequency expanded) signal. Bottom: processed signal. The sampling frequency is 48 *kHz*, $N = 1024$ and $(\ell + \Delta\ell) = 12.3$. 232

7.25 Time expansion of a female speech passage, *'time'*, by a factor of two. Top: original signal. Middle and bottom: processed signal. The original sampling rate is 16 *kHz* but the processing was performed using 32 *kHz* sampling rate and $N = 1024$. 234

7.26 Time expansion of a female speech passage, *'to'*, by a factor of two and three. Top: original signal. Middle: time-expanded version by a factor of two. Bottom: time-expanded version by a factor of three. The original sampling rate is 16 kHz but the processing was performed using 32 kHz sampling rate for two-times expansion, 48 kHz sampling rate for three-times expansion, and in both cases $N = 1024$. 235

List of Tables

4.1 Detection of transitions between quasi-stationary regions and non-stationary regions of a few test items. The number of transitions indicated is relative to the total number of processed segments (each one being about 12 ms long) which reflects the duration of each test item. 86

4.2 Reduction of the (Odd-DFT) time aliasing and stop-band leakage when $N = 1024$, relative to the reference window. 112

4.3 Reduction of the (TDAC) time aliasing and stop-band leakage when $N = 1024$, relative to the reference window. 117

5.1 Typical quality parameters of different types of PCM audio. 120

5.2 5-point grading scale of subjective impression regarding quality (MOS) and impairment (ITU-R scale) of coded material. 121

6.1 Characteristics of the Huffman CodeBooks used in ASC. The figures under Efficiency denote the reduction in bits per input value. 174

6.2 Signal to Noise Ratio of signals coded at 32 $Kbit/s$ by ASC (without [-HC] and with [+HC] the harmonic analysis and coding technique), and by the standard MPEG Layer 3 and Layer 2. 182

6.3 Coder specifications. 184

6.4 Proponents to the MPEG-4 Audio competitive tests. 186

6.5 Test items used in the 24 $Kbit/s$ and 40 $Kbit/s$ compression tests. 189

7.1 Center frequency of each band-pass filter of the equalizer. 201

7.2 An example of equalization settings. 201

7.3 Qualitative evaluation of the time-scaling algorithm for 4 increasingly difficult signals. The SNR figures indicate the errors in the time domain introduced by the spectral modification process relative to the ideal output signal (see text). $N = 1024$ and the duration of the signals is 4 audio frames ($4\frac{N}{2}$). 233

Abbreviations

AAC Advanced Audio Coding

AbS Analysis by Synthesis

ADPCM Adaptive Differential Pulse Code Modulation

ASC Audio Spectral Coder

ATC Adaptive Transform Coding

CD Compact Disc

CELP Code Excited Linear Prediction

DAB Digital Audio Broadcasting

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

FFT Fast Fourier Transform

ITU International Telecommunications Union

JND Just Noticeable Distortion

LAR Log Area Ratios

LPC Linear Predictive Coding

LSP Line Spectral Pairs

MBEV Multiband Excited Vocoder

MDCT Modified Discrete Cosine Transform

MPEG Moving Pictures Experts Group

MOS Mean Opinion Score

NICAM Near Instantaneous Companding Audio Multiplex

OA Overlap-Add

OS Overlap-Save

PAC Perceptual Audio Coder

PARCOR Partial Correlation Coefficients

PCM Pulse Code Modulation

QMF Quadrature Mirror Filter Bank

QNP Quantization Noise Profile

SFM Spectral Flatness Measure

SNR Signal to Noise Ratio

TDAC Time Domain Aliasing Cancellation

THD Total Harmonic Distortion

VQ Vector Quantization

Chapter 1

Introduction

Sound. It's not just the air vibrating.
Sound means feelings. It refreshes our minds.
Sues our hearts. It can make us happy, or sad, or excited.

sentences pronounced in a sound test item (male speech).

1.1 Context and Motivation

1.1.1 The Main Challenge

A high quality digital audio signal corresponds to the discrete representation, in both time (sampling) and amplitude (quantization), of an analog audio signal whose bandwidth and dynamic range are representative of the listening capabilities of the human auditory system. Figures such as 20 kHz and about 100 dB are usually considered, respectively, for the bandwidth and the dynamic range of high quality audio signals. However, a more important criterion of quality is a less quantifiable measure associated to the subjective impression of signal purity, *i.e.* the inaudibility of all forms of distortions, noises and artifacts not belonging to the original signal. Therefore, high quality does not preclude the existence of other signal components, namely quantization noise, as long as they remain inaudible.

The basic format of a digital audio signal is Pulse Code Modulation (PCM) whose data rate increases linearly with the sampling frequency and the sample resolution. For example, the raw data rate of a monophonic audio channel, as stored on a Compact Disc (CD), is about 706 $Kbit/s$ since the sampling frequency is 44.1 kHz and the sample resolution is 16 bit per sample. Doubling the sample resolution brings only a residual quality improvement, while halving the sample resolution dramatically reduces the quality. In fact, this corresponds to raising or to lowering the noise floor of the digital signal, equally at all frequency components by the same amount, which is clearly not commensurate with the perception of an identical amount of increased or decreased quality. The reason is that this approach to increase or decrease the data rate is not adapted to the characteristics of human perception. In fact, regarding subjective quality, the PCM format is only able to prevent all possible forms of perceived degradation, in a strict conservative sense, and at an extraordinary cost of data rate. It is therefore very inefficient and not acceptable for a large number of applications willing to offer high quality audio but lacking large and cheap storage capabilities, or lacking high bandwidth transmission channels.

An efficient digital representation of high quality audio signals should therefore minimize the use of storage or transmission resources, without sacrificing the subjective quality of the audio signal, as perceived by a human listener. In addition, such an efficient alternative to the PCM format should be obtained through a coding method implying a minimized cost in system complexity and delay.

This challenge was the first and main topic driving the research in this Dissertation. Multichannel audio coding was also admitted in the first research plan but it was later abandoned as will be explained later on. For this reason, it will henceforth be implicit that the terms compression or coding are synonymous with the reduction of the PCM data rate of a monophonic audio signal.

1.1.2 Source and Sink Coding

The opportunity to compress an audio signal lies in some knowledge of useful models concerning the two end points of a communication chain, as represented in Figure 1.1: the *source* and the *sink*. A source model elicits the concepts of signal *redundancy* and source coding[1]. A sink model

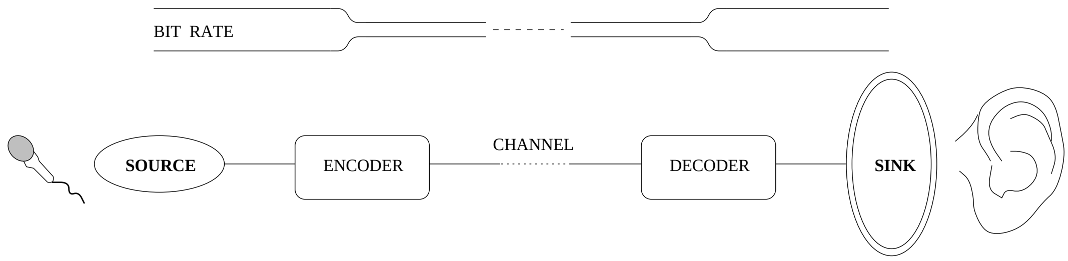

Figure 1.1: Illustration of a simplified communication chain where it is emphasized that the bit rate compression implemented at the encoder is followed by expansion at the decoder. An ideal channel is assumed.

[1]We will assume in this Dissertation that source coding consists in a processing acting upon the signal coming out of the source, using models of the source. This definition encompasses a narrower scope than that usually assumed in the Information Theory literature.

elicits the concepts of signal *irrelevancy* and sink coding[2].

For the purpose of definition, we will consider that redundancy is a statistical feature of the source that refers to the predictable structure of a stationary waveform at its output. If the output samples are statistically independent (memoriless source), redundancy results solely from the non-equiprobable distribution of the sample amplitudes. If the source exhibits memory, *i.e.* if the output samples are statistically dependent, redundancy results, in addition, from the existing inter-sample correlation. This latter aspect is revealed, for example, by a non-flat power spectral density of the source signal. A compression gain can be achieved by reducing the redundancy of the source. This process can be made lossless or lossy, in which case a non-zero approximation error to the source signal is tolerated.

Taking into account that we are dealing only with discrete-time sources, lossless coding is only possible for discrete amplitude source signals since in this case, it is possible to estimate the minimum rate necessary for an exact (lossless) replication of the source signal. According to the classical theory of Shannon, this estimate reflects a fundamental limit called the *entropy* (or the average information content) of a memoriless discrete source:

$$\text{entropy} = H(x) = -\sum_{k=0}^{M-1} p_k \log_2 p_k \quad (\text{bit/symbol}) \,, \tag{1.1}$$

which is upper bounded by $\log_2 M$, where M is the number of symbols (or possible amplitude levels) of the discrete source signal. In Equation (1.1) p_k is the probability of occurrence of the k_{th} discrete amplitude level. The redundancy of the discrete amplitude signal is given by [Jayant84, Annex C]:

$$\text{redundancy}_{d.a.s.} \triangleq \log_2 M + \sum_{k=0}^{M-1} p_k \log_2 p_k \quad (\text{bit/symbol})\,. \tag{1.2}$$

It can be shown [Jayant84] that the entropy of a discrete amplitude source with memory is inferior to the entropy of a memoriless discrete amplitude source, which means that in this case, the redundancy given by Equation (1.2) will be underestimated.

A lossless reduction of the redundancy can be achieved by using suitable source models allowing an effective uncorrelation of the source symbols (if the source exhibits memory) and by using a variable length encoding (also referred to as *entropy encoding*), so as to assign short codes to frequently occurring symbols, and longer codes to less frequent symbols. It should be noted, however, that appropriate and practical methods of redundancy reduction related to each one of

[2]We will assume in this Dissertation that sink coding consists in a processing acting upon the signal coming out of the source, using models of the sink. This notion does not exist in the Information Theory literature since it is included in the general concept of source coding.

these two opportunities (*i.e.* inter-sample correlation and non-uniform amplitude distribution), imply a certain encoding delay and system complexity. In general, these system parameters increase for higher coding efficiencies, and may tend to infinity when attempting to reach the theoretical limits. Naturally, a realistic system design and operation impose constraints on the system complexity and delay.

Regarding the non-uniform amplitude distribution, methods of entropy encoding are applicable which include Huffman coding, arithmetic coding and run-length coding. Regarding inter-sample correlation, the applicable methods include predictive coding and subband (which includes transform) coding.

A lossless reduction of the redundancy is not possible for continuous amplitude (discrete-time) source signals. In fact, in this case, a first difficulty arises because the absolute value of the information of a specific amplitude level is infinite since its probability is zero. If the continuous amplitude source is stationary, memoriless and zero-mean, its information contents is given by the *differential entropy*:

$$h(x) = -\int_{-\infty}^{+\infty} p_x(x) \log_2 p_x(x)\, dx \;, \tag{1.3}$$

where $p_x(x)$ is the probability density function of the source. For a given signal variance σ_x^2, the differential entropy is maximum and equal to $\frac{1}{2}\log_2 2\pi e\sigma_x^2$, in the case of Gaussian sources, *i.e.* those for which

$$p_x(x) = \frac{e^{-\frac{x^2}{2\sigma_x^2}}}{\sqrt{2\pi\sigma_x^2}} \;. \tag{1.4}$$

Other probability density functions will lead to lower values of the differential entropy of the source.

It can be shown that if the continuous amplitude source is stationary, zero-mean and has memory[3], the differential entropy is maximized for Gaussian sources becoming [Jayant84]

$$h_G(x) = \frac{1}{2}\log_2 2\pi e\gamma_x^2\sigma_x^2 \;, \quad 0 < \gamma_x^2 \leq 1 \;, \tag{1.5}$$

where γ_x^2 is the *spectral flatness measure* reflecting the non-flat shape of the power spectral density function. This result shows that for a given signal variance, a colored spectrum lowers the entropy of the source (as in the discrete amplitude case), and thus gives rise to a redundancy that may be expressed as:

$$\text{redundancy}_{c.a.s.} \triangleq \frac{1}{2}\log_2 2\pi e\sigma_x^2 - \frac{1}{2}\log_2 2\pi e\gamma_x^2\sigma_x^2 \;, \tag{1.6}$$

[3]We will assume that the concept of memory ignores high order statistics and therefore, redundancy will be associated only to a non-flat power spectrum, *i.e.* we will consider high-order spectra are not structured.

where $\gamma_x^2\sigma_x^2$ is the variance of the white innovations process that generates the source [Jayant84, page 644]. This variance is also known as the minimum one-step-ahead prediction error variance. However, contrary to the discrete amplitude case, the redundancy reduction of a continuous amplitude source can only be achieved in practice by taking advantage of the color (*i.e.* of the non-flatness) of the spectrum of the source signal, and not by taking advantage of the non-uniform probability density function.

The redundancy reduction can be achieved by passing the source signal through a filter modeling the reciprocal of the source spectral shape. This filter is usually referred to as the *whitening filter* and its purpose is the same as in the discrete amplitude case: to uncorrelate[4] the source signal and, as a result, to deliver an approximation to the non-redundant part of the source signal, whose variance will be lower bounded (assuming Gaussian sources) by the variance of the innovations $\gamma_x^2\sigma_x^2$. As a particular case, when the signal spectrum is flat ($\gamma_x^2 = 1$), the variance of the non-redundant part will be equal to the source variance, which means the redundancy is zero.

It is impossible to transmit digitally a signal produced by a continuous amplitude source. However, digital transmission is possible if one tolerates the introduction of an error due to the discretization (quantization) of the source magnitudes, using a limited number of bits per sample. The net effect of this operation is the forced reduction of the entropy of the source. This also represents an increase of the quantity in Equation (1.6), which we may call extended redundancy. This conclusion reflects a result of the Rate-Distortion Theory[5] [Jayant84]: when transmitting the information of memoriless zero-mean Gaussian sources at a rate of R bit/sample, the average reconstruction distortion, according to the mean squared error criterion, will be lower bounded by D:

$$R_G(D) = \frac{1}{2}\log_2 \frac{\sigma_x^2}{D} = \frac{1}{2}\log_2 2\pi e\sigma_x^2 - \frac{1}{2}\log_2 2\pi e D \ , \tag{1.7}$$

where σ_x^2 is the variance of the source and D is the noise variance, both being random and zero-mean Gaussian variables. This result can also be seen as the entropy difference between source and noise. A rate lower than R is needed for the same reconstruction distortion if the source is non-Gaussian or if the source exhibits memory. In fact, the lowest possible rate is given by the Shannon lower bound [Jayant84, page 640]:

$$R(D) = h(x) - \frac{1}{2}\log_2 2\pi e D \ , \tag{1.8}$$

where $h(x)$ is the differential entropy of the source. As we pointed out previously, in the case

[4]This is equivalent to flatten the spectrum of the source, *i.e.* to make it white.

[5]When referring to Rate-Distortion Theory, we are implicitly assuming the usual simplifying conditions, namely stationarity, ergodicity, and Gaussianity, and the unweighted mean squared reconstruction error as the design objective of a coding system.

of discrete amplitude sources, D can be made equal to zero (distortion-free), which means the bit rate in bit/sample equals the entropy of the source, *i.e.* $R(0) = H(x)$. However, as higher values of D (> 0) are tolerated, the rate R decreases correspondingly, which means the extended redundancy increases proportionally. This intuitive conclusion can be easily shown for continuous amplitude sources with memory. In fact, in this case and as results from Equation (1.7), the difference of the source and noise entropy becomes $\frac{1}{2}\log_2 2\pi e\gamma_x^2\sigma_x^2 - \frac{1}{2}\log_2 2\pi eD$, which leads, by Equation (1.6), to the extended redundancy:

$$\text{extended redundancy}_{c.a.s.} = \frac{1}{2}\log_2 2\pi e\sigma_x^2 + \frac{1}{2}\log_2 2\pi eD - \frac{1}{2}\log_2 2\pi e\gamma_x^2\sigma_x^2 \,. \qquad (1.9)$$

This result implies that the upper bound for the allowed distortion is defined by the entropy of the source, in which case no information at all would be conveyed to the decoder.

The substantive conclusions at this point are that redundancy results from two components:

1. a non-uniform probability distribution (discrete magnitude sources) or a non-Gaussian probability density function (continuous amplitude sources) of the source magnitudes,

2. the memory of the source that is reflected in a colored (*i.e.* non-flat) power spectral density of the source signal.

Therefore, the reduction of redundancy is only possible if there is a previous knowledge concerning these aspects or if there is, at least, a way to estimate them reliably. In either case, feasible source coding techniques can be used to obtain a digital representation of the source signal. In the case of discrete amplitude sources, the associated bit rate approaches the source entropy with zero distortion. In the case of continuous amplitudes sources, the implied reconstruction distortion, for a given data rate, approaches the theoretical limits given by the Rate-Distortion Theory.

The compression of high quality audio signals based uniquely on the reduction of redundancy faces several difficulties. In fact, as usually considered for simplicity, the Rate-Distortion Theory assumes signal Gaussianity, stationarity, the mean squared error criterion and establishes, based on these assumptions, theoretical bounds of coding performance, *but* without suggesting specific solutions or methods to approach them, within practical constraints on system delay and complexity [Jayant93, page 1387]. Furthermore, high quality audio signals represent any possible natural and/or synthesized acoustic event whose behavior may deviate strongly from stationarity and Gaussianity. To complicate even further the scenario, a wide source signal diversity may preclude the identification of a sufficiently broad and efficient source production model. This does not apply obviously to specific classes of audio signals for which simple and yet efficient models can be found. This is the case of speech signals whose predictability and spectral smoothness is efficiently modeled by a simple auto-regressive based model.

Another example that could be invoked is the standardized Musical Instrument Digital Interface (MIDI) protocol, which was specifically developed to model musical sounds and events. However, its main feature that contrasts with the focus in this Dissertation, is that MIDI consists in a symbolic representation of music that reflects its semantics. The signal fidelity is not an issue because the quality of the synthesized sounds is totally dependent upon the model lying on the synthesizer, independently of the MIDI generator ("source"). Although very efficient in a compression perspective, considering its natural application scope, it is nevertheless interesting to note that for an improved sound quality at the synthesis stage, MIDI uses plain PCM audio samples as recorded from real musical instruments. For these reasons, we will not make further references to MIDI in this Dissertation.

Besides the inadequacy of high quality audio signals to the usual conditions of the Rate-Distortion Theory, another problem is the maximum achievable coding gain. In fact, taking for example into consideration only the correlation between audio samples, it is known that if the power spectral density function of the source is non-flat, the theoretical coding gain over PCM, for an arbitrarily small reconstruction distortion, is given by the inverse of the spectral flatness measure (γ^2). This theoretical gain is reflected in the redundancy of the source, as can be concluded from (1.6) because this equation can be reduced to $\frac{1}{2}\log_2(1/\gamma^2)$. An implication of this kind of theoretical bounds is that they identify a maximum signal redundancy that may not be sufficient to attain compression factors of a significant practical interest. This is particularly true in the case of lossless compression. In fact, in this case and to our knowledge, compression factors as low as 3:1 have not yet been achieved for Compact Disc quality (discrete amplitude) audio signals, by exploiting both the non-uniform amplitude distribution and the correlation between samples.

Coding gains of undiscutable practical relevance are in the order of 10:1. This ambitious goal can only be achieved if the objective entropy of the source is allowed to be reduced significantly. However, the resulting increase of the reconstruction distortion, considering the usual simplifying conditions of the Rate-Distortion Theory, reflects implicitly the mean squared error criterion, which means the spectrum of the increased distortion is white[6]. As this noise shaping does not match the subjective impression of quality, as we pointed out previously, a small increase of white distortion may lead to an unacceptably poor signal quality.

The practical meaning of this conclusion is that in order to reach ambitious compression

[6]We assume here small distortions and we also consider that special measures are not taken to enforce a specific non-white spectral weighting of the noise.

factors, a substantial increase of the distortion is inevitable, but this extra noise must be shaped according to the properties and tolerances of the human auditory system, in order to minimize its perceptual impact on the signal quality. In other words, for the purpose of compressing high quality audio signals, a new concept (and tool) of entropy is needed that accounts for the characteristics of human perception. This new concept is the *Perceptual Entropy* [Johns.88a] and the associated new coding paradigm is known as *Perceptual Coding* [Jayant93]. As this approach looks predominantly at the sink of the communication chain rather than at the source, it is usually associated to the notion of *sink coding* (although Perceptual Coding does not exclude source coding), as opposed to the notion of *source coding*. An illustrative comparison of the coding gains achieved with source and sink coding is given in Figure 1.2.

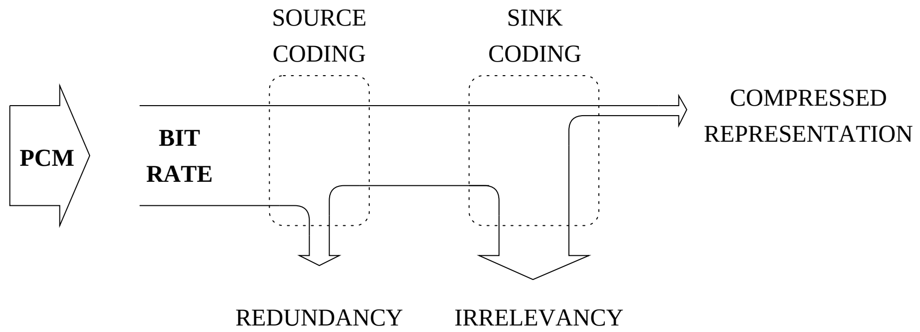

Figure 1.2: Illustration of the relation between the coding gains allowed by source coding (redundancy reduction) and by sink coding (irrelevancy reduction).

In general, the importance of sink coding in communicating audio-visual information in an efficient manner, stems also from the simple fact that stronger assumptions can be made about the sink (human auditory system or vision) rather than about the source.

1.1.3 Perceptual Audio Coding

The reduction of audio information, in a perceptual sense, consists in the exploitation of the properties and tolerances of the human auditory system [Moore89][Zwicker90], namely the frequency sensitivity and selectivity, the sensation of loudness, and masking effects, represented by psychoacoustic models, to reduce the *irrelevant* components of the audio signal, *i.e.* those components that do not contribute to the auditory impression (or sensation) of the associated acoustic stimulus. Thus, these irrelevant components of the audio signal may be reduced at the very beginning of the signal communication chain and the resulting available information capacity (for storage or transmission), may be used instead to achieve a better coding quality of the *relevant* audio components, or to code additional audio signals. This operation is called *irrelevancy* reduction and is *strongly* based on the perceptual phenomenon of masking.

Masking refers to the total or relative inaudibility of one sound component, the *maskee*, due to the presence of another one, the *masker*, and depends on their relations in amplitude, frequency, time [Moore89][Zwicker90] and space [Blauert83]. Masking properties are in general much better characterized in the frequency domain than in the time domain.

In the context of perceptual coding of audio, the masker is the input audio signal, whether involving coherent (tone-like) or incoherent (noise-like) components, and the maskee is the quantization noise. Therefore, the main function of a perceptual coder that dictates its success, is the accurate estimation of the highest possible (signal dependent) quantization noise profile that does not cause a noticeable impairment when actually injected into the original signal. This noise profile, also called *Threshold of Masking* [Johns.88b] or *Just Noticeable Distortion* [Jayant93], must be *optimally* shaped in frequency, time and space (in the case of a stereo or a multichannel combination of correlated audio signals).

1.1.3.1 The Roots of the Paradigm

The perceptual audio coding paradigm found its roots in the research activity related to speech coding[7], in the late seventies. In fact, it was noticed in particular that spectral regions of the speech signal with high energy, typically formant regions, had the ability to hide or to mask higher amounts of quantization noise than spectral regions of low energy. The resulting advantage was that for a given coding bit rate, the higher concentration of noise under the spectral regions of higher energy, allowed the coding of perceptually important low energy spectral regions that would not be coded at all under a purely objective coding criterion. Several published results and developments corroborated these findings and suggested that accounting explicitly for perceptual phenomena was essential to achieve a good subjective quality of the encoded speech signal, particularly when encoding at very low bit rates. As a consequence, instead of an objective distortion measure as the SNR, an alternative design criterion based on a perceptually related subjective distortion measure, has been looked at with an increasing interest, and has been implemented with convincing results. In fact, it has been verified that if used in a predictive or transform coder, a perceptually related subjective distortion measure allows to *hide* (mask) the quantization noise completely under the signal or, if the bit rate is unsufficient, allows to reduce its loudness and thus its perceived disturbance, as much as possible.

The first results obtained in the speech coding context and new technological advances, namely the increasing processing power (and decreasing cost) of DSP processors, inspired later work on wideband speech [Quacke.91] and wideband audio [Brand.87][Johns.88b], by incorpo-

[7] In fact, given the technological and computational limitations of the research activity 20 years ago, digital audio was, in practice, limited to the bandwidth of speech signals.

rating more complex psychoacoustic models to quantize the audio signal, while simultaneously benefiting from several other techniques like vector quantization and entropy coding for an overall improved performance. The intense research activity in this area was also triggered by the perspective of new applications like satellite and terrestrial digital audio broadcast, digital storage in archives allowing remote access, studios, consumer electronic products (*e.g.* the Digital Compact Cassette and the magneto-optical Minidisc), multimedia applications (*e.g.* virtual reality) and audio-visual services (*e.g.* *HDTV*[8]). The feasibility of some of these new scenarios of applications was in turn validated by the possibilities opened by the development and availability of digital communication channels such as those offered by ISDN[9] and ATM[10] based networks.

More specifically, in the area of wideband speech and high quality audio coding, deeper research efforts led to the conclusion that only a reasonably identifiable part of the information that reaches the ear is decisive to create an auditory sensation or impression. We can identify this fraction of the information as a *perceptual entropy* limit and, as a consequence, we may discard the remaining information because it is *irrelevant* in a perceptual sense. If this operation is successfully implemented by a coding scheme, the resulting coding quality can be considered subjectively lossless, which means the coding procedure is transparent since no sensible differences can be detected between the original and the decoded audio signal.

A tractable measure was developed that expressed the information contents of an acoustical stimulus from a perceptual point of view [Johns.88a]. This measure, called Perceptual Entropy (PE), became an interesting instrument to evaluate the minimum information rate needed to code (*i.e.* to properly quantize), in a transparent way, a suitable time/frequency representation of an arbitrary audio signal. This simple definition of PE encompasses the specification of at least two coding tools but also presumes, perhaps paradoxically, ideal coding conditions.

- The two coding tools are:

    1. a time-frequency decomposition of the audio signal by means of an analysis/synthesis filter bank that is representative of the auditory analytical properties,
    2. a psychoacoustic model that is representative of the most important auditory characteristics and processes of masking.

  In fact, a perceptually meaningful time-frequency representation of the audio signal must be available so as to capture the most relevant signal information (in the auditory sense), in order to drive the psychoacoustic model. In turn, this model yields an estimated Thresh-

[8]High Definition Television.
[9]Integrated Services Digital Networks.
[10]Asynchronous Transfer Mode.

old of Masking that characterizes the maximum amount of quantization noise effectively masked by the audio signal in any possible dimension (time, frequency, space).

- The PE limit presumes ideal coding conditions essentially for two reasons.

  1. It is assumed that there exist means to achieve an accurate noise shaping of the maximum permissible quantization noise corresponding to the Threshold of Masking. As this noise shaping is implemented by a practical analysis/synthesis filter bank and a practical quantizer, the resulting limitations should not be ignored. In practice, the analysis/synthesis filter bank may correspond to the same analysis scheme used to obtain the time/frequency decomposition driving the psychoacoustic model.

  2. The PE limit expresses the information concerning only the quantized representation of the audio signal, which is usually referred to as *main information*. As a typical audio coder codes a spectral representation of the audio signal, the main information consists of the quantized spectral coefficients. The information concerning bit allocation and quantization profile is usually referred to as *side information*. Despite the fact that this information is necessary at the decoder to reconstruct the audio signal from the coded representation, it is ignored by the PE measure.

Nevertheless, the usefulness of the PE measure lies in the fact that it establishes a reference for the performance of a coding criterion oriented to the characteristics of perception. Moreover, contrary to the Rate-Distortion Theory, an intrinsic advantage of the PE measure is that it suggests a specific solution to approach the estimated bounds.

Johnston [Johns.88a] proposed a simple procedure to compute the PE. It is based on a short-time discrete time/frequency transformation by means of an uniform filter bank, on a simplified psychoacoustic model, and on an uniform scalar quantizer. As the noise energy associated to a scalar quantizer with a symmetric and uniform probability density function ($= \frac{1}{\Delta}$), over the quantization interval of width Δ, is given by $\frac{\Delta^2}{12}$, a relation is made to the permissible noise energy at the Threshold of Masking predicted by the psychoacoustic model for each spectral coefficient k:

$$\mathrm{THR}(k) = \frac{\Delta_k^2}{12}\,. \tag{1.10}$$

Assuming a real and orthonormal time-frequency decomposition in N spectral coefficients, Δ_k corresponds to the quantization step-size for coefficient $X(k)$. If $X(k)$ is signed, the number of bits needed to represent its quantized version is $\log_2\left[2\,\mathrm{nint}\left(\frac{|X(k)|}{\Delta_k}\right)+1\right]$, where the function nint(.) returns the nearest integer of the argument. The short-term PE is therefore estimated as:

$$\mathrm{PE} = \frac{1}{N}\sum_{k=0}^{N-1}\log_2\left[2\,\mathrm{nint}\left(\frac{|X(k)|}{\sqrt{12\,\mathrm{THR}(k)}}\right)+1\right]. \tag{1.11}$$

The short-term PE estimates obtained by Johnston for a wide variety of audio material ranged from 0.1 to 2.1 bit/sample. Considering that the PCM magnitude resolution (*i.e.* the dynamic range) of high quality audio is at least 16 bit/sample, it can be concluded the PE estimates indicate a compression gain better than 8:1 is achievable using perceptual concepts and tools. This figure is particularly revealing if it is compared with practical gains (in the order of 3:1) achieved by reducing the redundancy only.

In a coding scheme facing bit rate constraints imposed by the operating conditions or the limitations of the transmission channel, the injected quantization noise rarely corresponds exactly to the Threshold of Masking. It will lie below this threshold if the available bit rate is above the PE of the signal. This condition is known as *overcoding*. However, if the available bit rate is below the PE of the signal, the injected quantization noise will lie above the Threshold of Masking and thus the coding process may not be transparent. In this case, usually referred to as *undercoding*, some additional criterion must then be used to shape the additional unmasked quantization noise so as to minimize its loudness.

It should be noted that, in practice, we can only estimate the PE which depends on the accuracy of the underlying psychoacoustic model. However, even if the PE limit could be exactly determined for a given audio signal, it could never be reached by a practical coder because neither any practical analysis/synthesis filter bank would be able to mimic the complex time and frequency signal analysis as performed by the human hear, nor any practical quantization strategy would mirror the complex non-linear behavior of the human auditory system, particularly when we admit audio signals with high dynamic range (about 100 *dB*) and high spectral bandwidth (*e.g.* 20 *kHz* or higher as in the case of very sharp attacks). Therefore, for practical reasons, perceptual audio coders use simplified psychoacoustic models, simple quantization schemes and computationally efficient signal analysis/synthesis filter banks.

1.1.3.2 The Implementation of the Paradigm

The importance of the Perceptual Entropy concept is considerable since it can lead to significant coding gains by reducing the irrelevancy of the signal, even if little is known or can be assumed about the source signal, which is generally the case of an arbitrary high quality audio signal. Furthermore, as the coding gains resulting from reducing the redundancy exhibited by the source are typically much lower, the overall coding gain achieved by a perceptual coder is, for the most part, based on irrelevancy reduction, but also benefits, to a less extent, on the reduction of redundancy. This relative importance is reflected in the implementation of a perceptual coder.

Taking the coding of a single audio channel as a reference, a typical perceptual audio coder can be functionally represented by the diagram depicted in Figure 1.3. Each building block of

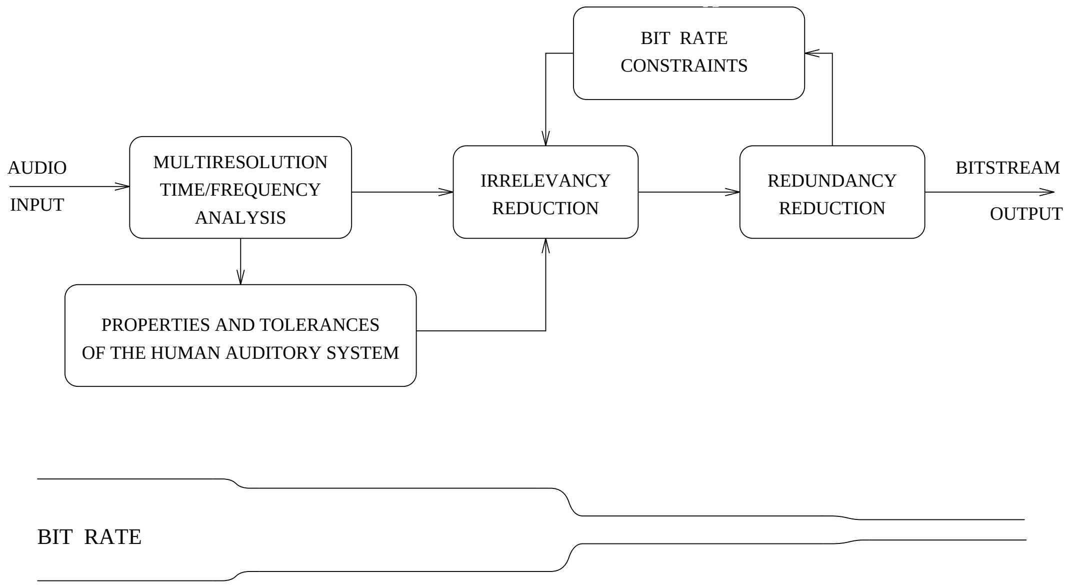

Figure 1.3: Simplified structure of a Perceptual Coder. The opportunities to reduce the bit rate are illustrated by the lower diagram. It shows that irrelevancy reduction represents the most significant contribution to the compression gain.

the diagram identifies a set of coding functions as follows.

- A time-frequency transformation block implements a convenient spectral decomposition of the audio signal. This transformation is necessary to derive important information about the signal such as power spectral density, tonality and harmonicity. Strictly speaking, this building block also reduces the redundancy of the audio signal since the transformed signal is in general much more uncorrelated. The same or an alternative transformation is also necessary to implement an accurate shaping of the quantization noise. The desired properties of this decomposition as well as several practical solutions will be reviewed in Chapter 4.

- Several important characteristics of auditory perception are represented by a psychoacoustic model. This model estimates the profile of the perceptually optimal noise shaping (particularly in frequency) corresponding to the just noticeable noise floor for a given input signal. This model may be extended to the stereophonic or multichannel context, where there is an additional necessity to prevent *unmasking effects* (Chapter 3) and where further gains can be obtained by properly shaping the relative spatial location between signal and quantization noise.

- The irrelevancy reduction is implemented by a quantization operation that makes use of the information delivered by the Threshold of Masking so as to shape the quantization noise as desired and permitted by the actual analysis/synthesis filter bank.

- The quantized spectral data and most of the side information exhibit some residual redundancy. Entropy coding may therefore be used for an overall improved coding gain. Finally, all data is assembled in a bit stream, possibly with some optional error protection information. When constraints are placed on the bit stream such as constant bit rate, an output buffer is needed and a closed loop interaction between the (re)quantization and entropy coding operations is established in order to match coding quality with bit rate limitations.

The decoder is typically simpler than the encoding counterpart because it involves only the demultiplexing and entropy decoding operations, the inverse quantization and the frequency-time transformation. The reference model depicted in Figure 1.3 assumes forward adaptation since this is the most frequently used approach in the implementation of the perceptual audio coding paradigm. However, backwards adaptation is also possible.

1.1.4 The MPEG Standardization Activities

The availability of digital networks and low cost optical storage gave rise, in 1988, to the establishment of an International Technical Group within the International Standardization Organization with the purpose of developing standards for the coded representation of moving pictures, associated audio and their combination. This group, known as the Moving Pictures Experts Group (MPEG, ISO/IEC JTC 1/SC29/WG 11), completed two phases of standards: the MPEG 1 in 1992 and the MPEG 2 in 1994. As far as the audio component in concerned, the first phase has specified 256 $Kbit/s$ as the maximum bit rate to code a pair of audio channels with CD quality. However, the allowed range varies from 32 $Kbit/s$ for a single channel, till 448 $Kbit/s$ for a stereo pair. The MPEG-Audio algorithm is based on a combination of the most efficient features of different algorithms it derives from, and is structured in three layers, each one offering a different compromise between complexity, delay, subjective quality and coding bit rate. For a given bit rate, the highest layer (Layer 3) is the most complex, delivers the best quality but implies a higher system delay. A detailed analysis of the MPEG algorithm can be found in [Noll97][Bosi96][Brand.92b][Ferre.94]. Layer 2 has been adopted for Digital Audio Broadcasting (DAB) [Dehery91].

The second phase of the MPEG-Audio standard has extended the coding of the basic stereo pair to additional audio channels, till a total of five channels, plus a Low Frequency Enhancement Channel (LFE). This *surround sound* combination may be replaced by the coding of seven commentary channels, *e.g.* for multilingual applications.

This phase had an indirect impact on our research activity. In fact, multichannel audio coding was one of the topics included in the first version of the research plan concerning this Dissertation, back in 1992. It happened, however, that several tests within the MPEG-2 Audio standardization activities (1992-1994) revealed clear difficulties in the coding of matrixed multichannel audio while seeking compatibility with the stereophonic format. This fact was later used as the main argument for the start of a special MPEG-2 Audio Non Backwards Compatible (NBC) standardization phase (1994-1997), which produced a standard renamed as Advanced Audio Coding (AAC). Joining to this fact the argument that a multichannel storage and reproduction system was not available to support our research, the multichannel topic was dropped, and the focus was instead concentrated on improving the coding performance of monophonic audio.

The fourth phase of the MPEG activities started in October 1995 (the third phase was assimilated by the second phase) and is currently in progress. The International Standard relative to this phase is expected to be approved by November 1998.

The MPEG-4 addresses a number of requirements resulting from the convergence of different areas: wireless communications, interactive computer applications and audio-visual industries. This phase assumes a new approach to standardization in the sense that it will produce a standard structured on the concept of functionalities for the access, communication and manipulation (which includes composition) of natural and synthesized audio-visual information. Concerning natural audio, the MPEG-4 called for proposals addressing new or improved functionalities including improved coding efficiency, bit rate scalability, robustness in error-prone environments, low delay coding, time-scale modification and pitch modification. It soon became obvious that an interesting audio coder should also offer the possibility to support multiple functionalities (besides compression), by reutilizing, in an efficient manner, its building blocks (designated as "*tools*") already existing for the basic functionality of compression.

As in all the previous standardization phases, the MPEG-4 had a first competitive period of activities from October 1995 till February 1996, followed by a collaborative period which will lead to the final standard. During the competitive period, different proposals addressing one or more functionalities being called for, were independently submitted to the MPEG-4. These were subsequently tested and ranked. The best ranking proposals were then selected to start the collaborative period of activities, by contributing to a Reference Model[11] that evolves to the final standard through a proposal-test-update procedure known as Core Experiment.

1.1.5 Additional Challenges

Even if in 1992 digital storage devices using audio compression such as the MiniDisc and the Digital Compact Cassette were not yet commercial realities, the MPEG-1 Audio was already approved as an International Standard and used, for example, in Compact Disc-Interactive (CD-I) equipment, and in a number of Digital Audio Broadcasting (DAB) pilot test fields. However, more work was recognized to be necessary to improve further the compression efficiency of existing audio coders. This was in fact one of the requirements considered for the phase 4 of the MPEG-Audio standardization activities. Not only became improved coding efficiency the main topic of our research as we pointed out in Sub-section 1.1.1, but also its substantiation in a challenge: the development of a new *Audio Spectral Coder* (ASC), to be described in Chapter 6, so as to participate in the MPEG-4 Audio standardization activities. The outcome of this participation will be described in Section 6.5.

Our initial (November 1995) proposal to the MPEG-4 Audio focussed on the following functionalities: improved coding efficiency, low delay coding, and improved temporal random access. However, as we were not backed by enough funding, we were forced to discontinue our active participation in the MPEG-4 Audio standardization activities beyond the competitive period. Nevertheless, as the MPEG-4 standardization approach implicitly suggested that the same core coder should be able to offer multiple functionalities, we were inspired to face an additional challenge: to exploit the structure of ASC in order to implement additional functionalities. The results of this research will be described in Chapter 7 and comprise three post-processing functions: flexible noise synthesis and modulation, equalization, and time-sale expansion.

1.2 The Thesis in the Dissertation

During the writing of this Dissertation, the terms Thesis and Dissertation appeared frequently as a source of confusion. We are now convinced that these terms have different meanings. We will assume Thesis refers to a set of ideas, premisses or theories whose validity we will try to demonstrate. We will assume further that Dissertation denotes the presentation of the arguments supporting our Thesis, as well as the documentation and explanation of the results of the research conducted to verify and prove our Thesis.

[11] The term Verification Model is also used to designate the Reference Model. It describes the operation of decoders and encoders, and is used to carry out simulations and optimization [Noll97].

In this context, our Thesis is represented by three main ideas.

1. The design of a perceptual audio coder as well as the optimization of its performance is a multidimensional problem that is better addressed with the help of a graphical simulation environment allowing interactivity, and integrating multiple evaluation capabilities: microscopic (*e.g.* revealing the bits spent in each quantized coefficient), macroscopic (*e.g.* indicating the average bit rate per coded frame), objective (*e.g.* Signal-to-Noise Ratio measures), and subjective (*e.g.* allowing to listen to the coded audio).

2. Contrary to a common practice in perceptual coding of audio using techniques in the frequency domain, the quantization noise in each non-uniform (*Bark* related) quantization band should *not* be white but should instead be modified according to the spectral contents of the signal in that band, which implies, for example, that an existing fine harmonic structure in the audio signal must be taken into account in the shaping of the quantization noise, over all coded spectrum.

3. A perceptual coder may offer additional functionalities besides compression such as artificial noise generation and modulation, equalization and time-scale expansion, by taking advantage of the structure of some of its processing building blocks, which may eventually need to be redesigned in order to foster computational efficiencies.

We will not takle each aspect separately because they will be addressed in several chapters of this Dissertation. Instead, we will make references to them in each chapter when appropriate.

This Dissertation includes original contributions that will be presented in Chapters 2, 3, 4, 6, and 7. These contributions reflect several publications, two of them corresponding to Journal papers, and seven of them corresponding to International Conference papers.

1.3 Organization of the Dissertation

This Dissertation is divided into eight chapters which include discussions of coding techniques, processing tools, and system functionalities, some of which may be regarded as post-processing functions. As Perceptual Audio Coding is a rather unknown research area to the Portuguese academic community, several chapters include a reasonable detail whose sole purpose is to attempt to provide more complete information and background to those readers who are not familiar with the area but are nevertheless willing to take a first contact by reading this Dissertation. Readers already familiar with some of the presented material may easily skip it without loss of continuity.

Chapter 1 (this Chapter) discusses the purpose of and the opportunities for compressing high quality digital audio information. A brief explanation is also given regarding the historical context, the objectives, the evolution and the achievements of the research described in this Dissertation.

Chapter 2 presents the concept of coding performance and the problem of developing and optimizing the performance of a new perceptual audio coder. As a possible solution to address these concerns, a specially designed interactive simulation and evaluation environment is proposed, its implementation is described and its functionalities are explained.

Chapter 3 addresses the characterization and modeling of important properties and tolerances of the human auditory system, with an emphasis placed on masking effects and the processes of pitch and timbre perception. A few psychoacoustic models are compared and a simplified model is proposed. Aspects of binaural hearing are also briefly reviewed and the practical limitations of a psychoacoustic model are discussed.

Chapter 4 reviews the theory of multirate filter banks from several perspectives: uniform or non-uniform structures, coding gain, ability to address multiresolution, and efficient implementation allowing extraction of important features of the audio signal. Considering the importance of these aspects in the context of perceptual coding of audio, it is explained why a specific uniform cosine modulated filter bank is of particular interest. Taking advantage of its properties, a new method is presented for the identification of the tonal components of an audio signal. An optimized prototype window is also proposed as a result of a study regarding the multiplicative effects arising from a modification of the spectral coefficients.

Chapter 5 reviews the most common coding approaches and techniques as applied to speech, wideband speech, and high-quality audio signals. In particular, the importance given by each type of coding scheme to the modeling of perceptual phenomena is highlighted. An emphasis is naturally placed on the *state-of-the-art* concerning high-quality audio. Both proprietary and standardized algorithms are compared.

In Chapter 6 we describe in detail the structure of a new perceptual audio coder (ASC) that has been developed. Its performance will be illustrated using the results of its participation in the competitive phase of the MPEG-4 Audio standardization activities.

Chapter 7 shows how it is possible to take advantage of the structure of the ASC analysis/synthesis filter bank so as to implement three additional functionalities: synthesis and modulation of stationary and impulsive noise, equalization in 20 non-uniform bands, and time-scale expansion.

We present in Chapter 8 our final thoughts on the subject of perceptual audio coding and on the main results presented in this Dissertation. We identify possible directions of evolution of our work, as well as new (but related) areas of *applied* research.

Chapter 2

Designing for Maximum Performance

The St. Thomas assertion is well known for emphasizing
the importance of the image: "*seeing is believing*".
Even though the flow of time has whispered such concepts
as *multimedia*, audio is still given its ancestral
secondary role, when it is accompanied by image or video.

Since hearing *and* seeing is understanding,
audio processing is worth a thousand images.

2.1 Introduction

High quality audio coders rely on the efficient reduction of both statistical redundancies and perceptually irrelevant components of the audio signal, in order to reduce significantly the data rate of the coded audio representation. The ratio between the data rate of the PCM representation and that of the coded representation is usually called compression ratio, compression factor or compression gain. This figure is an important parameter of the coding performance.

In this Chapter we address the main dimensions of the performance concept as well as their consideration in the design and optimization of a high quality audio coder.

As the practical importance of some of the performance dimensions is often ignored, the problem of assessing the global coding performance is analyzed in Section 2.2. As a high quality audio coder corresponds quite often to a complex algorithm, its optimization at high compression factors implies the identification and the *joint* optimization of critical processing functions of the algorithm that interact each other. A possible solution to address this problem is proposed and its implementation is described in Section 2.3. It consists in a flexible interactive simulation and evaluation environment allowing both objective and subjective observability capabilities, either in a microscopic, or in a macroscopic perspective. It is argued that this is the fastest and the most suitable approach to controllability. Section 2.4 presents our point of view concerning the complexity issue in the process of developing a new high quality audio coder. Section 2.5 concludes this Chapter.

2.2 The Coding Performance

Contrarily to what is sometimes assumed in the design of an application involving an audio bit rate reduction system, the overall coding performance of a specific coding scheme does not depend only on coding bit rate and subjective quality, although these are probably the most important aspects related to quality of service and communication bandwidth. Other parameters are also very important which may eventually dictate the unsuccess or unsuitability of a particular coding solution. Among these we find coding delay, algorithmic complexity (and resulting power consumption), error robustness, time resolution when accessing coded bit streams, easy bit stream editability and switching. If we consider just the first three as the most representative, then a performance metric can be established that relates five variables through some function $\mathcal{F}$ as follows [Jayant93]:

$$performance = \mathcal{F}(bitrate, quality, delay, complexity, robustness)\ . \tag{2.1}$$

The *robustness* is seen in this context as a measure of how little vulnerable a coding strategy is to random errors, for example, in avoiding the spread of errors inside the same coded frame, or the inter-frame propagation of errors. Error detection and correction is not included here because these functions are efficiently implemented by methods that are exogenous to (and therefore independent from) the coding process, although they may benefit from specific input, such as the identification of the most sensitive information for selective error protection purposes.

The coding performance is proportional to the subjective *quality* and *robustness* but is inversely proportional to the *bitrate*, coding *delay*, and system *complexity*. Several variables may exhibit mutual dependence such as *delay* and *complexity*. For the contour associated to the same *performance* level, there may be several possible trade-offs. For example, subjective *quality* can be traded-off for coding *delay* while maintaining all the remaining parameters. Similarly, *bitrate* and subjective *quality* can be traded-off for *robustness*.

As these considerations let us foresee, it is generally true that for a given coding scheme, a higher coding quality is obtained at the cost of a higher bit rate, coding delay, algorithmic complexity and error vulnerability. Ignoring the *robustness* issue for the moment, the performance of a given coding scheme may be estimated considering three parameters when the encoding-decoding process is transparent, *i.e.* when the quality of the decoded signal is *subjectively* indistinguishable from that of the original signal: the compression factor, the coding delay and the algorithmic complexity.

As we shall see in Section 5.5.2, according to this criterion and using current technology, compression factors in the order of 4:1 or 5:1 are easily reached by algorithms of modest complexity and coding delay (in the order of 10 ms). Higher compression factors in the order of 6:1 to 8:1 are reached by algorithms with increased complexity and longer coding delays (in the order of 50 ms and more). As even higher compression factors are sought for high quality audio signals, such as 12:1, a relaxation of the requirements regarding coding delay and coding/decoding complexity must be allowed. However, this may also turn out to be unsuitable for a number of applications, namely those involving real-time implementation constraints.

In this context, an optimized audio coding algorithm would deliver a maximum transparent compression factor, *i.e.* transparent coding for a maximum reduction of the bit rate, while minimizing the coding delay, the algorithmic complexity, and the error vulnerability. This means that all parameters should be controllable. However, it is common practice in high quality audio coding to presume that the *delay*, the *complexity*, and the *robustness* parameters are frozen or are combined in a structural quantity $f_{struct}()$, so that the parameters effectively controlled are: coding *quality* and *bitrate*. This means Equation (2.1) reduces to the following specific form:

$$performance = \mathcal{F}(bitrate, quality) \times f_{struct}(delay, complexity, robustness)\,. \qquad (2.2)$$

It is also for this reason that the Mean Opinion Score (MOS) tests [Jayant84], for example, relate subjective quality and coding bit rate. When comparing the MOS scores of different coding schemes, it must therefore be insured that they all possess a comparable quantity $f_{struct}()$, otherwise the comparison is unfair and not valid.

In order to improve the operational flexibility of ASC (Chapter 6), and in order to study other possible trade-offs than that implied in Equation (2.2), we admit that the *delay* parameter is also controllable. Therefore, relative to Equation (2.2), the *performance* function applicable to ASC is a more general one as follows:

$$performance = \mathcal{F}(bitrate, quality, delay) \times f_{struct}(complexity, robustness)\,. \qquad (2.3)$$

In fact, we assume the same (frozen) processing complexity for all coding situations, the same error robustness criterion (to limit the spread of random errors to a single coded frame, *i.e.* no inter-frame coding), but we allow the size of the audio coding frame, and hence the coding delay, to be user selectable.

2.3 A Visual Approach to Simulation

Our previous experience in audio coding [Ferre.92][Johns.92][Ferre.95a] led us to infer a very pragmatic rule: in order to optimize the global performance of a coding algorithm, plain system observability must precede system controllability. If this simple principle is ignored, two important consequences may take place:

1. the time devoted to algorithm optimization may easily get out of control;

2. the optimization process might be more the result of an *art* than a scientific procedure [Jayant93].

The main arguments supporting this perspective are presented in the next sub-section. In Subsections 2.3.2 and 2.3.3, we propose and describe the implementation of a visually oriented interactive simulation and evaluation environment. This environment has been specially designed to circumvent the above difficulties, namely to reduce the time needed to develop, to debug, and to optimize a new audio coder.

This environment has been programmed using the paradigms of an object oriented language (C++), and taking advantage of the native graphic and audio features of a Silicon Graphics work station (IrisIndigo).

2.3.1 The Rationale

High quality audio coders correspond, in general, to complex algorithms comprising several interacting processing functions. When coding at high compression ratios, it becomes very difficult and time consuming to optimize the coding performance. In fact, under stressful coding conditions, the coder operation is adversely affected by functional dependencies, interactions and mutual efficiency limitations existing among the different processing functions of the coder. Furthermore, these difficulties depend on the actual audio data. As a result, taking into consideration the dynamics of high quality audio signals, the performance of the encoding algorithm may be particularly poor on sporadic but very critical signal passages.

Optimizing each individual processing function of the coding algorithm is in general synonymous with attaining a local maximum. The iteration of this approach is likely to lead to an endless process without reaching a global extreme (either a global maximum or a global minimum). In fact, the sum or product of local optimizations may not lead to a global system optimization. Moreover, the search for an "average optimal" coding performance becomes an extremely complex task because quite frequently, at least a brief listening test is necessary to assess and validate results.

A possible way to improve the coding performance of a perceptual coder at high compression factors, consists in finding and understanding critical inter-dependencies between processing functions, and trying solutions to jointly optimize their operation and performance, in a global perspective. This obviously calls for exceptional analysis and control capabilities not compatible with traditional background running of the whole encoding and decoding processes.

As a pragmatic approach to overcome the above problems, we postulate that a fast evaluation and control of the coding performance requires a simulation environment with powerful and flexible interactive observability and analysis capabilities. According to this postulate, we have implemented an interactive simulation and evaluation environment in order to assist the development, the test and the optimization of ASC. The main features of this environment are described in [Ferre.95a] and will be presented in the next two sub-sections. It allows microscopic perspectives, *e.g.* by graphically revealing how bits are spent for all components of the information, and by displaying the relevant data associated to the main coding decisions and decoding steps. It also allows macroscopic perspectives, *e.g.* by revealing statistical results and tendencies. In short, all the coding/decoding process may be graphically inspected, in an interactive way, for any desired coding frame and/or sequence of frames. These possibilities are also articulated with listening capabilities which may be used simultaneously with the encoding/decoding process or afterwards.

Figure 2.1 shows how the above opportunities can be organized to schedule a phased development and optimization of a new perceptual audio coder. The objective of this organization is to have available in quasi-real-time an objective evaluation as well as a subjective evaluation of the coding performance, obtaining a fast correlation between results and combined causes. As a consequence, there is a permanent feedback giving valuable insight as to what steps should be followed in order to optimize the performance of the coding system.

2.3.2 Interactive Evaluation Environment

Two interactive analysis tools of particular convenience were implemented:

1. a graphical tool for the simultaneous analysis and comparison of any desired number of Finite Impulse Response (FIR) filters,

2. an audio editor with special analysis features.

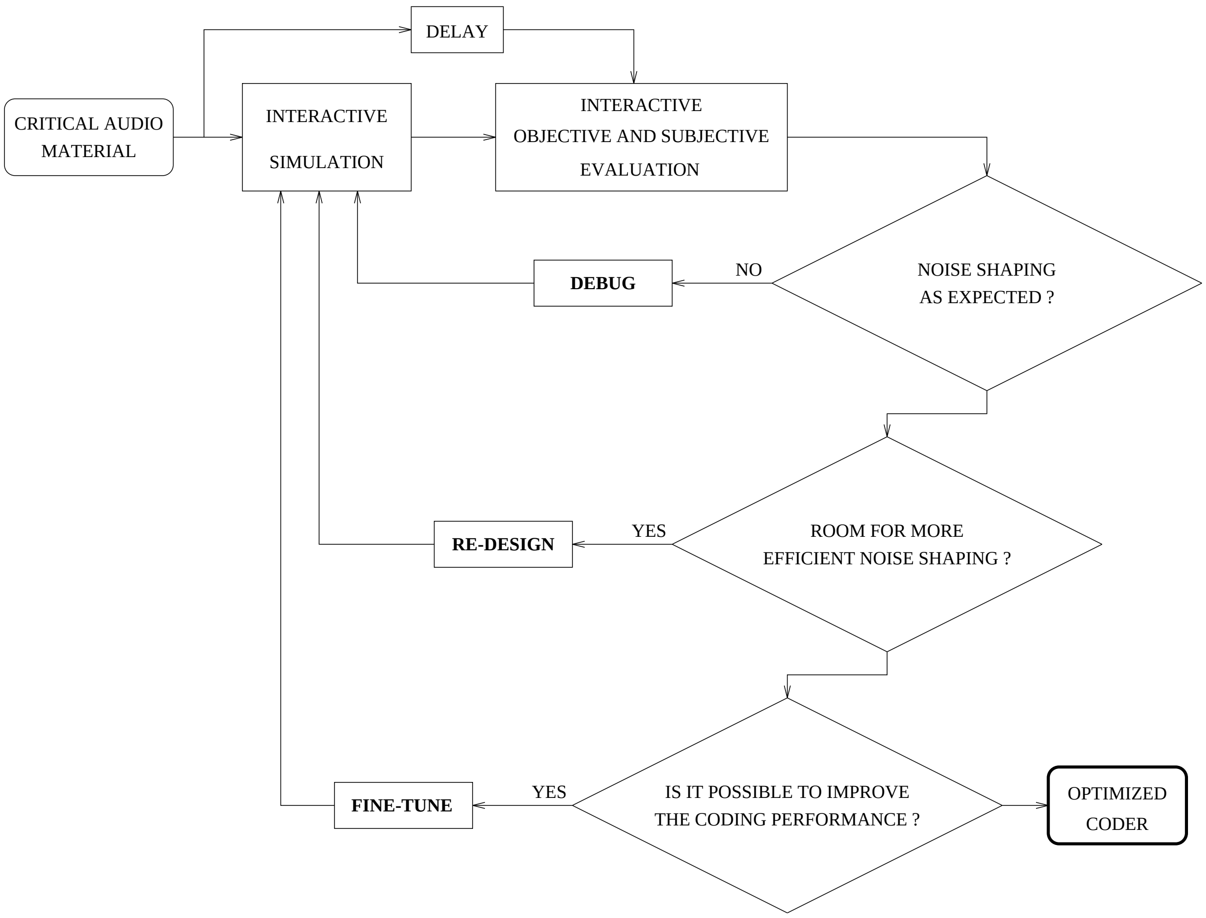

Figure 2.1: A possible approach to the design of a new coding system. The interactive simulation and evaluation environment reduces the time needed to design, to debug and to optimize the coding system.

2.3.2.1 A Tool for the Analysis of FIR Filters

When studying filter banks, we were faced with the need to compare the impulse response, the frequency response, and the group delay response of multiple FIR filters. In order to have more flexibility than that allowed by the available tools, we have developed an interactive environment (FILTRES) that can be invoked with several parameters. Each parameter corresponds to a system filename containing the coefficients of the impulse response of a FIR filter. FILTRES detects and signals the type of each filter: even or odd length, symmetric or anti-symmetric, or neither of these two. In each case, FILTRES uses the appropriate processing modules [Parks87] to compute the frequency response and the group delay response. For each filter, the three types of response are graphically represented using the same color.

A detailed inspection of the displayed information such as zooming is possible using only the mouse. For example, each window (be it the impulse response, the frequency response, or the group delay response) can be freely and independently resized, and is automatically refreshed. *Pop-up* menus can be called to align the frequency responses of the filters in a specified frequency region, by the same maximum or minimum value. In Figure 2.2 we show an example were we compare a 256-point Hanning window, a sine-taper window, and a new window we call Ferreira. The mouse may be used to inspect, within any window, the abscissa and ordinate coordinates

Figure 2.2: Illustration of an interactive tool (FILTRES) allowing the simultaneous analysis of multiple FIR filters. This tool classifies each filter and displays its impulse, frequency and group delay responses.

of a desired location.

2.3.2.2 Audio Time and Spectral Analysis Tool

The audio capabilities have been included in an independent software module (ATSAT) that controls the input and output audio ports of the workstation. Besides the capabilities of a conventional audio editor, it also includes extensive spectral analysis features. These are very important since they permit a fast comparison based on the spectrogram, or on the log-averaged spectrum, between the original and the decoded audio signal, or between the original and the difference signal corresponding to the coding noise[1].

[1] It should be noted however that although all of the coding noise results from the quantization, any aliasing effects introduced by the analysis/synthesis filter bank (Chapter 4) make that the shape of the spectrum of the coding noise, is only an approximation to the shape of the spectrum of the quantization noise injected into the signal.

The frequency resolution and selectivity of the spectral analysis can be easily chosen among several predefined possibilities, on any desired selection of the audio signal. This visual inspection, which may be regarded as a form of an objective evaluation, can be easily complemented by listening either in mono or stereo mode to the same selection of the audio signal.

Figure 2.3 compares the coding at 32 $Kbit/s$ of a harmonic signal (*pitchpipe*) sampled at 44.1 *kHz*, using two different algorithms: ASC (Chapter 6) and the MPEG-1 Audio Layer 3 standard (Section 5.5.2). ATSAT allows both sounds to be heard monaurally or binaurally. Any new

Figure 2.3: Illustration of ATSAT, an audio editor allowing extended spectral analysis features and listening capabilities. This example shows two coded versions of a high quality harmonic audio signal (see text).

desired feature can be easily included in the audio editor and analysis tool, and interactively used together with the existing features. ATSAT can be simultaneously used with the audio encoding algorithm, thus keeping track of all the audio data previously encoded and reconstructed, or can be used after the encoding and decoding session is completed.

2.3.3 Interactive Simulation Environment

Plain observability of the simulated system is achieved by graphically representing the main audio data, as well as the relevant system variables implied in the complete encoding *and* decoding of each audio frame. The whole simulated system may be inspected in an interactive (mouse driven) way. The graphical representation does not add a significant delay to the simulation process because of the direct call to the high performance native graphic subroutines. Furthermore, it is possible to exit the graphics mode at any time and let the simulation run at full background speed.

The system observability is based on a multiwindow representation. Each window is an object with multicolor overlapped data. For the simulation of single channel encoding, five windows are represented as illustrated in Figure 2.4. This example somehow reflects the structure of

Figure 2.4: Illustration of the interactive simulation environment designed to develop and optimize ASC. Each window represents the main audio data and system variables related to the coding and decoding of an audio frame (see text).

the ASC encoder and decoder (Chapter 6). In order to describe the displayed data and the interactive functionalities of the simulation environment, a few concepts are anticipated here without context.

- The Time Window represents the input frame of audio data which is overlapped with 50% of the decoded audio frame, since a TDAC filter bank [Princen87] with 50% overlap is used. The Time window sequence used to filter the input data is also displayed. Typically one long window is used in stationary regions of the signal, and a sequence of short windows is used in non-stationary regions of the signal in order to prevent pre-echo effects.

- The Linear Frequency Window represents important signal features such as the harmonic structure, and the power spectral density of the original signal computed either with a complex transform (Odd-DFT), or with a real transform (TDAC), delivering the coefficients for quantization. The power spectral density of the decoded audio signal before the reconstruction stage is also represented.

- The Non-linear Frequency Window (Bark Window) maps the information displayed on the previous window on a non-linear (Bark) frequency scale. Also displayed are the ideal quantization profile as well as the final quantization profile that complies with bit rate constraints and that is used to quantize the spectral coefficients.

- A Bit Allocation Window shows the distribution of bits by all components of the encoded data, including Huffman codes and CodeBooks.

- Finally, an Information Window displays all relevant values regarding the axis ranges pertaining to each window as well as other local statistical figures such as the Signal-to-Noise-Ratio (SNR) of the decoded audio segment, the bit rate, and the ratio of main to side information.

It is possible within each window to call a pup-up menu and to select, for instance, the detailed analysis of the displayed data in a desired range. In this case, new range values must be typed within a small editor window that appears for this purpose. It is also possible to interactively inspect the exact coordinates of any point within any graphical window. Other existing interactive capabilities include skipping the visual inspection of a desired number of subsequent encoded audio frames, and ignoring visual inspection after a specified segment.

Given the high flexibility of the simulation environment, it is possible to study, for any given bit rate constraint (variable or fixed), the influence of the frequency resolution of the encoding process, the sampling frequency of the audio signal, the criterion used to improve the time resolution of the encoding process in non-stationary regions of the signal, and the shape of the analysis/synthesis time windows satisfying the perfect reconstruction property (Chapter

4). These parameters are set at the start of a coding/decoding session and are therefore fixed during that session.

2.4 Complexity Concerns

In an interesting article *"Speeding DSP Algorithm Design"* in the July 1996 issue of the IEEE *Spectrum* magazine, it is suggested that in order to '*reduce the gap between the algorithm development and implementation phases of a DSP design project [...] it is vital to consider implementation aspects early in the design process*'. Back in 1992, we had this same opinion concerning the way the algorithm complexity issue should be addressed during the algorithm research and development phase. We could have taken either one of two possible approaches:

1. To perform research and algorithm development regardless of the computational complexity of all the processing functions involved, *i.e.* ignoring the complexity parameter in the performance equation. At a later stage, the algorithmic complexity could be reduced trying not to significantly sacrifice other system parameters, namely the coding quality.

2. To start the algorithmic development using processing functions satisfying the basic functionalities and requirements and having reduced complexity. Then, the system operation should be optimized in a global perspective by evaluating all pertinent trade-offs, and by adding new functions or by increasing the computational complexity of specific functions of the algorithm, only if there is statistical evidence of meaningful improvements in global performance.

The first approach is quite common but, in our opinion, it has the potential to lead to sub-optimal results since only one dimension of the performance problem is acted upon at a time.

We decided to follow the last approach since we believe it has the potential to be more exhaustive in finding a better global performance level. For this reason, some of the processing options taken during the research reported in this Dissertation, are likely to appear too simplistic, poorly elaborated or even lacking the elegance of sophistication. In all cases, the implicit argument is that elegance is rather seen as the use of coding solutions of reduced complexity delivering results of statistical significance similar to those obtained with alternative solutions of higher complexity. As will be shown in Chapter 3, this philosophy applies notably to the modeling of the properties and tolerances of the human auditory system.

2.5 Summary

In this Chapter, we have explained our understanding of the challenge inspired by the concept of coding performance, and we have described the way it has been tackled in the context of this Dissertation.

An interactive simulation and evaluation environment has been implemented. Its importance is transversal to all Chapters, and has been instrumental for the research reported in Section 4.5, and in Chapters 6 and 7.

We believe new research results in the area of audio coding and post-processing depend strongly on the insight provided by the simulation environment and the associated analysis tools.

Chapter 3

The Human Auditory System

The rationale underlying the
fascinating analytical and cognitive processes
of the human auditory system is a very simple but dramatic one:
to differentiate is to survive.

3.1 Introduction

The complex analytical capabilities of the human auditory system bear the ability to detect a distant but threatening sound in a deep forest, the ability to evaluate the tuning quality of a single instrument in a performing orchestra, or the ability to identify a single voice in simultaneous conversations in a party. These examples illustrate that the human listening acuity is particularly impressive in environments where many concurrent sounds compete for attention.

To understand how an acoustic stimulus gives rise to an auditory sensation is quite a stimulating challenge given that the most important processing stages of the human auditory system are located in the brain and, therefore, are not directly accessible. This problem is reasonably circumvented by using psychoacoustic methods to characterize both qualitatively and quantitatively, specific aspects of auditory perception. The goal is to create models based on the statistical analysis of the subjective answers provided by trained listeners when subjected to specially prepared acoustic stimulus. However, this characterization and the underlying models are far from complete, precise and objective. In fact, they are based on subjective data derived from simplified listening tests which are not fully representative of the rich diversity of the daily listening experience. Despite these limitations, the information provided by psychoacoustic studies represents a valuable instrument to estimate which cues and which components of an arbitrary audio signal are relevant in driving the auditory attention, ultimately leading to the auditory impression. Such an estimate is extremely important because it paves the way to a compact representation of the audio information allowing an efficient allocation of transmission or storage resources, according to the relevance or irrelevance of each component of an audio signal.

This Chapter reviews a few relevant aspects concerning the structure and behavior of the human auditory system, and discusses some published results of psychoacoustical studies. This information can be used to build a model which is fairly representative of the most obvious properties and tolerances of the human auditory system. We show in Section 3.2 that some of these properties are closely related to the structure of the peripheral region of the human auditory system, lying between the pinna and the auditory nerves. However, those properties are enhanced at higher levels of the human auditory system. To gain a better understanding of these properties, we discuss in Section 3.3 and in Section 3.4 the combined physiological and psychological aspects of auditory perception:

- we introduce important concepts underlying the phenomenon of masking;
- we compare known models of simultaneous masking and propose a new simplified model;
- we address the perceptual importance of the time information existing in an audio signal;
- we address pitch and timbre perception, and emphasize its importance to sound identification, discrimination and quality;
- we briefly review a few aspects of binaural hearing.

Section 3.5 discusses the accuracy of a psychoacoustic model, and Section 3.6 concludes this Chapter.

3.2 The Peripheral Structure

The peripheral region of the human auditory system is responsible for converting audible oscillations of air particles, generally referred to as sound, into neural information which is conveyed to higher levels of the auditory system located in the brain. Besides this transduction function (mechanical energy into electrical pulses), the peripheral region performs also a pre-processing of the acoustic signal, the most notable aspect being a frequency analysis.

The structure of the peripheral region is represented in Figure 3.1. Three different regions are identified: the outer ear, the middle ear and the inner ear.

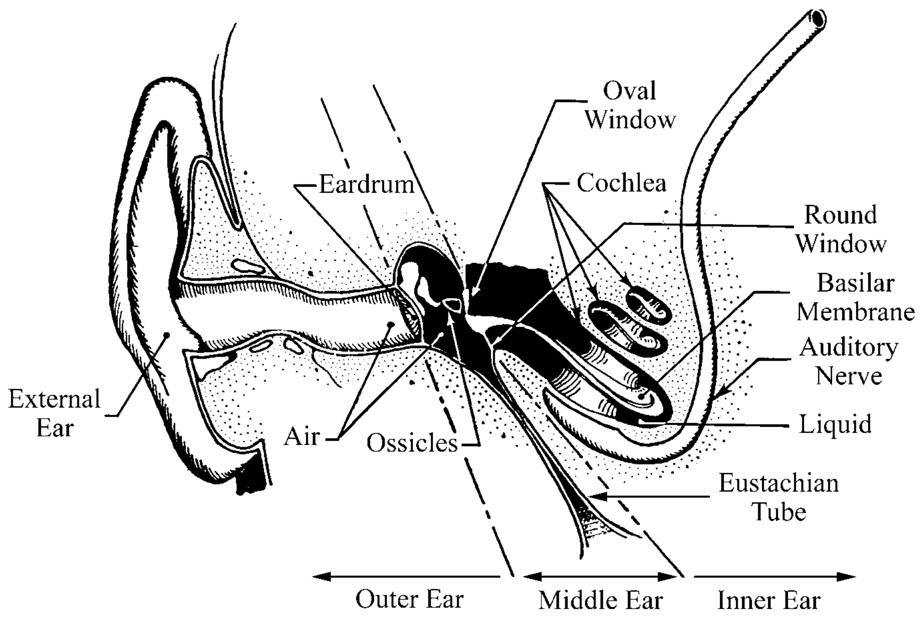

Figure 3.1: A sketch of the peripheral region of the human auditory system (adapted from [Beranek90]).

3.2.1 The Outer Ear

In the outer ear, the pinna collects sound energy and transmits it through the outer ear canal, approximately 2.5 *cm* long, to the ear drum, giving rise to an oscillatory motion. Even though the pinnae (auricles) are fixed in position and lie close to the head, they provide stereophonic hearing allowing us to judge the direction of sound. This discrimination occurs because the intensity, the timbre and the arrival time of sound waves differ as they reach each one of the two ears, as a consequence of their distance to the sound source, and of the shadowing effect of the head. The shape of the auricles helps to differentiate sound coming from behind and sound arriving directly in front [Silver.92].

3.2.2 The Middle Ear

The middle ear transmits the mechanical energy of the ear drum oscillations through a set of three ossicles (the malleus, the incus and the stapes) to the oval window, a membrane at the entrance of the fluid-filled inner ear.

The ossicles match the impedance of the air to that of the fluid inside the cochlea, and also protect the ear against aggressive sound intensities by modifying the transmission gain. This gain is about 15 in normal conditions [Beranek90] (20 according to [Schroe.75]). It results from the mechanical advantage of the ossicles transmission (lever ratio [Schroe.75]) and from the ratio of areas of the ear drum and of the oval window.

The middle ear is filled with air whose pressure is equalized with that of the outside environment through the Eustachian tube. Equalization is naturally done but sometimes must be enforced (*e.g.* when diving) in order to restore the correct resting position of the ear drum so that maximum hearing sensitivity is available at all times.

3.2.3 The Inner Ear

The inner ear corresponds to the cochlea, a bone structure shaped like a snail shell with two and a half turns, whose internal structure is depicted in Figure 3.2. The cochlea is partitioned in three channels or scalae (the scala vestibuli, the scala timpani, and the cochlear duct) filled with two different fluids. These scalae run in parallel from the base to the other extreme or apex of the cochlea (the helicotrema).

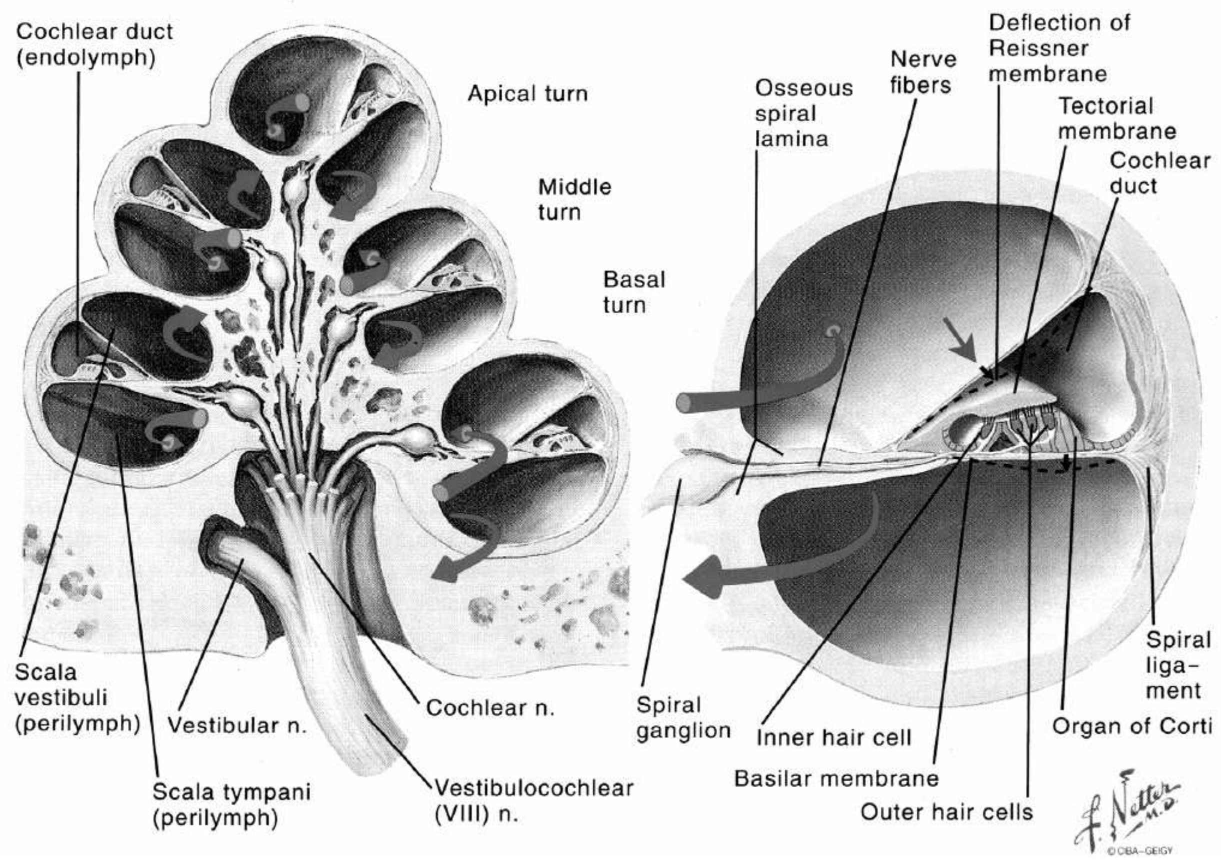

Figure 3.2: A sketch of the internal structure of the cochlea (adapted from [Silver.92]).The arrows illustrate the movement of fluids as a consequence of an inward movement of the stapes footplate at the oval window.

The oval window membrane propagates the mechanical energy from the footplate of the stapes into a fluid called perilymph in the scala vestibuli. The scala vestibuli is connected to the scala timpani at the apex. At the other end of the scala timpani, in the base of the cochlea, there is a round window which acts as a pressure release. The third channel, the cochlear duct, lies in between the above scalae and is separated from the scala vestibuli by the Reissner membrane, and from the scale timpani by the basilar membrane. The cochlear duct is filled with a fluid called endolymph, whose ionic concentration is such that its potential is about 80 mV relative to the perilymph.

The basilar membrane is about 35 mm long, is narrower and thicker at the base, wider and thinner at the apex, and supports the Organ of Corti which contains the sensory cells. These cells, called hair cells, are arranged in one row of inner hair cells on the inner side of the Organ of Corti, and three rows of outer hair cells near the middle of the Organ of Corti. Inner and outer hair cells are structurally different and are differently innervated by afferent[1] as well as efferent[2] nerve fibers, which suggests they are also functionally different. For example, it is a matter of consensus that an active feedback mechanism exists[3], involving in particular the outer hair cells. This mechanism explains the large dynamic range of the auditory system, as well as its sharp frequency selectivity at low stimulus levels [Zwicker90]. However, both types of hair cells are overhung by (but not firmly attached to) a tectorial membrane which modulates the electrochemical activity of each hair cell. This is achieved by forcing a variable bending of the hairs (stereocilia) of each cell according to the relative motion between the basilar membrane and the tectorial membrane, as a result of the hydromechanical oscillations. The measurable results of the electrochemical activity of each hair cell are electrical pulses detectable inside the nerve fibers [Schroe.75]. In the absence of acoustic stimulus, the steady generation of pulses (firing rate) occurs at a rate of about 50 pulses per second. During strong steady-state acoustic stimulus, this firing rate can go up to 150 or more pulses per second. During the onset of a strong stimulation, the firing rate can, for a short time, go up as high as 1000 pulses per second [Schroe.75]. An important aspect is that after the onset of a strong stimulation, the firing rate generated by a hair cell decays strongly for about 1 ms, in order to restore its chemical conditions ("refractory period" [Schroe.75]).

3.2.4 The Transduction Process

An oscillation of the oval window, produced by an oscillation of the ear drum, gives rise to a traveling wave that travels through the perilymph and the endolymph, and along the basilar

[1] Afferent nerve fibers carry information from the cochlea to higher levels of the auditory system.

[2] Efferent nerve fibers carry information from the upper centers of the auditory system to the cochlea.

[3] This also explains the spontaneous otoacoustic emissions (cochlear self-oscillations or *tinnitus*), as well as the evoked otoacoustic emissions (cochlear post-stimulus resonances)

membrane. Given that the dimensions of the cochlear scalae and the physical properties of the basilar membrane vary from the base to the apex of the cochlea, the basilar membrane suffers a maximum excitation at a distance from the base which is related to the frequency of the oscillation. In fact, Von Békési [Békésy60] discovered that the magnitude of the traveling wave gradually increases from the oval window in the direction of the helicotrema, reaches a maximum and then vanishes quite rapidly beyond this maximum. High frequencies cause the membrane to "resonate" near the round window, while low frequencies cause the membrane to "resonate" near the helicotrema. Greenwood [Greenw.90] derived a simple formula giving the position (Z) of the peak, measured in millimeters from the helicotrema, as a function of the frequency in *Hertz*:

$$Z = 7.24 \ln\left(1 + \frac{f}{165}\right) \quad (mm)\,. \tag{3.1}$$

Therefore, the basilar membrane can be seen as a wide-band mechanical filter which is able to decompose a complex sound into a number of frequency components. As a result, a particular group of hair cells and associated nerves are more vigorously excited by a particular frequency than by other frequencies. This peripheral frequency decomposition is only partial because a sharper frequency analysis is performed at higher levels of the human auditory system.

Man has about 40000 hair cells when very young and about 30000 by middle age due to *presbycusis*, the progressive loss of acuity, particularly at high frequencies, that occurs with advancing age. Several types of medicine and behavioral factors, such as exposure to loud sounds, also contribute to accelerate the hearing losses. The information delivery by the living hair cells is conveyed to the brain by nerve fibers that preserve the distribution of frequency that occurs peripherally in the cochlea (tonotopic organization). The dynamic range of the hair cells, as well as of the associated nerve fibers, is about 60 dB. Despite this limitation, man perceives intensity over a range of more than 100 dB. The main reason is that the peripheral pre-processing provides some dynamic compression, which means the peripheral region exhibits a non-linear behavior. This is demonstrated, for example, by the *two tone suppression effect* [Moore89] and by the appearance of combination tones[4] when two pure tones are applied to the outer ear [Schroe.75]. Another important aspect is that at low frequencies, the fibers respond according to the instantaneous phase of the basilar membrane motion, while at high frequencies, above about 3.5 kHz, this phase synchronization no longer appears to exist [Zwicker90] (4.0 kHz to 5.0 kHz according to [Moore89]). We will return to this topic in Section 3.3.5.

While these are all very interesting aspects providing valuable insight into a few properties of the human auditory system, the truth is that this refers only to a pre-processing stage. In

[4] Given an acoustic stimulus consisting of two pure tones whose frequencies are f_1 and f_2 ($f_2 > f_1$), one particularly important combination tone is generated by the auditory system whose frequency is $2f_2 - f_1$ [Moore89].

fact, the most important processing functions, such as spatial perception, take place beyond the cochlea, at higher centers of the human auditory system. Unfortunately, these centers are inaccessible and therefore remain largely unknown. For example, it is known that as the nerve impulses ascend the auditory pathways, the interaction and synchronization between the two ears increase [Silver.92]. But little is known about how this occurs, let alone with what degree.

Psychoacoustics is therefore an effective tool although not entirely exact, in relating input (the acoustic stimulus) to the output (the auditory sensation). This approach allows the derivation of simplified but very useful models, without getting concerned with the immense detail and likely inextricable complexity of the physiological and psychological aspects of the auditory system. This will be our approach throughout the rest of this Chapter.

3.3 Monaural Hearing

3.3.1 The Hearing Area and Loudness

An auditory sensation is evoked by acoustic signals whose physical properties, namely intensity and frequency, make them audible. The hearing area is displayed in Figure 3.3 for monaural listening and represents the frequency and intensity of audible pure tones in free field conditions.

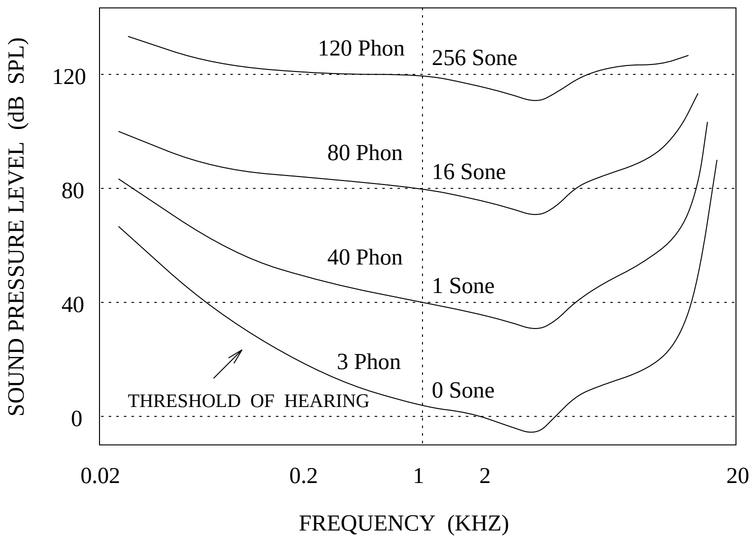

Figure 3.3: A representation inspired by the results of Fletcher and Munson [Beranek90] of equal loudness contours for pure tones. The lower contour corresponds to the threshold of audibility. Zwicker [Zwicker90] argues this contour corresponds to *3 Phon* (and not to *0 Phon*) and *0 Sone* at 1 *kHz*, whereas Moore [Moore89] suggests *2 Phon* and *0.003 Sone*.

The ordinate scale is Sound Pressure Level which is a logarithmic measure of sound intensity in Watt per square meter (W/m^2) relative to 10^{-12} W/m^2. The abscissa scale is *Hertz* in a logarithmic organization to accommodate a frequency range of about three decades. A more convenient abscissa scale is *Bark*, as will be discussed in the next sub-section, in the sense that it represents the natural organization of frequency as perceived by the human auditory system. The bottom curve in Figure 3.3 represents the Absolute Threshold of Hearing, *i.e.* the minimum sound intensity of a pure tone making it just audible. This threshold is extremely important because it represents the reference zero for the calibration of audiometric equipment [ISO96b]. Intensities in the order of 120 dB (but only 90 dB for the region of highest sensitivity around 3.8 kHz) represent the upper limits of a safe hearing experience. Higher intensities originate discomfort, pain, and if the exposition is sufficiently long, may cause irreversible hearing losses.

An intermediate intensity level gives rise to a sensation level or loudness, a measure of perceived intensity whose unit is *"Phon"*. An equal loudness contour corresponds to the physical intensity of tones of different frequencies that sound equally loud. The reference loudness is normally associated to the intensity of a 1 kHz tone. This measure is usually referred to as loudness level. A related measure is loudness (*tout-court*) whose unity is "*sone*" and is said to be more consistent with the perceived (subjective) relation between loudness levels. By definition, 1 *sone* corresponds to the loudness level of a 1 kHz pure tone whose intensity is 40 $dB\ SPL$. For higher intensities doubling the loudness in *sones* corresponds approximately to increase the loudness level by 10 $dB\ SPL$. This reflects Steven's law [Stevens57]: the intensity sensation increases with physical intensity according to a power law. In fact, given the reference intensity $I_0 = 10^{-12}$ W/m^2, the following correspondence can be established for a 1 kHz tone whose intensity is I:

$$Loudness_{Sone} = 2^{\frac{10\log\frac{I}{I_0}-40}{10}} \approx \frac{1}{16}\left(\frac{I}{I_0}\right)^{0.3} . \tag{3.2}$$

This relationship assumes a tone intensity higher than 40 $dB\ SPL$. For uniform noise, the multiplicative constant in Equation (3.2) changes to $\frac{2}{3}$ and the exponent of the power law is adjusted to 0.23 [Zwicker90][Schroe.79].

For intensities lower than 40 $dB\ SPL$, a decrease of the loudness level by 10 $dB\ SPL$ corresponds approximately to a decrease of the loudness in *sones* by a factor of 3. But at very low intensities, the same 10 $dB\ SPL$ decrease in loudness level corresponds to a decrease in loudness by a factor of nearly 20 [Beranek90]. It is important to note that the above considerations hold even if the 1 kHz pure tone is replaced by an equally intense narrow-band noise centered around 1 kHz. This condition implies that the intensity density of the noise must decrease proportionally as the bandwidth increases. There is however a critical bandwidth above which an equally intense band-pass noise sounds louder instead of equally loud. This limit corresponds to the critical bandwidth of human hearing (which amounts to 160 Hz at the center frequency of 1 *kHz*), as will be detailed in the next sub-section.

Although loudness is usually explained as above considering ideal conditions such as frontally-incident plane sound-field and pure tone signals, it depends in general on other variables such as bandwidth, spectral content and duration [Zwicker61]. For example, it has been tested that the loudness of a 2 *kHz* tone burst is constant (independently of the duration to a certain extent) if its duration is higher than about 200 ms but the loudness decreases for smaller durations [Zwicker90].

Two examples of models of loudness have been proposed by Zwicker [Zwicker90] and Moore [Moore97]. Although they follow similar rules and procedures, there are two basic differences:

1. while the model derived by Zwicker computes loudness as a function of the critical band rate (frequency scale in *Bark*), the model derived by Moore performs the calculations on the *Equivalent Rectangular Bandwidth* scale (frequency scale in *ERB*), which differs significantly from the Bark scale only at low frequencies,

2. in Zwicker's model the Absolute Threshold of Hearing is mainly associated to a noise floor internal to the auditory system, to which any external sound intensity is added, whereas Moore's model takes the Absolute Threshold of Hearing essentially as the result of the transmission gain through the outer and middle ear.

A third model of loudness [Schroe.79] simplifies the inclusion of the Absolute Threshold of Hearing by taking it as an absolute lower minimum below which any sound excitation remains inaudible.

As pointed out by Schroeder, Atal and Hall [Schroe.79], the loudness concept is an instrumental one since the whole point in audio coding is to reduce as much as possible (ideally to zero) the loudness of the coding noise.

3.3.2 Critical Bandwidths and Band-Rate

One of the most notable pre-processing functions performed by the peripheral region of the auditory system (cochlea), is the frequency decomposition of a complex sound. This explains, to a great extent, our ability to discriminate different sound components.

The frequency resolution of the auditory system is strongly related to (but not completely determined by) the physiological properties of the basilar membrane and the arrangement of sensory cells (hair cells). In fact, these aspects explain the non-uniform frequency resolution of the auditory spectrum, the lower frequency range of the audible spectrum having much finer frequency resolution than the high frequency region of the audible spectrum. Therefore, a comparison is made to a filter bank with narrow bandwidth filters on the low frequency region, and with large bandwidth filters on the high frequency region. The characterization of these natural filters has been the subject of intense research mainly performed by Fletcher [Fletch.40], who first presented the concept of *critical band filters*, and by Zwicker [Zwicker61][Zwicker90]. Strongly linked to these "natural" filters or critical band filters is the concept of masking which is present in everyday listening experience.

Masking refers to the reduced audibility (*i.e.* partial or total inaudibility) of one sound component (the *maskee*) due to the action of a more prominent one (the *masker*). The amount of masking depends on the relations in amplitude, frequency, time and space [Blauert83][Zwicker90] between masker and maskee. In particular, the *Threshold of Masking* corresponds to the limit case when the masking conditions are such that the maskee is in the threshold between audibility and inaudibility. Concerning time, masking is further characterized as *simultaneous masking*, *backward masking* or *forward masking.*

simultaneous masking Refers to the masking experience when both masker and maskee co-exist temporally. Without further specification, masking refers implicitly to simultaneous masking.

backward masking Occurs when the maskee precedes the masker. This masking effect is associated to differences in the way faint and loud signals are processed by the auditory system, and is reasonably sensible for time differences between the onset of the maskee and that of the masker, as high as 20 ms.

forward masking Occurs when the maskee follows the masker and is usually explained in terms of a persistence of the pattern of neural activity invoked by the masker. Significant forward masking effects are sensible for as long as 200 ms.

Attempts to characterize the critical bands were based essentially on loudness and masking experiments. For example, Fletcher used a band-pass noise masker centered around a sinusoidal maskee. The noise had constant spectral density. As its bandwidth was increased, the intensity of the maskee was adjusted in order to keep it at the Threshold of Masking. The results revealed that the maskee intensity increased as the bandwidth of the masker increased, but only till a critical bandwidth beyond which the maskee intensity was constant. This critical bandwidth was therefore defined as the bandwidth around the center frequency of a signal (maskee), which has effective masking properties. The experiment suggested that the auditory system could be modeled as a bank of band-pass filters, first defined as *critical ratios* [Scharf70] but later redefined as critical bands, whose bandwidths increase for higher center frequencies.

Another experiment similar to the previous one was performed by Zwicker: the loudness of a complex signal with a given bandwidth was compared to the loudness of a equally intense sinusoid lying on its center. The experiment revealed that below the critical bandwidth, the loudness was identical but above that limit, the loudness of the complex signal increased. This can be easily demonstrated for moderate and high intensities [Moore89], using the loudness model presented previously in Equation (3.2).

Other experiments such as the two-tone masking [Zwicker90], the sensitivity to phase [Moore89] and musical consonance [Scharf70], also confirm the critical band phenomenon, in a consistent way. Furthermore, these psychoacoustic results seem to agree with physiological measurements on the basilar membrane which indicate that a critical bandwidth corresponds to a constant distance along the basilar membrane (0.9 *cm* according to Moore [Moore89], 1.5 *cm* according to Schroeder *et al.* [Schroe.79], and 1.3 *cm* according to Zwicker [Zwicker90] and Scharf [Scharf70]), regardless of the center frequency.

An important issue is whether the critical bands are discrete and slightly overlapping (or even non-overlapping), or continuous and strongly overlapping. Although Fletcher first presented a table with 24 discrete critical bands abutted, which has been refined by others [Zwicker61], it appears that critical bands are continuous rather than discrete. Nevertheless, for practical and computational reasons, a discrete organization of critical or sub-critical bands, is usually considered to characterize and model perceptual phenomena. The usual unit of the critical band rate is *Bark*, 1 *Bark* corresponding to the width of one critical band. Figure 3.4 shows the mapping between the two frequency units: *Bark* (non-linear frequency scale) and *Hertz* (linear frequency scale), as assumed in this Dissertation. The bandwidth in *Hertz* at the center of each critical band is also represented. As an approximation, the critical bandwidth is about 100 *Hz* below the center frequency of 500 *Hz*, and above this limit the critical bandwidth is about 20 % of the center frequency.

As presented so far, the critical bandwidths were determined using simple stimuli and evaluation methods, reflecting essentially the contribution of the peripheral auditory region. Therefore, these results do not represent the limitations in frequency resolution, nor the discrimination ability of the human auditory system when subjected to more complex stimuli (*e.g.* the discrimination of partials in complex tones[5]). In fact, there is evidence of a sharpening process following the basilar membrane which results in an enhancement of frequency selectivity [Moore89, page 107]. For example, the just noticeable frequency variation[6] exhibits the same trend as the critical

[5]A complex tone consists in a combination of pure tones, whether related harmonically or not.

[6]This corresponds to a frequency modulation experiment whose results show a just noticeable frequency vari-

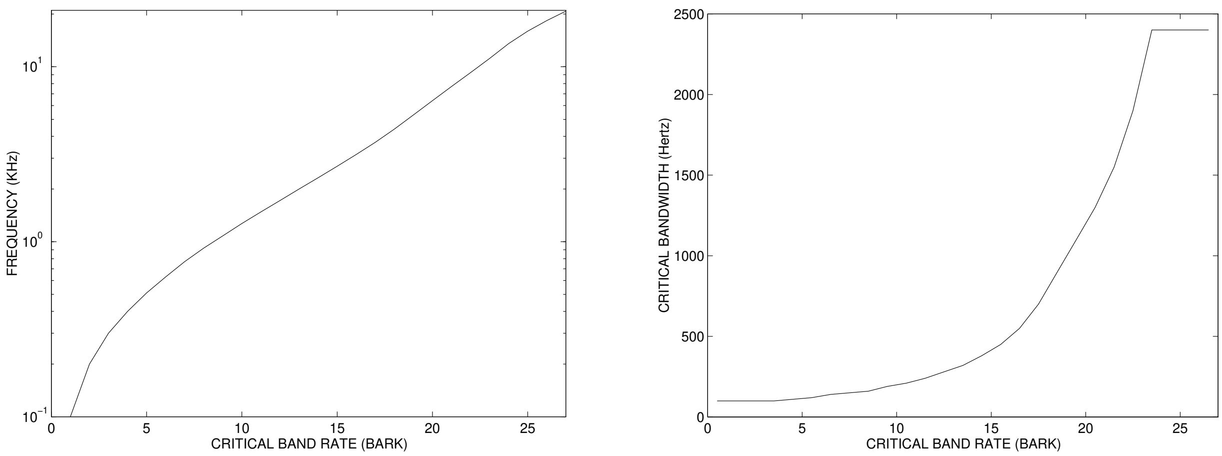

Figure 3.4: Organization of the natural auditory filters. On the left: relation between the linear frequency scale (Hertz) and the critical band rate (Bark). On the right: bandwidth in Hertz of each critical band.

band rate but is about 25 times as precise (*i.e.* smaller) [Zwicker90, page 144].

Furthermore, the influence of the stimuli duration has not been evaluated since durations in excess of 100 ms have generally been utilized producing consistent critical bandwidths estimates. However, specific studies involving a combination of short/long maskers and maskees, indicate that critical bandwidths are wider at short stimulus duration, which suggests that the highest frequency selectivity of the human auditory system requires a certain build-up time (up to about 100 ms) [Scharf70].

3.3.3 Excitation and Masking

Although critical bandwidths represent the effective widths of the natural auditory filters whose simplified models are the critical bands presented previously, it does not imply that the auditory filters have a rectangular shape. A more realistic approach should consider finite slopes on either side on any band-pass auditory filter. This assumption is in fact corroborated by a number of experiments revealing masking effects between signals lying more than 1 $Bark$ apart.

As illustrated in Figure 3.5, the shape of auditory filters may be estimated, for example, by adjusting the intensity of a masker, either a sinusoid or a narrow-band noise, as a function of its center frequency, so as to mask a pure tone whose frequency and intensity are fixed.

A noise masker is usually preferred to avoid the perception of beats between the masker

ations of 3 Hz below the frequency of 500 Hz. However, just noticeable frequency differences (which corresponds to a different experiment) are only about 1 Hz in this range of frequencies.

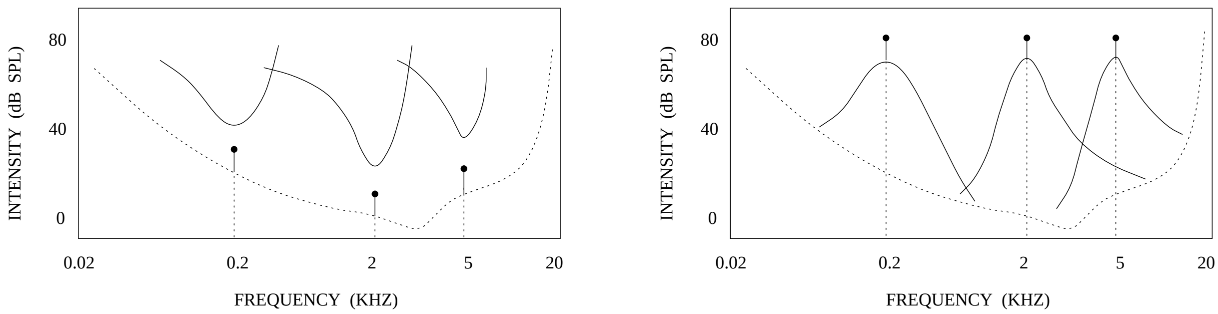

Figure 3.5: Masking experiments revealing the shape of the natural auditory filters for three different center frequencies: 200 *Hz*, 2 *kHz*, and 5 *kHz*. On the left, the solid lines represent the intensity of a masker, as a function of frequency, just masking a tonal maskee of fixed intensity and frequency (filled circles). On the right, the solid lines represent the intensity of a maskee just masked by a fixed frequency and intensity masker (filled circles). The dashed line represents the Absolute Threshold of Hearing.

and the maskee sinusoids, when their frequencies are very close together. The resulting curves are often called *psychoacoustical tunning curves* (PTC) because as the maskee intensity is very low, it is assumed it will produce activity primarily on a single auditory filter, and also because these curves bear an interesting similarity with the *neural tunning curves* [Gelfand90] which represent the intensity of a sinusoidal stimulus, as a function of frequency, that produces a constant activity level on a single auditory nerve filter.

The PTC reveal that the masker reaches a minimum intensity when its frequency equals that of the maskee. On the other hand, it is also evident that the masker masks more effectively (*i.e.* a lower intensity is needed for masking) when its center frequency is lower than that of the maskee, than the opposite situation. These conclusions could also be reached using a different experiment: given a masker of fixed frequency and intensity, the intensity of a sinusoidal maskee is adjusted, as a function of frequency, so as to stay just masked. The resulting curves, also illustrated in Figure 3.5, are called *masking curves* or *masking patterns* (or even *masked audiograms* [Scharf70][Moore89]). Although graphically they look different from the curves of the previous experiment, they are, in fact, symmetric representations and express the same auditory phenomenon: the spread of masking due to the shape of auditory filters.

When plotted on the linear frequency scale (*Hertz*), masking patterns vary substantially according to the center frequency [Zwicker90]. However, if represented on the critical band rate (*Bark*), their shape is reasonably maintained irrespective of the center frequency[7]. This is illustrated in Figure 3.6 using excitation patterns. As we will explain below, excitation patterns have a shape similar to that of masking patterns.

[7]In fact small variations do occur particularly when the center frequency is very low but this can be ignored without incurring in crude approximation errors.

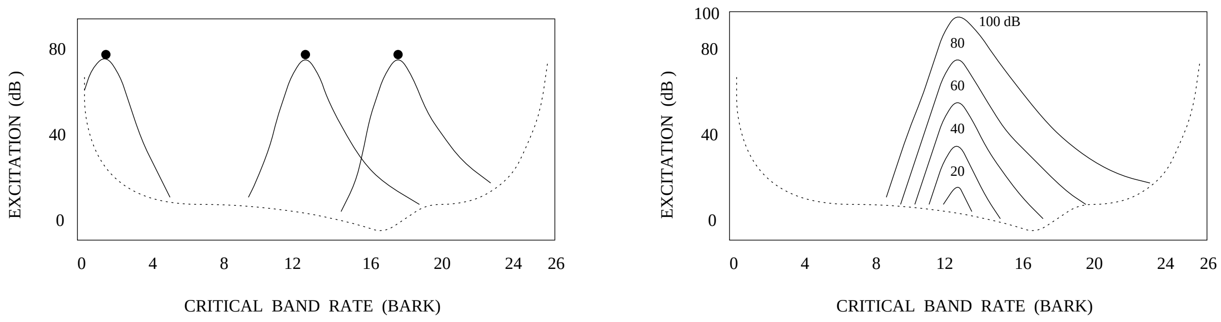

Figure 3.6: Excitation patterns. On the left, solid lines represent the individual excitation patterns generated by tones (filled circles) at 200 *Hz* (2 *Bark*), 2 *kHz* (13 *Bark*), and 5 *kHz* (about 18.5 *Bark*). On the right, solid lines represent the variation of the excitation pattern with the intensity of a 2 *kHz* masker. The dashed lines correspond to the Absolute Threshold of Hearing.

The lower slope of the masking pattern is approximately 27 $dB/Bark$ and the upper slope depends on the masker intensity. Typical values are -20 $dB/Bark$ for very low intensities of the masker and about -5 $dB/Bark$ for high intensities of the masker, as illustrated in Figure 3.6.

Like the PTC, masking patterns reflect the influence of several auditory filters[8] which "capture" different amounts of the masker energy according to their frequency distance to the center frequency of the masker. Therefore, the masking pattern can be seen [Moore89][Zwicker90] as a result of the activity evoked by the masker on each auditory filter, whose distribution is called the excitation pattern. For a single masker, the simplest approach to obtain the excitation pattern consists in a simple upward shift of the masking pattern by the *masking index* (in Decibel). The masking index expresses the fraction of the masker excitation that "overrides" the excitation produced by the maskee; in other words, it expresses the difference (in dBs) between the excitation due to the masker and the excitation due to the maskee. Equivalently, the masking pattern can be obtained from the excitation pattern by a simple downward shift of this pattern by the masking index (in dBs).

3.3.4 Models of Masking

The concept of masking is of central importance in this Dissertation since efficiency in audio coding also implies the effective *hiding* or masking of the maximum amount of quantization noise. The upper limit, in transparent coding, corresponds to the Threshold of Masking. Es-

[8]It should be noted that a particular case of hearing impairment of cochlear origin is related to an abnormal broadening of the auditory filters. This problem, which is accompanied by an elevation of the Absolute Threshold of Hearing (increased hearing losses evident on a audiogram test), leads to difficulties in detectability and sound discrimination (including speech), which are not corrected by conventional hearing aids.

timates of the Threshold of Masking can be obtained using a spectral description of the signal masker, mathematical functions modeling the excitation (or masking) patterns, and models of the masking index for coherent (tone-like) maskers, as well as for incoherent (noise-like) maskers.

A possible computational procedure for the estimation of the Threshold of Masking resulting from an arbitrary acoustic stimulus would involve the following main steps:

1. the power spectral density of the stimulus falling within each critical band is first computed,

2. from the critical band densities, the excitation (or masked) levels are then calculated using the excitation curves with appropriate slopes (which depend on the stimuli levels),

3. finally, the Threshold of Masking for the whole audible spectrum is derived using the masking index, the excitation levels, and the Absolute Threshold of Hearing.

In the case of stationary acoustic stimuli, this simplified procedure encompasses a number of interesting computational advantages:

1. the signal spectral density on the critical band rate can be derived from the density computed with an uniform and computationally efficient filter bank,

2. the critical band rate allows for the uniformization of the excitation functions (Figure 3.6) along the frequency axis, thus reducing the inter-critical band spreading of excitation (or masking) to a simple convolution operation between the appropriate spectral densities and a single model of the excitation function (also known as the *spreading function* [Schroe.79][Johns.88b]),

3. as a corollary of the previous advantage, the effects of the "addition of masking" (as suggested in Figures 3.5 and 3.6) are implicitly addressed by accumulating[9] the spread excitation or masked densities,

4. given that in audio coding the maskee is the quantization noise, which is likely to have incoherent (noise-like) properties, the consideration of the tone or noise-like nature of the masker can be easily taken into account when computing the overall excitation (or masking) pattern, since the shape of the spreading function is practically independent on the nature of the masker as well as of the maskee.

[9] For example, the addition of two equally intense absolute excitation densities produces the desired excitation increase by 3 dB [Zwicker90, page 151] which seems to be a rather conservative [Zwicker90, page 97] but safe approach to the purpose of audio coding.

In spite of these simplifications, special precautions are needed because the above general rules apply only to simultaneous masking and stationary signals. This means that non-stationary occurrences must be monitored and dealt with so as to not violate neither forward nor backward masking requirements. In practice, this implies that overestimated Thresholds of Masking should be avoided because if they are used to quantize a non-stationary signal, the delicate temporal information of the signal would not be preserved [Scharf70].

On the other hand, special care should be taken in the coding of harmonic signals whose timbre depends on the relative magnitude of the harmonicaly related tonal components. If only critical band based densities are considered to estimate the Threshold of Masking, it may happen that several of the tonal components are "averaged" because they fall on the same critical band. This may result in the destruction of the relative tonal magnitudes and consequently, of the original timbre. Therefore, contrary to a common assumption [Veldhu.92, page 88], we do *not* sustain that when shaping the quantization noise on a critical band basis, '*the exact spectral shape of the* [masked] *noise target within the critical band is not important*'.

Several psychoacoustic models have been proposed in the literature to estimate the Thresholds of Masking [Zwicker90][Schroe.79][Johns.88b][Veldhu.92]. Two representative models are described in the Informative Part of the MPEG-1 Audio Standard [ISO93] and are known as Psychoacoustic Model 1 and Psychoacoustic Model 2. We will present briefly these models using a similar notation so as to highlight their origin and main differences.

3.3.4.1 MPEG-1 Audio Psychoacoustic Model 1

A simplified formulation of the Psychoacoustic Model 1 is given by:

$$THR(k) = \theta(k) + \sum_{j=1}^{N_T} P_T(j)\, SF(j,k)\, 10^{-\frac{tm(j)}{10}} + \sum_{j=1}^{N_N} P_N(j)\, SF(j,k)\, 10^{-\frac{nm(j)}{10}} \;, \tag{3.3}$$

where $THR(k)$ represents the absolute power of the Threshold of Masking in band k, $\theta(k)$ represents the Absolute Threshold of Hearing in band k, $P(j)$ represents the absolute value of the signal power in band j, $SF(j,k)$ represents the absolute value (≤ 1.0) of the spreading function in band k when it is centered in band j, $tm(j)$ represents the masking index (in dB and ≥ 0.0) due to a tone masker in band j, $nm(j)$ represents the masking index (in dB and ≥ 0.0) due to a noise masker in band j, and finally, N_T and N_N represent the number of tone maskers and noise maskers respectively. The band indexes (j and k) refer to the division of the audible spectrum in accordance with the critical band rate.

This model performs a preliminary classification and separation of bands according to the coherent or incoherent nature of their signal content. This corresponds to label each band as a tone masker or as a noise masker.

The Psychoacoustic Model 1 expresses the accumulated spreading of the masked spectral power, due to tone and noise maskers. It follows, in a simplified way, the Zwicker model [Zwicker90], namely by considering the Absolute Threshold of Hearing as an additive masked noise internal to the auditory system [Zwicker90, page 100], and by reflecting Zwicker's postulate [Zwicker90, page 99] that masking occurs at any place along the critical band rate scale, if the increment in excitation level, over 1 $Bark$, is less than 1 dB (which corresponds to a masking index for noise maskers of 6 dB at high signal frequencies and around 2 dB at low signal frequencies).

3.3.4.2 MPEG-1 Audio Psychoacoustic Model 2

Instead of a binary classification of bands as coherent or incoherent, the Psychoacoustic Model 2 allows a continuous classification between the two limits, using a tonality[10] index $\alpha(j)$ for each band j. Entirely coherent bands are assigned $\alpha(j) = 1.0$ and entirely incoherent bands are assigned $\alpha(j) = 0.0$. A simplified formulation of the Psychoacoustic Model 2 is given by:

$$THR(k) = \max\left\{\theta(k), 10^{-\frac{\alpha(k)tm(k)+[1-\alpha(k)nm(k)]}{10}} \times \frac{\sum_{j=1}^{N} P(j)SF(j,k)}{\sum_{j=1}^{N} SF(j,k)}\right\}. \tag{3.4}$$

The variables have the same meaning as presented previously for Psychoacoustic Model 1, although in the implementation details they might differ slightly, *e.g.* the spreading function $SF(j,k)$ has a different implementation according to each model.

This model expresses the masking of the accumulated spreading of the weighted signal excitation. It is clearly inspired on the model developed by Schroeder, Atal and Hall for speech signals [Schroe.79]. The Absolute Threshold of Hearing is taken as a reference to the minimum masked power in each band. Rather than spreading masked levels, this model spreads the excitation and then the masked levels are estimated on each band according to the local tonality index. This index is used to interpolate between the masking index due to a tone masker, and the masking index due to a noise masker. The masking index due to a tone masker, for example, is very similar to that initially proposed by Schroeder *et al.* [Schroe.79]: $10 \log tm(x) \simeq 15 + x \quad (dB)$, where x is frequency on the critical band rate ($Bark$).

[10] This is different from musical tonality which refers to the predominance of a single tone or pitch in a musical sequence [Roeder.95].

3.3.4.3 A Simplified Model of Simultaneous Masking

The two psychoacoustic models presented previously for the estimation of the Threshold of Masking, differ quite substantially in their core structure. There is not a clear indication that one is more accurate than the other and we do not find convincing arguments supporting that a more complex model performs better than a simpler one in the context of audio coding. This is certainly an area where there is not a clear border between art and science [Jayant93]. Regarding psychoacoustic modeling, our approach draws on the following pragmatic grounds:

- a psychoacoustic model should address the most evident perceptual aspects of general consensus (*e.g.* the critical band rate, the level and nature of the masker, the spreading function, the temporal relations between masker and maskee), and its sophistication should be constrained when an increase in complexity does not bring a sensible and clear improvement in performance,

- as most published psychoacoustic data has been derived in conditions very different from those characterizing the operation of a psychoacoustic model in an audio coder, it is necessary to validate and to adjust psychoacoustic data and models to the actual coding conditions,

- when deriving information from the audio signal in order to feed the psychoacoustic model, care should be taken to reuse as much as possible processing functions already available in the audio coding algorithm, thus avoiding the proliferation of dedicated modules, as well as the resulting penalty in the global complexity of the audio encoding/decoding algorithm (not to mention the design elegance).

The interactive and graphical capabilities of our simulation environment has made the assessment of prospective psychoacoustic models very easy, as well as their simplification to a form tunned to the actual coding conditions. These are mainly dictated by the nature of the filter bank whose spectral coefficients are quantized according to the output of the psychoacoustic model.

It would be rather tedious, and perhaps uninformative, to describe all tests performed with pure tones and narrow-band noise signals, for example, in order to verify the slopes of the spreading function and in order to find models for the masking indexes associated to noise maskers as well as to tone maskers. We will simply present the core structure of our psychoacoustic model, so as to ease its comparison to the models presented previously.

Our model expresses the accumulated spreading of the masked spectral densities of the signal:

$$THR_D(k) = max\left\{\theta_D(k), \sum_{j=1}^{N} D(j)\, SF(j,k)\, 10^{-\frac{\alpha(j)tm(j)+[1-\alpha(j)nm(j)]}{10}}\right\}. \tag{3.5}$$

Instead of considering the signal power in each band as the models presented previously did, we consider the signal densities $D(j)$ in each band j. This approach eases the modeling of the spreading of masking, particularly in the high frequency region of the audible spectrum where critical bands have larger bandwidths in *Hertz*, and also simplifies the mapping from spectral densities in the linear frequency scale (*Hertz* scale) to densities in the non-linear frequency scale (critical band rate), and *vice-versa*.

The masking indexes are as illustrated in Figure 3.7, and were derived in real audio coding conditions using pure tones, narrow-band noise and broadband audio signals. This method of derivation takes into consideration the combined influence of other models such as the spreading function, other coding parameters such as the frequency resolution on the critical band rate (taken as 1/3 $Bark$), as well as all algorithmic constraints such the precision of the noise shaping allowed by the time/frequency tools of the audio encoder/decoder[11]. The models presented in

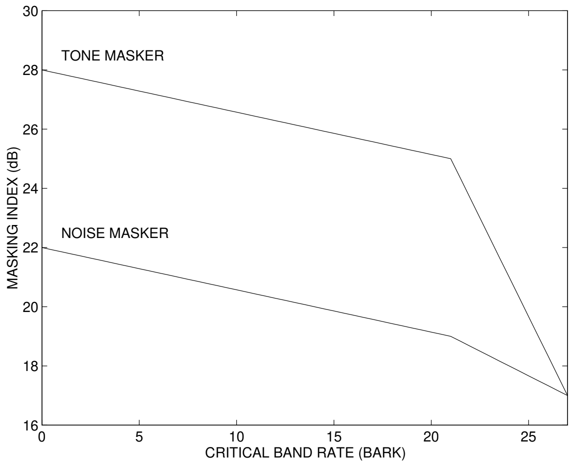

Figure 3.7: Masking indexes for tone maskers (upper model) and for noise maskers (lower model), as derived in real audio coding conditions.

Figure 3.7 agree qualitatively with Hellman's results on the asymmetry of masking between noise and tone [Hellman72]. In fact, he found that a narrow-band noise masks a tone much more effectively than a tone masks a narrow-band noise. This asymmetry has also impelled us to derive the above models: we actually needed a model for noise masking noise (and not noise

[11]In a transform based audio coder, this is mainly related to the structure of the filter bank and the selectivity of the analysis/synthesis time window.

masking tone[12]) which, to our knowledge, was not available in the literature.

As in the Psychoacoustic Model 2, we also take the Absolute Threshold of Hearing as a reference to the minimum masked density in each band. However, except for very low bit rate coding, $\theta_D(k)$ is ignored because it assumes one knows *a priori* the sound pressure level at which the sound will be played. The spreading function is a rather conservative model based simply on two straight lines, the lower slope being 30 $dB/Bark$ and the upper slope being -20 $dB/Bark$, independently of the type and level of the masker. The role of the tonality index, $\alpha(j)$, is the same as in Psychoacoustic Model 2. Its computation will be detailed in Section 4.4.

In summary, the central aspect of the psychoacoustic model we have just presented is that it places the emphasis on the local accumulation of masked levels generated by each signal component, rather than on the local masking of the accumulated excitations generated by each signal component. We believe this approach is more representative of psychoacoustical experiments and data reflecting the complex auditory mechanisms and processes.

3.3.5 Phase Locking and Time Resolution

So far, we have been mainly concerned with simultaneous masking in connection with stationary stimuli. We have assumed essentially the association between the frequency content of the stimulus and the place of maximum excitation on the basilar membrane. On a macroscopic level, this association is reflected by the notion of critical band excitation and masked densities.

There is evidence however, even at the microscopic level, that time information is also used to code stationary acoustic information. In fact, the firing rates of nerve fibers (Section 3.2.3) exhibit periodic patterns whose maxima and minima appear to be related to a multiple integer of the period of a tonal stimulus, as if the firing patterns were "phase locked" to the tonal stimulus waveform [Gelfand90]. This "phase locking" phenomenon decreases for frequencies higher than 2 *kHz*, and disappears completely above 5 *kHz* [Moore89]. This fact is explicitly related in the literature to aspects of pitch and timbre perception. For example, it appears that the musical interval of an octave is only clearly perceived when both lower and higher tones are below 5 *kHz*.

Also interesting is the fact that in response to short non-stationary signals such as clicks, the decaying firing patterns of nerve fibers also exhibit a period related to the frequency (in the click stimulus) to which the fiber responds[13]. This fact may eventually be related to interesting

[12] However, we could have taken published models of noise masking tone as a conservative replacement for a noise masking noise model, but then we would need to check its performance.

[13] Although evidence of these facts has been obtained through direct measurements on the cochlea of certain animals such as cats, monkeys and pigs, the structural and functional similarities between the cochlea of these animals and the human cochlea, is generally accepted and assumed.

aspects of pitch and timbre perception associated to synthetic signals such as square waves, pulsating noise, or simply filtered noise [Zwicker90].

Several other experiments have indicated that the latency in the response of each nerve fiber decreases as the tonotopic center frequency of the nerve fiber increases. This is probably related to the time the traveling wave takes (which is higher for lower frequencies) to reach the point of maximum displacement in the basilar membrane. Latencies in the order of 1 ms to 3 ms are suggested by several experiments [Gelfand90]. Although we could not find a direct relation in the literature, this same range of values is also associated to the monaural time resolution of the human auditory system [Moore89].

It is well established that time plays an important role in the perception of both stationary and non-stationary signals. For example, it is known that the recognition of musical instruments depends quite strongly on the onset transient and on the temporal structure of the sound envelope [Moore89]. With this respect, we performed a very simple experiment[14] using a short sound of castanets that is displayed in the upper part of Figure 3.8.

We replaced progressively the tail of this sound by zero valued audio samples, towards the origin of its onset, as long as we could recognize the resulting sound as a castanets sound. We concluded that below 3.7 ms after the onset, which corresponds to the signal displayed in the lower part of Figure 3.8, we could not recognize the sound which sounded more like a scratch or a click. Furthermore, we concluded that this segment duration was fairly independent on its position as long as it started on a zero crossing region. In fact, despite the obvious reduction in loudness, a 3.7 ms segment was perfectly recognizable even when it started 7 ms after the onset of the original signal. We also concluded that replacing about 1 ms of the transient at the beginning of the original signal by zero valued audio samples (so as to start and stop on zero crossing regions), had no sensible effect on the quality of the signal. This unexpected finding is somehow consistent with the conclusions of the previous experiment.

We have also concluded that if a segment of about 1 ms taken from the low level tail of the original signal (the short segment from 8 ms till 9 ms in Figure 3.8) is appended just before the onset of the original signal, then a loss in sharpness is clearly perceived. This seems to indicate that the onset of a signal contains delicate temporal information, and that *pre-echoes* are more likely to modify or reduce its quality than *post-echoes*. Although this could be explained in the light of the backward and forward masking effects (Section 3.3.2), a connection of this aspect with the "precedence effect" may also be plausible [Moore89].

[14] Only one trained subject using headphones was involved.

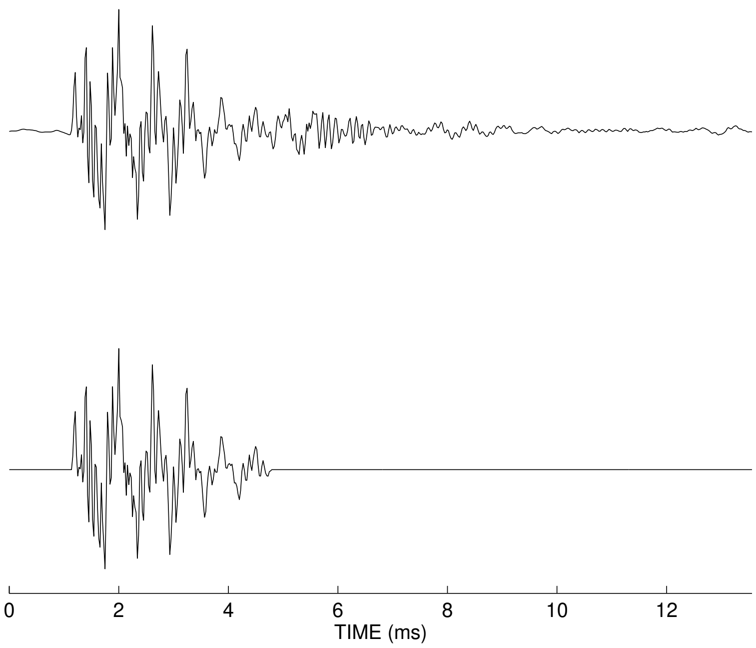

Figure 3.8: The sound of castanets shown on top, was time windowed to the lowest segment (bottom) that maintains sound recognizability. Despite a change in loudness, this 3.7 ms window could even be shifted to the right as much as 7 ms into the signal, that recognizability was not affected, provided that the borders of the time window occur in zero crossing regions.

3.3.6 The Perception of Pitch and Timbre

Many naturally occurring sounds including human speech and those produced by musical instruments, are characterized by a well defined structure of pure tones. Most frequently, this structure corresponds to tones whose frequencies are integer multiples of a fundamental frequency. The pitch of the harmonic complex tone is mainly defined by the fundamental frequency although the sensation of pitch is not independent of the loudness of the sound stimulus. In fact, several tests indicate that when the intensity increases, the pitch of a low frequency tone goes down, but the pitch of a high frequency tone goes up. The frontier seems to be around 2 kHz [Hartma.96].

The combination of the harmonic tones defines the pitch and the timbre of the sound. Timbre reflects the quality of the sound allowing to distinguish among sounds from different instruments, even if their pitch and loudness is the same [Roeder.95]. Timbre depends mainly on the relative magnitudes of the harmonics [Moore89][Roeder.95] and, as such, results from the auditory processing of a complex stimulus mapped on several critical bands.

Some of the harmonics may fall individually within separate critical bands, *i.e.* some harmonics are *resolved* by the peripheral auditory system. On the other hand, groups of harmonics may fall within the same critical band, *i.e.* they are *unresolved* by the peripheral auditory system. This is easily understood by recalling that the critical band partition of the audible spectrum is non-linear with high frequency bands having larger bandwidths [Zwicker61].

While the peripheral auditory system seems to convert, to filter, to decompose, and to code acoustic information; the central auditory system seems to reassemble, to integrate, to interpret, and even to synthesize acoustic data, delivering ultimately the acoustic impression. In fact, although it is generally accepted that the frequency resolution of the human auditory system is related[15] to the critical band rate (Bark rate) [Moore89], several auditory sensations, like pitch and timbre, reveal the existence of enhanced processing functions, probably at higher levels of the human auditory system, with the ability to relate unresolved harmonics. An interesting evidence of this ability is demonstrated by the well known "residue pitch" experiment [Moore89][Zwicker90].

This experiment shows basically that when the lower harmonics (and, in the limit, all resolvable harmonics) are removed from a complex tone, a sensation of the pitch corresponding to the non-existing fundamental still exists. This seems to indicate that the auditory system "synthesizes" the pitch through a clever pattern matching process. These findings have been observed for pitches as high as 5 kHz. Above this frequency the neural synchrony no longer appears to operate [Moore89, page 164] which might explain the absence of the "virtual" pitch.

Other experimental data related to the perception of pitch and timbre reveals that [Hartma.96]:

- if one tone component of a harmonic complex (speech or instrument sound) is changed in some way, either by altering its relative phase or its level, it changes significantly the quality of that complex as a whole,
- if one tone component of a harmonic complex lying above the fundamental is suppressed, the listener hears two pitches: one corresponds to the harmonic complex and the other corresponds to the missing harmonic.

Other convincing examples could be given as the case (well known from our daily experience but poorly understood) of the capability of the auditory neural system to discriminate the timbre of two simultaneously sounding complex tones [Roeder.95], even if partly overlapped. The overall conclusion is that the perception of pitch and timbre plays a central role in defining individual objects in the acoustic world and in separating them from other objects.

The enhanced analytical properties of the human auditory system concerning pitch and timbre perception are not well addressed by masking experiments on which most perceptual

[15] Recent theories, however, support that below 5 kHz the just detectable change in frequency is lower than predicted by the critical bands, and is higher than predicted above that limit.

models are based. For example, to our knowledge, there are no results studying the specific case of critical band masking of unresolved harmonics belonging to a complex auditory stimulus. In fact, most psychoacoustic results are drawn on the basis of very simple stimuli or stimulus attributes. This is the case of well known results of masking of tones by noise and masking of noise by tones [Hellman72][Tobias72]. Since masking is strongly linked to the concept of critical bands (critical band masking and inter-critical band masking), problems could arise because most models of masking assume a relative independence among the excitation or masking levels on a critical band basis. Given a harmonic tone, these models might predict that partials (tone components) in one band are completely masked by the partials in an adjacent band, therefore ignoring their contribution to the global pitch or timbre, which violates the conclusions presented previously for a number of experiments.

In support of this perspective, our previous experience in the area of wideband audio coding shows that most audio signals exhibit quasi-stationary regions with a well defined harmonic structure. In most cases, these regions represent the majority of the signal duration. Musical and speech signals are good examples.

We used a few monophonic test sounds (44.1 kHz sampling frequency and 16 bit/sample) to make a statistical analysis of the regions exhibiting a harmonic character. We describe next their type and duration (surrounded by square brackets):

1. *harpsichord*[6.0 sec.] –consists in a sequence of ten ascending musical notes (harpsichord solo), each note lasting about 500 ms between its onset and the transition to the next note,

2. *female vocal*[5.2 sec.] –consists in a vocal excerpt of a well-known song from Suzane Vega, which is limited to the following sentence: *"It is always nice to see you says the man behind the counter"*,

3. *castanets*[2.3 sec.] –consists in a castanets solo containing 15 occurrences identical to the one displayed in Figure 3.8; the maximum distance between occurrences is about 250 ms and the minimum distance is about 100 ms,

4. *pitchpipe*[6.0 ms] is a pitchpipe solo consisting of three different musical notes separated from each other by about 300 ms, each note lasting about 1.6 sec.

For each one of these sound items, we determined the frequency of occurrence of short segments in which a number (higher than two) of tonal components harmonically related could be detected. The analysis is based on the spectral decomposition of short segments with 1024 audio samples, and using 50 % overlap. The details of the algorithm will be presented in Section 6.2.4.

The results are summarized in Figure 3.9. For each sound item, the segments having more than 12 tonal components harmonically related are condensed in a single result (bar). The maximum number of harmonics found is also indicated.

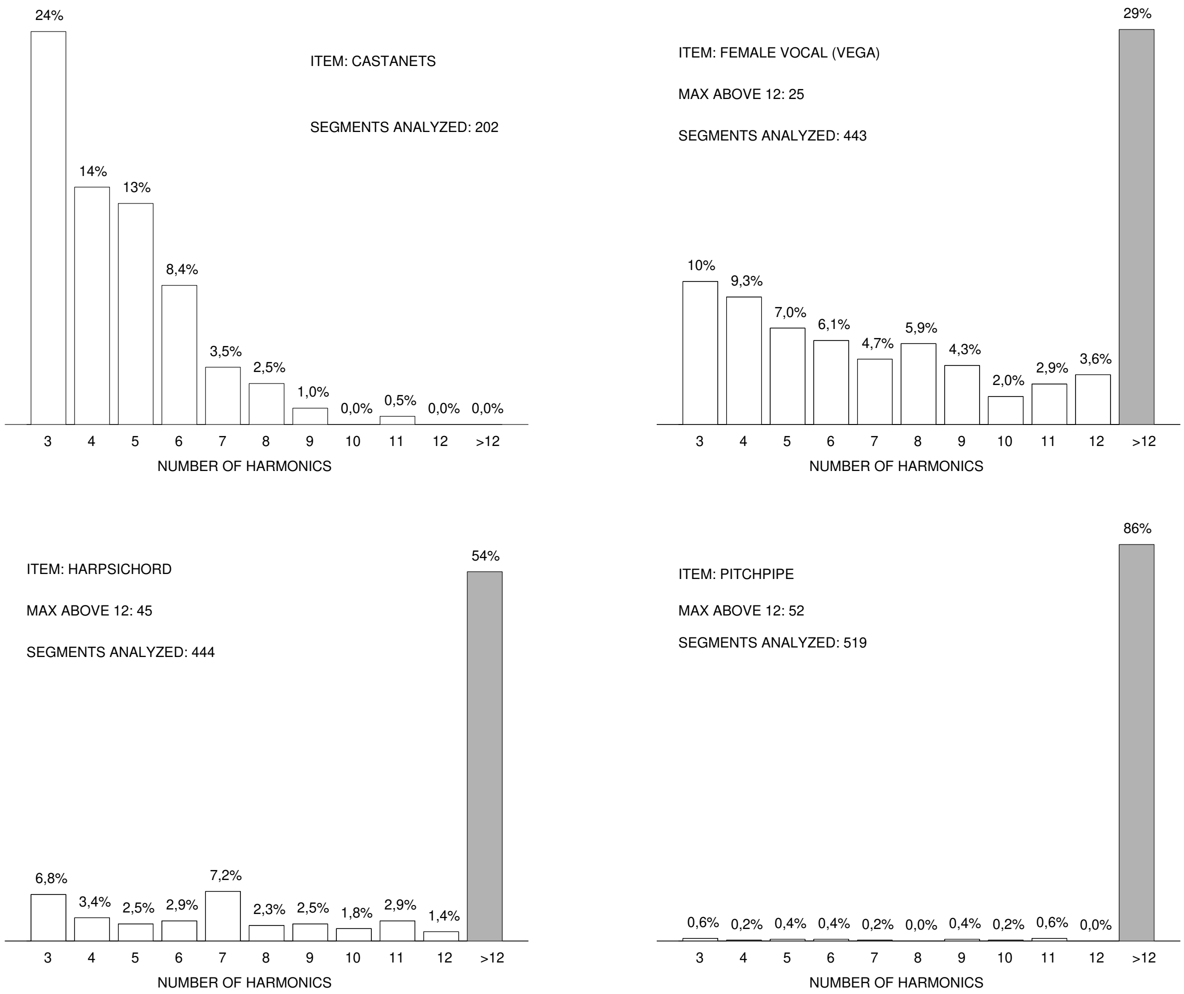

Figure 3.9: Harmonic analysis of different sound items. Each bar represents the relative occurrence of short audio segments where a specific number of tones harmonically related were detected. The upper left results characterize a strongly non-stationary signal. The upper right, the lower left and lower right results characterize signals with an increasing predominance of a well defined harmonic structure.

Castanets is essentially a non-stationary sound item. Despite this, a significant number of harmonic segments was found to have up to 6 harmonics. A more detailed analysis of this strange result revealed that the harmonic segments were captured during the "stationary" regions between the strong castanet hits, where standing waves due to echoes are evident.

Harpsichord and pitchpipe are eminently harmonic sound items since the vast majority of their segments exhibited more than 12 harmonics. This tendency is particularly evident in pitchpipe than it is in harpsichord because the latter test item has more musical note transitions than the former. In either case, the maximum number of harmonics is quite surprising.

The female vocal item reveals that 85 % (sum of all percentages) of all segments exhibit at least 3 harmonics and about 30 % exhibit more than 12 harmonics. A vocal signal differs significantly from speech in the sense that singing a sentence leads to a temporal expansion, in particular, of voiced sounds (which have in general a clear harmonic structure). Despite this, it can be concluded that harmonic sounds predominate[16] over non-harmonic sounds in the active region of vocals and speech. Moreover, the voiced sounds in speech are perceptually more important.

In summary, these results suggest that the harmonic structure is a very important feature in most audio signals, defining their pitch, timbre and overall quality. A model of masking should therefore address this issue considering the signal structure on the linear frequency scale and over a range encompassing several critical bands.

We believe simplified perceptual models currently used by most perceptual audio coders can not accurately estimate the masking pattern of a complex harmonic tone. This is particularly true in the case of unresolved harmonic components since these models do not take into account the global harmonic structure of the audio signal, which contradicts the evidence of the behavior of the human auditory system on this matter. In Chapter 6 we will propose a solution to overcome this problem.

3.4 Binaural Hearing

Most of the material presented in this section was based on two books authored by Zwicker [Zwicker90] and Moore [Moore89]. An in-depth study of binaural hearing and spatialization would necessarily imply the reference to more specialized texts such as the classical book from Blauert [Blauert83], and Begault's book on 3-D sound [Begault94].

3.4.1 Sound Differentiation with the two Ears

So far we have characterized the physiological structure and the psychophysical behavior of the human auditory system in response to *diotic* stimulation (*i.e.* identical acoustic signals presented to the two ears). We made a critical analysis of both established and known psychoacoustics experiments and results, as well as models thereof. Our main motivation results from the fact that we are particularly concerned with monophonic audio coding and therefore, monaural hearing

[16]We are considering a typical European language such as English, French or Portuguese.

is of particular relevance. In spite of this, we feel that we should also present here, although briefly, a few aspects of binaural perception in order provide some insight into suitable and possible ways to complete or to simplify perceptual models. In the past, we have investigated the application of aspects of binaural perception to stereophonic audio coding [Johns.92][Ferre.92].

Binaural hearing corresponds to the natural hearing mode in which different acoustic signals are presented to the two ears (*dichotic* stimulation), allowing for localization and differentiation of sound sources. This is a consequence of the fact that the sound reaching the two ears exhibits several differences which are used as cues by the binaural auditory system to make judgements about the surrounding space, the nature of the acoustic source, and its position relative to the listener. We will first address the physical and physiological aspects that elicit those differences. Then, we will describe established psychoacoustic results revealing a number of properties of the binaural auditory processing. Finally, we will explain why the monaural masking models addressed in Section 3.3.4, if not used strictly in monaural hearing conditions, should consider the release from masking due to binaural hearing.

3.4.2 Physical and Physiological Implications

The shortest distance between the two ears in about 23 cm. Considering the velocity of the propagation of sound in the air, a sound wavefront reaches the two ears with a maximum time delay of about 670 μs, which corresponds to the period of 1500 Hz.

Sound frequencies much lower than 1500 Hz have a wavelength higher than the diameter of the head. In this case, the effect of the head and shoulders is a mere diffraction which does not cause relevant intensity changes on the sound reaching the two ears. On the contrary, sound frequencies much higher than 1500 Hz have wavelengths lower than the diameter of the head and in consequence, are reflected when hitting the head and shoulders. This shadowing effect has a strong impact in the intensity of the sound reaching the two ears, which may differ as much as 20 dB.

It may be concluded from the above considerations that the *interaural time difference* (ITD) and the *interaural intensity difference* (IID) are two effective cues allowing the binaural auditory system to judge the direction of a sinusoidal sound source, according to its frequency. Several experiments have in fact indicated that the IID is the most effective cue for sounds whose frequency is much higher than 1.5 kHz, and that the ITD is the most effective cue for sounds whose frequency is much lower than 1.5 kHz. In the vicinity of 1.5 kHz there are ambiguities in the direction judgement if the head is not allowed to move freely.

Sounds more realistic than simple sinusoids, whether stationary or non-stationary, have a common feature: the onset (and obviously, but less important, the offset). Several experimental results suggest that the binaural auditory system looks particularly at the onset transient to derive as much information as possible concerning the direction and the type of the sound source. The ITD seems to play a crucial role and several tests with clicks have revealed, in addition, that the binaural time resolution is incredibly fine, in the order of 10 μs (as we have seen in Section 3.3.5, the monaural time resolution is about $1 - 2$ ms).

When complex acoustic signals contain only high frequency components, the auditory localization mechanism still responds by looking at the envelopes of these signals, and by comparing their relative timings. Tests with trains of clicks also indicate that only the onset of the click train (first click) is effectively processed by the binaural auditory system to derive localization information. Subsequent clicks of the train just induce a certain adaptation not invoking the full binaural acuity.

Tests with clicks and using headphones[17] also revealed it is possible to trade ITD for IIS. Specifically, these tests demonstrated that the lateralized acoustic image created by a relative time shift of the clicks presented to one of the ears, could be completely shifted back to the initial position by increasing the intensity of the clicks presented to the opposite ear. Although several theories have attempted to explain this phenomenon (for example using the ITD in the firings of phase locked neural fibers, or using the latency of neural responses to impulsive signals as a function of the IID), it seems that in normal conditions both aspects cooperate rather than compete in the localization task.

As mentioned previously, head movements can resolve ambiguity in localization by weighting selectively ITD and IID cues. For complex sounds, another aspect in head movement provides additional cues (particularly in the vertical localization): it is related to the spectral filtering due essentially to the shape of the pinnae (Section 3.2.1, Figure 3.1). In fact, the head and the pinnae act together as a direction dependent filter: some frequency bands are magnified while others are attenuated according to the direction of the sound incidence. This spectral filtering is particularly evident for frequencies above 6 kHz given their short wavelength when compared to the dimensions of the pinna.

This spectral filtering also raises two particularly important issues. The first is that the phenomenon is present even in monaural hearing which reveals, as a consequence, that the monaural listening mode has some capability of space perception. The second issue is about understanding how does the binaural auditory system distinguish a filtered spectrum from the original spectrum with the exact same spectral pattern. The answer seems to lay on the previous

[17] Headphone listening elicits only the perception of *lateralization* (by opposition to *localization –outside* the head) since the apparent location is perceived *inside* the head.

knowledge (stored information) about source characteristics and environment acoustics. If these are not familiar, the "learning" capability of the binaural auditory system seems to be effective shortly after the contact with an entirely new context.

3.4.3 Properties of Binaural Hearing

3.4.3.1 The Precedence Effect

When listening to a sound source in a room, besides the direct signal, multiple reflected replicas, or echoes, also reach the ears of the listener. However, these delayed replicas do not affect significantly the perception of the first arriving direct signal, as well as its localization. This precedence advantage of the leading sound happens because subsequent replicas are fused with the leading sound into a single acoustic impression, provided their distance to the leading sound is not excessive (about 5 ms for single clicks and less than 40 ms for generic speech or music sounds). In the case of fusion, the location of the sound source is mainly determined by the leading sound. If the replicas are made more and more distant from the leading sound, first a sensation of shift in the sound location takes place. Then, for even higher delays, in the order of 50 ms, distinct echoes are heard.

The importance of the precedence effect (or the "Haas effect") lies in that it eases the localization and identification of different sound sources, in adverse acoustical conditions such as in a strongly reverberative room.

3.4.3.2 The Cocktail Party Effect

This effect owes its name to a very familiar situation: when different people are engaged in simultaneous conversations, as in a party, a listener is able to distinguish the voice of a specific speaker if listening binaurally, but not if listening monaurally. This corresponds obviously to a masking effect which is explained in a simple way: masking is maximum if both masker and maskee are coincident in space. By listening binaurally, the location of the maskee (the specific speaker) is better differentiated from the location of the masker (the remaining voices) and therefore total masking does not occur. The practical advantage is that binaural hearing improves the detection and analysis ability of signals in noisy backgrounds. This release from masking will be better characterized in the next sub-section.

3.4.3.3 Binaural Masking Level Differences

The binaural release from masking consists in the improved detectability of the maskee when its location is diverse from the location of the masker, relative to the case where they coincide spatially. Thus, assuming a maskee is just masked by a masker when both signals are feed to only one ear, or both ears simultaneously, the maskee will become unmasked if:

- both signals are presented to the two ears but the phase of the masker or the phase of the maskee in inverted in one of the ears,
- the maskee is presented to both ears but the masker is presented to only one ear.

In these cases, in order to restore the masking conditions, the level of the masker must be increased again. The maximum difference between this new level and the monaural masking level is the masking level difference (MLD). Its value is about 15 dB at low frequencies (around 500 Hz) and about $2-3$ dB for frequencies above 1500 Hz.

MLDs are particularly important in the context of stereophonic audio coding [Ferre.92]. In this context there are two audio channels (maskers) and two quantization signals (maskees). These are uncorrelated if each audio channel is coded independently using monaural masking models, thus eliciting conditions for the release from masking. In order to prevent this, a possible solution is to modulate the quantization noise in order to follow the acoustic image of the stereophonic signal. Alternatively, the problem may also be addressed in a conservative way by using the additional MLD safety margin.

3.5 A Final Comment on the Accuracy of a Psychoacoustic Model

While there are a number of psychoacoustic results well consolidated, many others are controversial and are still a matter of debate. Selecting the most suitable psychoacoustic results and models to estimate how the ear uses and discards acoustic information, is a problem whose challenges are well illustrated by a Moore's thought [Moore89, page 205]:

> *Pinnae differ considerably between people, and it would be of interest to know how far we are able to use, or to learn to use, information from others people pinnae.*

Psychoacoustic models result from subjective data. Their validity is therefore dependent on many factors including the test conditions, the training of subjects, the idiosyncrasy of their natural language, as well as their musical preferences and traditions. As a result, a given psychoacoustic model can not be presented as being a precise model. Precision, in the sense of exactness, would imply that a particular model only fits a specific subject because it only represents and mimics his (or her) exact hearing properties and tolerances.

A given psychoacoustic model can only be said to be accurate. Accuracy, in the sense of correctness, says how well a model is faithful to the general trend of few but important aspects of the human auditory system having the largest consensus and acceptance among the Scientific Community. This has been our main approach in building a psychoacoustic model. It is our opinion that the advantage of this approach is that it provides performance while promoting simplification and complexity reduction.

3.6 Summary

We have presented in this Chapter a brief survey of the most important properties and tolerances of the human auditory system, from the point of view of perceptual audio coding.

We have proposed a new simplified model of simultaneous masking. This model has been integrated in the ASC psychoacoustic model.

The perceptual implications of important features of the audio signal such as stationarity and harmonicity, have also been addressed. These topics have a clear influence on the implementation and performance of audio coders, as we will discuss in Chapters 4, 5, and 6.

Chapter 4

Audio Decomposition Using Multirate Filter Banks

divide ut regnes - [to divide is to reign]

Machiavelli, Florentine political philosopher, 16th century.

4.1 Introduction

We have seen in the previous Chapter that the human auditory properties and tolerances can be accurately characterized, especially in the frequency domain. Therefore, a psychoacoustic model fairly representative of these properties and tolerances, may pave the way to an efficient reduction of the data rate needed to represent the audio signal, provided that an appropriate spectral decomposition of the audio signal is performed.

In this Chapter we will focus on multirate analysis and synthesis structures allowing a suitable representation of audio signals for analysis, compression, and post-processing.

In Section 4.2 we enumerate the desired properties of a multirate filter bank matching our objective. In Section 4.3 we review briefly the theory of multirate filter banks and address the design, the performance and implementation issues of prospective solutions in the light of the desired properties. In particular, we compare uniform and non-uniform filter banks considering such aspects as structure, coding gain and ability to address multiresolution. We explain why uniform filter banks deserve our preference despite the fact that non-uniform filter banks are a better approximation to the critical bands presented in the previous Chapter. In particular, the main advantages of cosine modulated filter banks are highlighted in order to explain their frequent use in high quality audio coding as well as in ASC.

Using an extended formulation of a cosine modulated filter bank, it is possible to take advantage of a number of resulting properties. As an example, we show in Section 4.4 how we can extract important features of the audio signal such as the identification of tonal components, using the core structure of a cosine modulated filter bank.

In Section 4.5 we detail the design of an analysis/synthesis prototype filter considering, in particular, the effects of a multiplicative modification of the maximally decimated spectral representation of an audio signal. Section 4.6 summarizes this Chapter.

4.2 Desired Properties of the Multirate Filter Bank

In the context of audio compression, an ideal analysis/synthesis filter bank should allow fine frequency and temporal resolution, perfect reconstruction (PR), critical sampling and the reduction of *blocking effects*. Last but not least, it should be possible to use fast algorithms for its implementation. In the following we will characterize each aspect separately.

Time/Frequency Resolution As we suggested in the previous Chapter, simultaneous masking is the most important aspect in psychoacoustic modeling because of its potential in the reduction of irrelevant components of the audio signal. It implies fine frequency resolution, particularly in the low frequency region of the audible spectrum. Although a non-uniform frequency resolution according to the *Bark* rate could be appropriate, the aspects of pitch and timbre perception reveal that an uniform frequency resolution could not only be suitable but eventually also desirable. The filter bank should therefore exhibit a frequency resolution finer than the narrowest critical bandwidth which is less than 100 *Hz*.

Fine frequency resolution implies coarse temporal resolution. Given that the former is necessary to achieve significant coding gains [Jayant84], the filter bank should enable fine temporal resolution (therefore sacrificing coding gain) only when coding quality is at risk, such as when forward and backward masking requirements are not respected.

Perfect Reconstruction The PR condition requires that when the signal is not modified during the coding process, the reconstructed signal should be an exact replica of the original signal, except for a system delay. Although this is not a stringent requirement from the perceptual point of view, it facilitates the understanding of the nature of the coding noise, in relation to a specific spectral modification (essentially due to quantization) of the audio signal.

Maximal Decimation Sub-sampling of the spectral components (or channels) is allowed because they represent a fraction of the original signal bandwidth, which means that their sampling frequency can be correspondingly reduced. Ideally, the sum of the data rate of all spectral components should not exceed the input data rate. In other words, the spectral components should be maximally decimated (critically sampled) in order to not impose a burden on the compression task.

Reduction of Blocking Effects Classical block transforms based on the coding of disjoint blocks of data (*i.e.* block-by-block without overlap), exhibit objectionable coding artifacts because of discontinuities in the reconstructed signal at block boundaries. These are known as *block edge effects* or *blocking noise.* To avoid these effects, the coding noise must be somewhat correlated at the edges of adjacent transform blocks. As we explain in Sub-section 4.3.1.4, overlapped transforms offer the possibility to attenuate strongly the coding noise [Malvar92].

Fast Algorithms The analysis/synthesis filter bank should be implemented using fast computational algorithms such as the Fast Fourier Transform (FFT) or the Discrete Cosine Transform (DCT). As we indicated in Section 2.4, we think that implementation concerns should be considered as early as possible in the design process. We may anticipate that the computational efficiency objective is elegantly satisfied by cosine modulated filter banks.

4.3 Theory of Multirate Filter Banks

Historically, filter banks have been classified as subband filters or transforms. These designations depended essentially from the relation between the length of the analysis/synthesis filters and the number of subbands, and from the real or complex nature of the filter coefficients, given that normally the input signal is real. We will henceforth use both designations indistinctly since they just correspond to different perspectives to the same problem (multirate filter banks), which is formally treated using an unified theory. Because of its simplicity and importance, we will focus on uniform filter banks.

4.3.1 Uniform Filter Banks

4.3.1.1 The General Structure of Uniform Filter Banks

Figure 4.1 represents a complete analysis/synthesis system based on an uniform filter bank. In this figure, $H_k(z)$ represents the Z transform $H_k(z) = \sum_{n=-\infty}^{+\infty} h_k(n)\, z^{-n}$ of the k_{th} filter. The input signal $x(n)$ is filtered by N subband filters $H_0(z)$ to $H_{N-1}(z)$, whose outputs are

$x_0(n)$ to $x_{N-1}(n)$. As the signal bandwidth of each of these outputs represents[1] about $2\pi/N$, the subband signals are sub-sampled[2] by M with $M \leq N$. Critical sampling (or maximal decimation) is obtained if $M = N$. The decimated subband signals $x_{D,0}(n)$ to $x_{D,N-1}(n)$ are

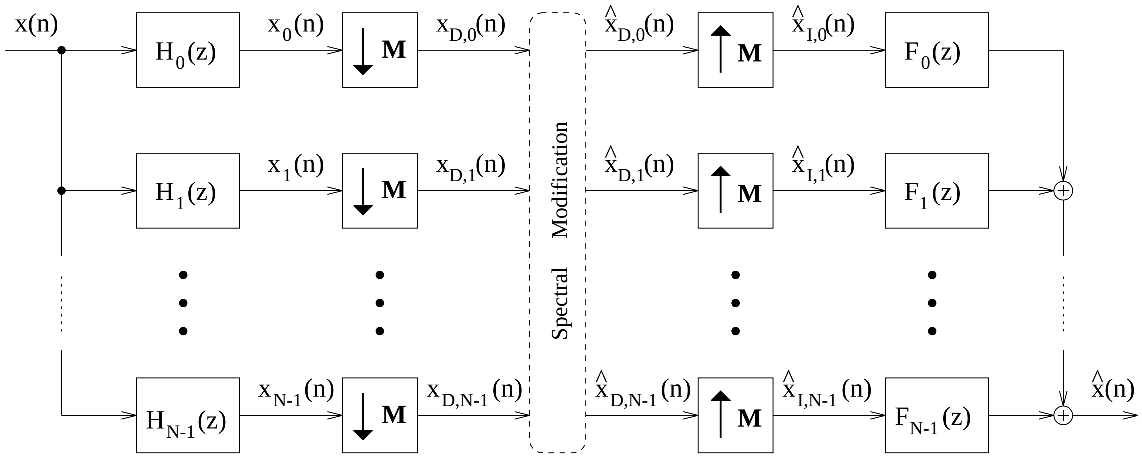

Figure 4.1: An uniform filter bank with N channels. Each channel comprises one analysis filter ($H(z)$) and one synthesis filter ($F(z)$). These are necessary to avoid aliasing after decimation by M, and to avoid spectral images after interpolation by M. The filter bank is maximally decimated if M=N.

then modified in some way to serve the purpose of signal compression. The modified signals $\hat{x}_{D,0}(n)$ to $\hat{x}_{D,N-1}(n)$ are interpolated by a factor of M, resulting in $\hat{x}_{I,0}(n)$ to $\hat{x}_{I,N-1}(n)$. These are then filtered by the corresponding synthesis filters $F_k(z) = \sum_{n=-\infty}^{+\infty} f_k(n)\, z^{-n}$ in order to eliminate the images of the spectrum of each subband channel. The output of all synthesis filters are summed to reconstruct a replica $\hat{x}(n)$ of the original signal. We will find the relation between $\hat{x}(n)$ and $x(n)$ or, equivalently, between their Z transforms $\hat{X}(z) = \sum_{n=-\infty}^{+\infty} \hat{x}(n)\, z^{-n}$ and $X(z) = \sum_{n=-\infty}^{+\infty} x(n)\, z^{-n}$.

4.3.1.2 The Relation Between Input and Output

The Z transform of the output of each filter is simply given by

$$X_k(z) = H_k(z)X(z)\,, \quad 0 \leq k \leq N-1\,. \tag{4.1}$$

[1]The analog sampling frequency corresponds to the digital frequency 2π. If both $x(n)$ and $h_k(n)$ are real, then the bandwidth of $x_k(n)$ is π/N.

[2]Contrary to what is usually assumed in the literature, we admit the decimation factor M is in general different from (but necessarily inferior or equal to) the number of channels N. This generalization easily accommodates the presentation of cosine modulated filter banks later on.

After decimation by M, there will be $M-1$ images of $X(z)$, creating potential aliasing problems [Crochi.83]:

$$X_{D,k}(z) = \frac{1}{M}\sum_{l=0}^{M-1} X_k\left(z^{\frac{1}{M}}W^l\right) = \frac{1}{M}\sum_{l=0}^{M-1} H_k\left(z^{\frac{1}{M}}W^l\right) X\left(z^{\frac{1}{M}}W^l\right) \ , \quad 0 \le k \le N-1 \ , \quad (4.2)$$

where $W = e^{-j\frac{2\pi}{M}}$. In the absence of spectral modification, $X_{D,k}(z) = \hat{X}_{D,k}(z)$ and the interpolated subband signals result as:

$$\hat{X}_{I,k}(z) = X_{D,k}\left(z^M\right) = \frac{1}{M}\sum_{l=0}^{M-1} H_k\left(zW^l\right) X\left(zW^l\right) \ , \quad 0 \le k \le N-1 \, . \quad (4.3)$$

Finally, $\hat{X}(z)$ can be obtained as [Vaidy.87a]:

$$\hat{X}(z) = \sum_{k=0}^{N-1} \hat{X}_{I,k}(z)F_k(z) = \frac{1}{M}\sum_{l=0}^{M-1} A_l(z)X\left(zW^l\right) \ , \quad (4.4)$$

where

$$A_l(z) = \sum_{k=0}^{N-1} H_k\left(zW^l\right) F_k(z) \, . \quad (4.5)$$

From this result a few conclusions can be drawn.

1. The reconstructed spectrum consists in the combination of the original spectrum and its $M-1$ uniformly shifted versions corresponding to a periodically time varying system with period M.

2. Aliasing is completely removed if $A_l(z) = 0$ for $1 \le l \le M-1$. However, this condition does not prevent amplitude and phase distortions.

3. If the aliasing is removed, the amplitude and phase distortions depend solely on the analysis and synthesis filters $\sum_{k=0}^{N-1} H_k(z)F_k(z)$. A careful design of these filters might reduce or eliminate completely those distortions.

Perfect reconstruction implies that all types of distortions are eliminated or exactly cancelled (in the absence of spectral modification). As a consequence, the output is an exact replica of the input except for a possible scaling factor, c, and a system delay n_0:

$$\hat{x}(n) = c\ x(n-n_0) \quad (4.6)$$

or, equivalently,

$$\hat{X}(z) = c\ z^{-n_0} X(z) \, . \quad (4.7)$$

For $N > 2$, except for trivial analysis and synthesis filters (which are not of practical interest), the above system analysis does not provide insight into practical filter design procedures[3] so as to reach PR. In fact, design difficulties (matrix inversion) and system instabilities arise (Infinite Impulse Response filters), unless we '*are satisfied with alias cancellation*' only [Vaidy.93, page 229]. In this case PR is not achieved, both analysis and synthesis filters are Finite Impulse Response (FIR) filters (ensuring system stability), but new difficulties arise such as filters of very large order. A more revealing system analysis involves the polyphase decomposition [Crochi.83].

4.3.1.3 Input to Output Relation Using the Polyphase Decomposition

Using the Type 1 polyphase decomposition, the analysis filters, of size L (we assume $L \geq N$), can be expressed as:

$$H_k(z) = \sum_{l=0}^{N-1} z^{-l} E_{kl}\left(z^N\right) , \tag{4.8}$$

and using the Type 2 polyphase decomposition, the synthesis filters, of size L (we assume $L \geq N$), can be expressed as:

$$F_k(z) = \sum_{l=0}^{N-1} z^{-(N-1-l)} R_{lk}\left(z^N\right) . \tag{4.9}$$

The resulting analysis/synthesis system is illustrated in Figure 4.2. Using the *noble identities* [Crochi.83], the filter bank can be further modified as depicted in Figure 4.3 which highlights two important aspects. First, the filtering operations (based on the coefficients of $\mathbf{E}(z)$ and $\mathbf{R}(z)$) are computed at the lowest possible rate. Second, the combination of the delay chain with the decimation blocks on the analysis side, and the combination of the interpolation blocks with the delay chain on the synthesis side, leads to very efficient and elegant implementation solutions using the *commutator* model [Crochi.83]. From Equation (4.8) we build the matrix $\mathbf{E}(z)$ as follows:

$$\mathbf{E}(z) = \begin{bmatrix} E_{00}(z) & E_{01}(z) & \cdots & E_{0,N-1}(z) \\ E_{10}(z) & E_{11}(z) & \cdots & E_{1,N-1}(z) \\ \vdots & \vdots & \ddots & \vdots \\ E_{N-1,0}(z) & E_{N-1,1}(z) & \cdots & E_{N-1,N-1}(z) \end{bmatrix} . \tag{4.10}$$

The matrix $\mathbf{R}(z)$ is built in a similar way from Equation (4.9), only with a difference in the order of the row and column indices. The rows of $\mathbf{E}(z)$ represent the analysis basis functions (*i.e.* the impulse response of the analysis filters $h_k(n)$) of the transformation, and the columns of $\mathbf{R}(z)$ represent the (time-reversed) synthesis basis functions (*i.e.* the impulse response of the

[3] N=2 corresponds to the two-channel Quadrature Mirror Filter for which many design procedures exist (*e.g.* based on *lattice* structures) [Vaidy.87b][Johns.80].

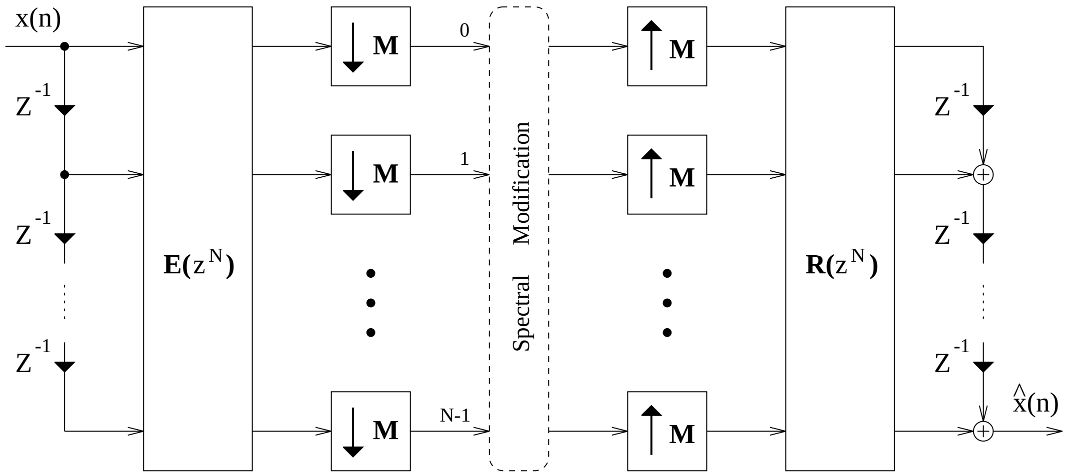

Figure 4.2: Polyphase representation of a filter bank with N channels. The filter bank is maximally decimated if M=N.

synthesis filter $f_k(n)$) of the transformation. Using these matrices, a necessary and sufficient condition[4] for perfect reconstruction [Vaidy.93, page 237] is that

$$\mathbf{R}(z)\mathbf{E}(z) = \mathbf{I}\,, \tag{4.11}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. If both $H_k(z)$ and $F_k(z)$ filters are FIR, then the PR requirement reduces to $det\ \mathbf{E}(z) = \alpha_0 z^{-k_0}$, $\alpha_0 \neq 0$, $k_0 =$ integer, *i.e.* the determinant of $\mathbf{E}(z)$ must be a delay. In this case, it results that $det\ \mathbf{R}(z) = \alpha_1 z^{-k_1}$.

A well-known case of particular interest arises when $L = N = M$ (block transforms) and when $\mathbf{E}$ is an orthogonal transform, *i.e.* a matrix for which $\mathbf{E}^{-1} = \mathbf{E}^T$. The most common example is the block DFT transform. In this case, the analysis and synthesis basis functions are simply given by:

$$E^*_{kl}(z) = R_{lk}(z) = \frac{1}{\sqrt{N}} W^{kl}\,, \quad 0 \leq k, l \leq N-1\,. \tag{4.12}$$

In more general terms, perfect reconstruction is achieved if $\mathbf{E}(z)$ is *paraunitary*[5]:

$$\tilde{\mathbf{E}}(z)\mathbf{E}(z) = d\mathbf{I}\,, \quad d = \text{constant} > 0\,, \tag{4.13}$$

where $\tilde{\mathbf{E}}(z)$ is the paraconjugate of $\mathbf{E}(z)$ and is defined as $\tilde{\mathbf{E}}(z) = \mathbf{E}^\dagger(1/z^*)$ [Vaidy.93]. $\mathbf{E}^\dagger(z)$ is

[4]The necessary and sufficient condition for alias cancellation (which is a less stringent requirement than perfect reconstruction) is that the product $\mathbf{R}(z)\mathbf{E}(z)$ is a *pseudocirculant matrix* [Vaidy.93, page 249].

[5]Paraunitariness of $\mathbf{E}(z)$ means that $\mathbf{E}(z)$ is unitary for all Z and not just on the unit circle of the Z plane (in which case $\mathbf{E}^\dagger(e^{j\omega})\mathbf{E}(e^{j\omega}) = d\mathbf{I}$, for all ω, and for some $d > 0$). A causal and stable paraunitary system is said to be a lossless system. Losslessness denotes the input-to-output conservation of energy which, for a causal and stable FIR system is equivalent to the power complementary property. For this reason, paraunitariness of $\mathbf{E}(z)$ is not a necessary condition for obtaining the PR property in FIR filter bank systems [Vaidy.93, page327].

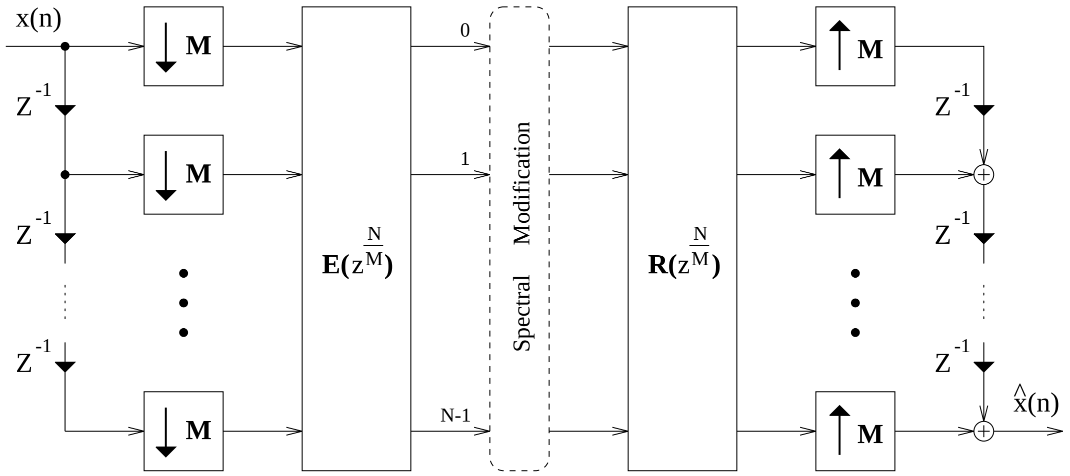

Figure 4.3: Polyphase representation of a filter bank with N channels where the *noble identities* were used to shift the filtering operations to the side of lowest possible computing rate. The filter bank is maximally decimated if M=N.

the transpose conjugate matrix of $\mathbf{E}(z)$. It follows that $\mathbf{R}(z)$ can be easily obtained:

$$\mathbf{R}(z) = cz^{-(K-1)}\tilde{\mathbf{E}}(z)\,, \quad c \neq 0\,. \tag{4.14}$$

The delay $K-1$ is necessary to ensure that $\mathbf{R}(z)$ [hence $F_k(z)$] is causal. If $\mathbf{E}(z)$ is FIR, paraunitariness implies important relations between the analysis filters $[h_k(n)]$ and the synthesis filters $[f_k(n)]$ (points 1,2, and 3 below), as well as other relevant results (points 4,5, and 6 below) [Vaidy.93]:

1. $h_k(n)$ and $f_k(n)$ have the same length L, $(L = KN)$,

2. $f_k(n)$ is obtained simply as $f_k(n) = ch_k^*(L-1-n)$, in the frequency domain this corresponds to $F_k(z) = cz^{-(L-1)}\tilde{H}_k(z)$,

3. the magnitude response of $F_k(z)$ is the same as that of $H_k(z)$ except for a scale factor: $|F_k(e^{j\omega})| = c\,|H_k(e^{j\omega})|$,

4. the analysis and synthesis filters are power complementary, *e.g.* for the former this means that the quantity $\sum_{k=0}^{N-1} |H_k(e^{j\omega})|^2$ is a constant,

5. each analysis (synthesis) filter has unit energy: $\int_0^{2\pi} |H_k(e^{j\omega})|^2\, d\omega = 2\pi$, which can also be represented as $\sum_{l=0}^{N-1} \left|H_k(e^{j\omega}W^l)\right|^2 = N$,

6. each analysis filter $H_k(z)$ is a spectral factor of a zero phase Nth-band filter [Vaidy.87b], *i.e.* $\tilde{H}_k(z)H_k(z)$ is a Nth-band filter[6].

[6] In particular, for $N = 2$, $H_0(z)$ and $H_1(z)$ are spectral factors of a *half-band* filter. Spectral factorization is therefore an interesting design method, especially for $N = 2$, but faces serious difficulties for $N > 2$.

Although this analysis simplifies significantly the design of the filter bank, *all* the analysis filters must first be *simultaneously* found, using for example a cascaded structure of Householder matrices and non-linear optimization techniques, in order to obtain good frequency responses [Vaidy.93, page 314]. The procedure may also involve factorization techniques of Nth-band filters. However, as pointed out by Malvar [Malvar92, page 126] *'even with the improved techniques discussed above, the design of perfect reconstruction FIR filter banks is far from being an easy task, mainly for large N,* e.g. $N > 16$*'* (N is the number of channels and not the length of each filter). This reflects the fact that the computational complexity of the optimization procedure increases quadratically with N. In speech and audio processing a number of channels higher than 16 is usually desired.

The next alternative to reduce drastically the number of optimization parameters, and therefore to simplify the design procedure while maintaining the PR requirement, is called cosine modulated FIR filter banks.

4.3.1.4 Cosine Modulated FIR Filter Banks

Cosine modulated filter banks have the property that all analysis filters and synthesis filters are derived from a prototype filter $H(z)$ by cosine modulation. In order to explain how this can be easily achieved in practice, we will illustrate exponential modulation first, and then we will merge the appropriate complex filters in order to obtain the real cosine modulation. The complex analysis filters $H_k(z)$ are obtained from the prototype $H(z)$ by a right shift in frequency:

$$H_k(z) \triangleq H\left(zW^{k+\frac{1}{2}}\right)\ , \quad 0 \le k \le N-1\ , \tag{4.15}$$

where W is redefined as $W = e^{-j\frac{2\pi}{N}}$. Using the polyphase representation of the prototype filter, Equation (4.15) becomes:

$$H_k(z) = \sum_{l=0}^{N-1} z^{-l} W^{-(k+\frac{1}{2})l} E_l\left(-z^N\right) = \sum_{l=0}^{N-1} \left(z^{-1}W^{-\frac{1}{2}}\right)^l W^{-kl} E_l\left(-z^N\right)\ . \tag{4.16}$$

We are using modulation by $W^{k+\frac{1}{2}}$ and not just by W^k because we want to merge filters with identical bandwidths (odd-channel stacking [Malvar92, pages 108,176]). Figure 4.4 illustrates that in the latter case, the analysis filter $H_0(k)$ would have a different bandwidth from that of the remaining filters. The complex synthesis filters $F_k(k)$ are obtained by:

$$F_k(z) \triangleq -W^{-(k+\frac{1}{2})} F\left(zW^{k+\frac{1}{2}}\right)\ , \quad 0 \le k \le N-1\ . \tag{4.17}$$

Using the polyphase representation of the prototype filter, Equation (4.17) becomes:

$$F_k(z) = \sum_{l=0}^{N-1} z^{-(N-1-l)} R_l\left(-z^N\right) W^{(k+\frac{1}{2})l} = \sum_{l=0}^{N-1} z^{-(N-1-l)} W^{\frac{l}{2}} R_l\left(-z^N\right) W^{kl}\ . \tag{4.18}$$

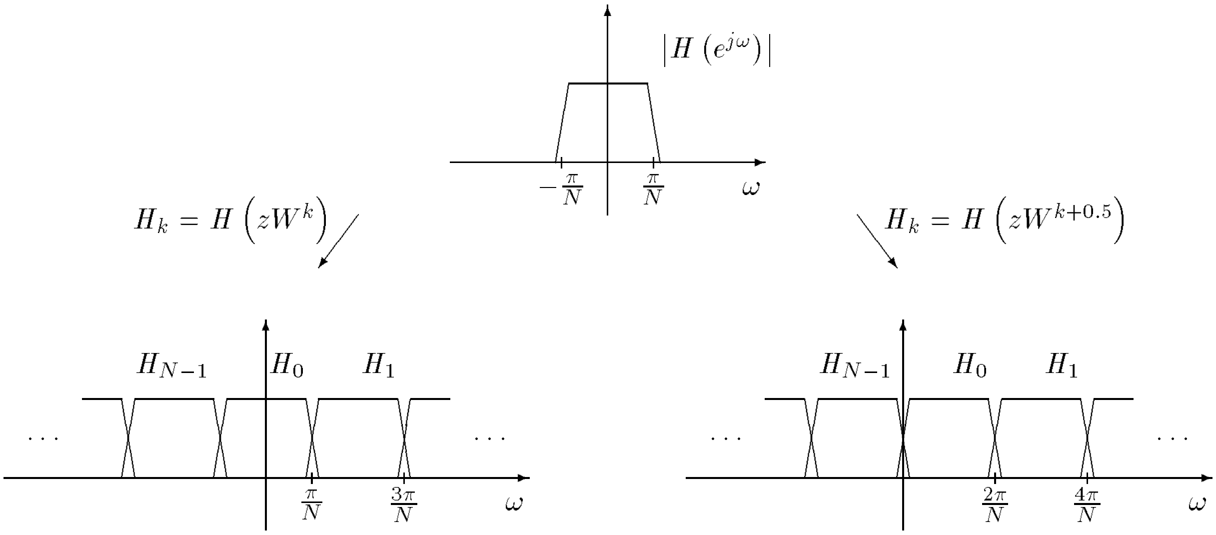

Figure 4.4: Modulation of a prototype filter (top) resulting in *even channel stacking* (left) or in *odd channel stacking* (right).

Cosine modulation is obtained by combining the appropriate complex filters $H_k(z)$ in the analysis side, and the appropriate complex filters $F_k(z)$ in the synthesis side, such as to produce M real channels. The resulting analysis and synthesis basis functions are real and we will designate them respectively as $p_k(n)$ and $q_k(n)$, $0 \leq k \leq M-1$. Using Equation (4.16) and Equation (4.18), and recalling that the notation W^{kl} represents the element $e^{-j\frac{2\pi}{N}kl}$ in row k and column l of the DFT matrix $\mathbf{W}$, the complete (expanded) analysis/synthesis system using the polyphase decomposition is illustrated in Figure 4.5.

From this figure, a number of remarks can be made and a number of conclusions can be drawn.

- The cosine modulation block represents a zero-delay linear combination of N complex signals into M real signals. The cosine demodulation block represents a zero-delay mapping of M real signals into N complex signals. Therefore, if the length of the prototype filter is $L = N$, the total system delay results as $N-1$, *i.e.* in the absence of spectral modification, $\hat{x}(n) = x(n-N+1)$.

- The matrix $\mathbf{W}$ is commutable with the polyphase filters and the decimation or interpolation blocks because this consists in a memoriless operation (since $\mathbf{W}$ is a zero-order polyphase component matrix).

- Since $\mathbf{W}^*$ is depicted on the analysis side (instead of $\mathbf{W}$), it would appear at first sight

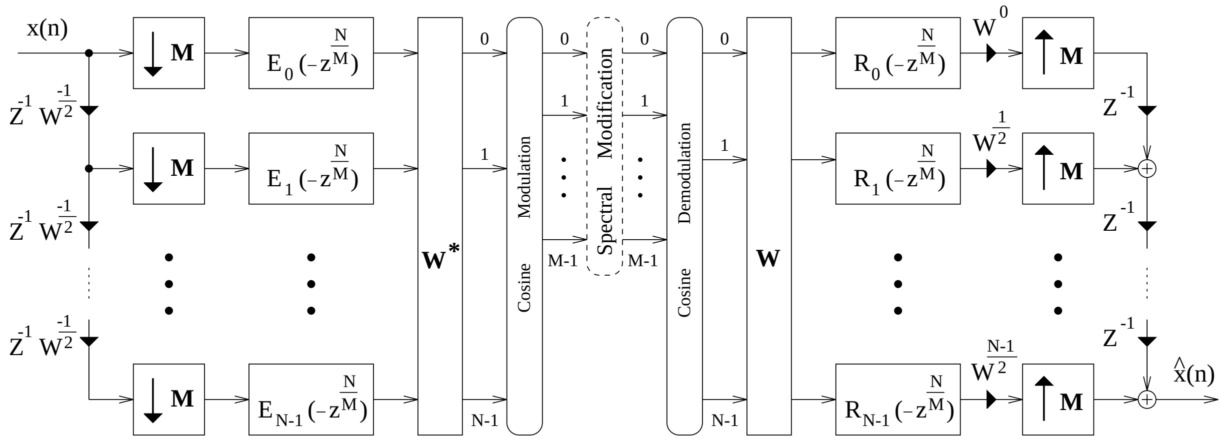

Figure 4.5: Cosine modulated filter bank with M channels. The system is critically sampled and achieves PR if the polyphase components satisfy the power complementary requirement. It is usual in practice to take N=2M.

that we are computing the inverse DFT. In fact $\mathbf{W}^{-1} = \mathbf{W}^{\dagger}/N$ and since $\mathbf{W}$ is symmetric (*i.e.* $\mathbf{W}^T = \mathbf{W}$), then $\mathbf{W}^{-1} = \mathbf{W}^*/N$. The proof that we are in fact computing the DFT can be obtained easily by realizing that the input data $v(n)$, just before the $\mathbf{W}^*$ matrix, is taken "backwards", *i.e.* we are computing $\sum_{i=0}^{N-1} v(N-1-i)W^{-ki} = W^k \sum_{l=0}^{N-1} v(l)W^{kl}$ [Vaidy.93, page 116]. Thus except for a scale factor, we are computing the DFT on the analysis side of the filter bank.

- Cosine modulation allows critical sampling while benefiting from 'block' overlap (which reduces the blocking noise). For example, if $L = N = 2M$, which is a case of particular interest to us, M spectral coefficients are generated for each new set of M temporal samples while keeping a 50% overlap between adjacent data blocks of length $2M$.

- The computational cost of the filter bank is that of the polyphase filter bank (which includes the DFT shown in Figure 4.5), plus the cosine modulation. This combined processing may benefit from efficient techniques such as the FFT (as implicitly suggested in Figure 4.5), or the Discrete Cosine Transform [Vaidy.93]. Another possibility (which we will exploit in Chapter 6) results directly from Equations (4.16) and (4.18) since $W^{k+\frac{1}{2}}$ is the complex exponential of the Odd-DFT transform [Bellan.89]. This transform can be implemented using the FFT algorithm without any pre-processing or post-processing [Ferre.92]. In fact, the only restriction is that the direct Odd-DFT should be implemented using the *decimation-in-time* method [Oppenh.75] (*i.e.* based on sub-sequences of the input samples); whereas the inverse Odd-DFT should be implemented using the *decimation-in-frequency* method (*i.e.* based on sub-sequences of the output signal). We will also use the Odd-DFT for other purposes, namely for feature extraction (Sub-section

4.4.2), equalization, and time-scale expansion (Chapter 7).

- The design of cosine modulated filter banks is extremely simplified because the optimization of the whole analysis system is reduced to the optimization of a single prototype filter, as we will show in the next sub-section.

4.3.1.5 The Design of the Prototype Filter

The first cosine modulation technique to be historically developed is known as the Pseudo-QMF [Chu85]. It achieves only approximate alias cancellation (only the aliasing from adjacent bands is cancelled), exhibits amplitude distortion, and eliminates phase distortion. Vaidyanathan [Vaidy.93, section 8.1] details (using a notation different from that we have been using) how the Pseudo-QMF merges the N complex outputs to produce M real outputs (as illustrated in Figure 4.5). The basis functions of the Pseudo-QMF are real and are given by:

$$p_k(n) = h(n)\cos\left[\frac{\pi}{M}\left(k+\frac{1}{2}\right)\left(n-\frac{L-1}{2}\right)+\theta_k\right]\ , \quad 0 \le k \le M-1\ , \tag{4.19}$$

$$q_k(n) = h(n)\cos\left[\frac{\pi}{M}\left(k+\frac{1}{2}\right)\left(n-\frac{L-1}{2}\right)-\theta_k\right]\ , \quad 0 \le k \le M-1\ , \tag{4.20}$$

where $\theta_k = (-1)^k\pi/4$, and $h(n)$ is the prototype filter.

It was later recognized that PR could be achieved with a system based on Equation (4.19) if several conditions are met:

1. $L = N = 2mM$, for some integer m,
2. the mirror image condition of the synthesis filter: $f_k(n) = h_k(L-1-n)$ (or equivalently, $F_k(z) = z^{-(L-1)}H_k(z^{-1})$),
3. the linear phase constraint of the prototype filter: $h(n) = h(L-1-n)$,
4. if the polyphase components $E_k(z)$, $0 \le k \le 2M-1$, of the prototype filter $H(z)$ satisfy the pairwise power complementary requirement [Vaidy.93, page 379]:

$$\tilde{E}_k(z)E_k(z) + \tilde{E}_{M+k}(z)E_{M+k}(z) = \text{constant}\ , \quad 0 \le k \le M-1\ . \tag{4.21}$$

In the following we will consider $L = N = 2M$.

A simple and elegant example of a cosine modulated filter bank achieving perfect reconstruction is the (oddly stacked) Time Domain Aliasing Cancellation Filter Bank (TDAC) [Princen87], or the Modulated Lapped Transform (MLT) [Malvar92]. Both analysis and synthesis basis functions are as follows:

$$p_k(n) = q_k(N-1-n) = h(n)\cos\frac{\pi}{M}\left(k+\frac{1}{2}\right)\left(n+\frac{1}{2}+\frac{M}{2}\right) \;, \quad 0 \leq k \leq M-1\,. \tag{4.22}$$

This result presumes implicitly that the prototype filter exhibits even symmetry $h(n) = h(N-1-n)$, and satisfies the power complementary requirement $h^2(n) + h^2(n+M) =$ constant. A simple prototype filter which satisfies these requirements, as well as the *polyphase normalization* requirement [Malvar92, page 177] (*i.e.* that concentrates the DC component of the signal on a single spectral line), is given by:

$$h(n) = \sin\frac{\pi}{N}\left(n+\frac{1}{2}\right) \;, \quad 0 \leq n \leq N-1\,. \tag{4.23}$$

This window is widely used in the audio coding context and is usually referred to as the 'sine window'. We will show in Chapter 7 that it can be derived from the Hanning window. We will *also* show in that Chapter how we can take advantage of other interesting properties of this window, when combined with the Odd-DFT, to perform time-scale expansion of audio signals, through a delicate spectral expansion operation.

Other solutions seeking the optimization of the coefficients of the prototype filter $H(z)$, have essentially focussed on frequency domain aspects such as the improvement of the stop-band attenuation [Vaidy.93, section 8.5.2][Malvar92, section 5.3]. In section 4.5 we will present an optimization procedure that takes into account time *and* frequency domain aspects.

We have selected the oddly stacked TDAC filter bank (or MLT) as the analysis/synthesis system of ASC. The main reasons were its simplicity, the fact that it meets all the desired properties presented in Section 4.2 without incurring in an excessive system delay, that there is the possibility of fast computation using the Odd-DFT as suggested by Equations (4.16) and (4.18), and because it accommodates the possibility of post-processing as can be concluded from Figure 4.5.

As a final note, we should make clear that although the prototype filter has linear phase, the basis functions *per se* do not because they do not exhibit symmetry. However, the complete analysis/synthesis is perfect reconstructing, which means it is linear phase.

4.3.1.6 Objective Coding Gain

In the audio coding context, the spectral decomposition of an audio signal consists in a decisive factor contributing to the efficient reduction of the data rate (in bits per second) of its PCM representation. This is achieved by quantizing individually each subband channel so as to shape the quantization noise according to an appropriate criterion. Our knowledge of the properties and tolerances of the human auditory system reveals that a suitable criterion is to achieve complete masking of the quantization noise or, if this is not possible, to minimize its perceptibility (*i.e.* its relative loudness). Thus, this is eminently a subjective criterion. However, historically, the emphasis has been placed on an objective criterion which is an important classic reference: the objective coding gain over PCM [Jayant84].

In order to define the objective coding gain and a similarly derived *perceptual coding gain*, a few preliminary considerations should be made. We assume that the spectral decomposition leads to uncorrelated subband signals[7] so that the variance of the full-band input signal, σ_x^2, can be expressed as a sum of the variances of the maximally decimated subband signals σ_{xk}^2:

$$\sigma_x^2 = \frac{1}{N} \sum_{k=0}^{N-1} \sigma_{xk}^2 \ . \tag{4.24}$$

Assuming further that paraunitary subband coders [Soman93] and scalar uniform quantizers are used, it follows that the reconstruction error variance, σ_q^2, can be expressed as a sum of the subband error variances, σ_{qk}^2, when the subband samples are quantized to R_k bits [Jayant84]:

$$\sigma_q^2 = \frac{1}{N} \sum_{k=0}^{N-1} \sigma_{qk}^2 = \frac{1}{N} \sum_{k=0}^{N-1} \epsilon_k^2 2^{-2R_k} \sigma_{xk}^2 \ . \tag{4.25}$$

The constant ϵ_k^2 is the quantizer performance factor. If all subbands are maximally decimated, the equivalent average number of bits used to encode (quantize) a full-band sample is simply given by:

$$R = \frac{1}{N} \sum_{k=0}^{N-1} R_k \ , \tag{4.26}$$

and the associated error variance results as $\epsilon^2 2^{-2R} \sigma_x^2$. The coding gain over PCM expresses the Signal-to-Noise Ratio (SNR) improvement resulting from a judicious subband bit allocation relative to a full-band fixed bit allocation (assuming that $\epsilon_k^2 = \epsilon^2 \ \forall k$, which is a reasonable assumption provided that the R_k values are not too coarse, *i.e.* safely above zero):

$$G = \frac{2^{-2R} \sigma_x^2}{\frac{1}{N} \sum_{k=0}^{N-1} 2^{-2R_k} \sigma_{xk}^2} \ . \tag{4.27}$$

For stationary signals, this gain increases with N and verifies $G \geq 1$, with the equality occurring for a flat spectrum. If an average bit rate R is given as a constraint, then the maximization of (4.27) is obtained through the minimization of its denominator, *i.e.* by minimizing the mean

[7]The input signal, the subband signals and all quantization errors, are assumed to represent zero-mean, wide sense stationary processes. The quantization error is also assumed to be uncorrelated with the input signal.

square value of the reconstruction error. This problem is solved using the method of Lagrange multipliers, which leads to the optimum bit allocation:

$$R_k = R + \frac{1}{2}\frac{\sigma_{xk}^2}{\left[\prod_{\ell=0}^{N-1}\sigma_{x\ell}^2\right]^{1/N}} . \tag{4.28}$$

This bit allocation results in equal quantizer noise variances for all subband signals, which means the reconstruction noise is white. In the case of uniform quantizers and not too coarse quantization, this means

$$\sigma_q^2 = \sigma_{qk}^2 = \frac{\Delta^2}{12} , \tag{4.29}$$

where Δ is the quantization step. The resulting optimum coding gain is the ratio of the arithmetic mean to the geometric mean of the subband variances [Jayant84]:

$$G_{MAX} = \frac{\frac{1}{N}\sum_{k=0}^{N-1}\sigma_{xk}^2}{\left[\prod_{\ell=0}^{N-1}\sigma_{x\ell}^2\right]^{1/N}} . \tag{4.30}$$

This gain increases with N, is upper bounded by the inverse of the *spectral flatness measure* [Jayant84], and assumes an optimum spectral decomposition through paraunitary subband coders, which include the brick-wall subband coder and the orthogonal transform coder as particular cases [Soman93]. Furthermore, as the arithmetic mean of positive numbers is always greater than or equal to their geometric mean, there is always a coding gain[8], unless the signal spectrum is flat, *i.e.* all the subband variances are equal. The orthogonal transform that maximizes the coding gain is the Karhunen-Loeve Transform (KLT). In fact, this transform diagonalizes the covariance matrix which means the subband signals are uncorrelated (which legitimates the independent quantization of each subband signals), and the geometric mean of variances is minimized. However, the KLT is signal dependent and does not have a fast computational algorithm which strongly reduces its practical interest. For this reason, invariant transforms that are computationally efficient and that approach the KLT performance for specific signals (such as images), are preferred in practice and have been used with success [Malvar92].

It has been demonstrated that the above results concerning the objective coding gain over PCM, also hold for *lapped* paraunitary subband coders [Soman93]. This will allow us to analyze easily the implications of modifying the optimization criterion to account for perceptual aspects, while benefiting from the advantages of overlapped cosine modulated filter banks (which represent a particular case of paraunitary filter banks).

[8]It should be noted however that these are theoretical figures since the cost of side information needed to characterize the optimum bit allocation is not considered.

4.3.1.7 Perceptual Coding Gain

The conventional coding gain measure presented above, was based on the maximization of the SNR or, equivalently, on the minimization of the mean square value of the quantization noise. As perceptual rules allow the estimation of a Masked Threshold variance (Section 3.3.4) in each band k, σ^2_{Tk}, it is possible to identify the perceptually relevant amount of the signal variance, $\sigma^2_{xk}/\sigma^2_{Tk}$, and the perceptually significant amount of the quantization noise variance above the Masked Threshold: $\sigma^2_{qk}/\sigma^2_{Tk}$. Therefore, a *perceptual coding gain* would seek the maximization of the following quantity:

$$\frac{\sum_{k=0}^{N-1} \sigma^2_{xk}/\sigma^2_{Tk}}{\sum_{\ell=0}^{N-1} \sigma^2_{q\ell}/\sigma^2_{T\ell}} = \frac{\mathrm{SMR}}{\mathrm{NMR}}\,. \tag{4.31}$$

The Signal-to-Masked-Noise ratio (SMR) is a constant quantity since it is derived from the signal variance and from the estimated threshold THR_k (Section 3.3.4, Equation (3.5)). This means that the above maximization problem is equivalent to minimizing the Noise-to-Mask-Ratio (NMR) [Wei97]. Ideally, the NMR is a negative quantity since total masking is achieved when the quantization noise is below the Masked Threshold. This leads to an important conclusion: the reference for subjective quality is the Masked Threshold variance and not (as before) the signal variance. As a corollary, this means that a SNR-like measure is not appropriate to judge the subjective quality of the signal after being quantized. Instead, if the THR estimate is reliable, negative NMR values indicate a high probability of excellent subjective signal quality after quantization, and increasing positive NMR values indicate high probability of the quantization noise being unmasked, therefore reducing the subjective quality of the signal. This criterion has been proposed [Brand.89] to evaluate the subjective quality, together with another measure, the *masking flag*, which indicates if the quantization noise is above the Masked Threshold in at least one subband. Obviously, defined this way, these quality measures are only as reliable as the underlying psychoacoustic model. A more definitive way to judge subjective quality, although considerably much more expensive in terms of resources and time, requires careful listening tests involving both trained and untrained listeners.

As the perceptually relevant SMR and NMR quantities are obtained by weighting respectively the signal variance, $\mathrm{SMR}_k = \sigma^2_{xk}/\sigma^2_{Tk} = \mathrm{w}_k\sigma^2_{xk}$, and the quantization noise variance, $\mathrm{NMR}_k = \sigma^2_{qk}/\sigma^2_{Tk} = \mathrm{w}_k\sigma^2_{qk}$, by the Masked Threshold variance; the optimal perceptual bit allocation and the corresponding perceptual coding gain are given by Equation (4.28) and Equation (4.30) respectively, with all σ^2_{xk} replaced by $\mathrm{w}_k\sigma^2_{xk}$ [Soman93][Wei97]. It should be noted that this is formally identical to the *frequency weighting* technique developed to improve[9] the subjective

[9] A special class of weighting functions were proposed by Tribolet and Crochiere [Crochi.79]: $\mathrm{w}_k = \sigma_k^{2\lambda}$, $-1 \leq \lambda \leq 0$. The parameter λ deviates progressively the spectral shape of the quantization noise from flat

coding quality of speech signals [Tribol.79][Crochi.79][Jayant84]. In this context, optimality means that the *weighted* noise variance is equal in each subband. In other words, the spectral profile of the quantization noise is *parallel* to the Masked Threshold. While this important conclusion is acceptable in practice when the quantization noise is below the Masked Threshold (perceptual overcoding), it is not too appropriate when the quantization noise is above the Threshold of Masking (perceptual undercoding). The concern in this case is to minimize the loudness of the unmasked quantization noise. In Chapter 6 we will present a few possible solutions to this problem.

4.3.2 Non-Uniform Filter Banks

4.3.2.1 The General Structure of Non-Uniform Filter Banks

Uniform filter banks are characterized by the fact that all subbands have equal bandwidth. This implies that the time/frequency resolution is the same for all subbands. Another implication, as illustrated in Figure 4.1, is that when the subbands are maximally decimated, the decimation and interpolation factors are the same for all subbands.

A non-uniform filter bank implements an unequal division of the signal spectrum so as to assign a different time/frequency resolution to different subbands. A non-uniform filter bank comprising N channels is illustrated in Figure 4.6. A resulting implication is that when the subbands are maximally decimated, the decimation and interpolation factors are the same for the same subband (or channel) but differ among different subbands. Similarly, the analysis and synthesis filters have the same bandwidth only for the same subband but differ among different subbands. These aspects can be related as follows. If channel k retains a fraction of the full bandwidth $[0,\ \pi]$ of a real signal given by $\mathrm{BW}(k)$, then, in order for all the subbands to cover the whole spectrum, it should be true that:

$$\sum_{k=0}^{N-1} \mathrm{BW}(k) = \pi \ . \tag{4.32}$$

If each subband is maximally decimated, then $\mathrm{BW}(k) = \pi/M_k$, which leads to:

$$\sum_{k=0}^{N-1} \frac{1}{M_k} = 1 \ . \tag{4.33}$$

($\lambda = 0$) to a shape parallel to that of the signal spectrum ($\lambda = -1$). Good subjective results were reported for $\lambda = -0.125$ [Tribol.79].

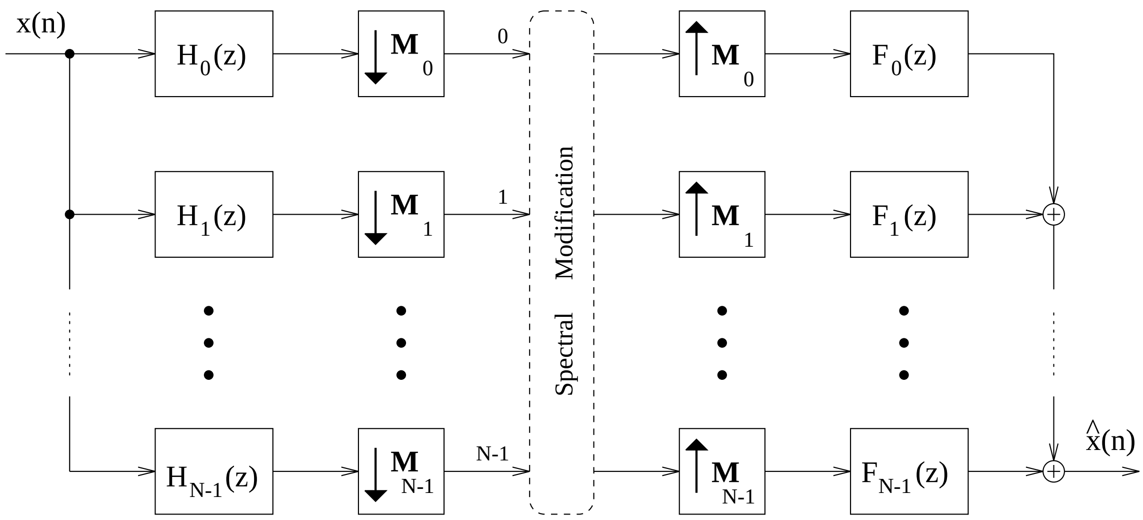

Figure 4.6: An non-uniform filter bank with N channels. M_k represents the maximal decimation/interpolation factor in channel k, $0 \leq k \leq N-1$ since the bandwidth of that (real) subband is π/M_k.

4.3.2.2 Objective and Perceptual Coding Gain

The classic measure of coding gain over PCM (4.27) derived for uniform filter banks (Sub-section 4.3.1), can be easily adapted to non-uniform filter banks. Similarly, a perceptual coding gain for non-uniform filter banks can also be established [Wei97]. In this Sub-section we will present the essential ideas of these formulations as well a number of related conclusions.

If the subband samples of a maximally decimated channel k (Figure 4.6) are coded (quantized) to R_k bits/sample, the resulting average bit rate of the encoded signal is

$$R = \sum_{k=0}^{N-1} \frac{R_k}{M_k}\ , \tag{4.34}$$

where the M_k values satisfy (4.33). If the error variance (due to quantization) in each channel is $\sigma_{qk}^2/M_k = \epsilon_k^2 2^{-2R_k}\sigma_{xk}^2/M_k$, a coding gain is obtained by minimizing $\sum_{k=0}^{N-1}\sigma_{qk}^2/M_k$, under the condition (4.34). σ_{xk}^2 represents the variance of the input signal in subband k, *before* the quantizer. The method of the Lagrange multipliers yields the optimum bit allocation [Soman93]:

$$R_k = R + \frac{1}{2}\frac{\sigma_{xk}^2}{\prod_{\ell=0}^{N-1}\left[\sigma_{x\ell}^2\right]^{1/M_\ell}}\ , \tag{4.35}$$

from which the coding gain over PCM can be derived:

$$G_{NUmax} = \frac{\sum_{k=0}^{N-1}\frac{\sigma_{xk}^2}{M_k}}{\prod_{\ell=0}^{N-1}\left[\sigma_{x\ell}^2\right]^{1/M_\ell}} = \frac{\sigma_x^2}{\prod_{\ell=0}^{N-1}\left[\sigma_{x\ell}^2\right]^{1/M_\ell}}\ . \tag{4.36}$$

These results indicate that, as for the uniform filter bank, the variances of the quantization errors are equal, *i.e.* $\sigma_{qk}^2/M_k = \text{constant}, \ \forall\ k$. On the other hand, under optimal bit allocation, the total reconstruction error is simply given by $\sigma_q^2 = \epsilon^2 2^{-2R} \prod_{\ell=0}^{N-1} \left[\sigma_{x\ell}^2\right]^{1/M_\ell}$.

It may be concluded that as the coding gain increases for higher values of N, the coding gain achieved by an uniform filter bank (for quasi-stationary signals) is higher than that associated to a non-uniform filter bank reaching a similar finest frequency resolution.

A perceptual coding gain can be easily obtained from the previous results by noting that instead of minimizing the noise variance, we want to minimize the noise variance weighted by the masked threshold, as we did in Sub-section 4.3.1.7 for the case of an uniform filter bank. In this context, optimality also presumes the quantization noise profile is parallel to the masked threshold. However, further investigation on the maximization of the perceptual coding gain [Wei97] indicates that sub-optimal results are obtained if the SMR in a subband is not white (which agrees with the theory that a spectral non-flatness elicits a coding gain). In this case, which is likely to occur particularly at high frequencies where the critical bandwidths are larger, the subbands should be split in order to increase the perceptual coding gain [Wei97]. Clearly, this necessity contradicts the critical band partition (Bark rate) of the audible spectrum. Paradoxically, this also suggests that an uniform filter bank is likely to be a better option, especially to the coding of quasi-stationary signals.

4.3.3 Multiresolution Analysis/Synthesis Systems

A clear difference should be made between a desired single (*i.e.* uniform) or multiple (*i.e.* non-uniform) time/frequency resolution in the decomposition of a signal; and the uniform or non-uniform filter banks used to decompose that signal. Multiresolution does not preclude uniform filter banks as we will illustrate next.

A non-uniform time/frequency resolution is often desired to capture the important details of stationary and non-stationary events. The former have a relatively long duration and are better analyzed (and processed) using a fine frequency resolution. The latter have a relatively large bandwidth and are better analyzed (and processed) using a fine temporal resolution because their localization in time is very important.

Thus, two possibilities can be conceived to obtain a non-uniform time/frequency resolution: dynamic switching between two (or more) uniform filter banks, or a fixed non-uniform filter bank. Figure 4.7 illustrates these concepts. An uniform time/frequency resolution is also illustrated as a reference (left side of Figure 4.7). This example denotes a signal decomposition by a 16 channel uniform filter bank. There is only one time scale and one frequency scale which are common

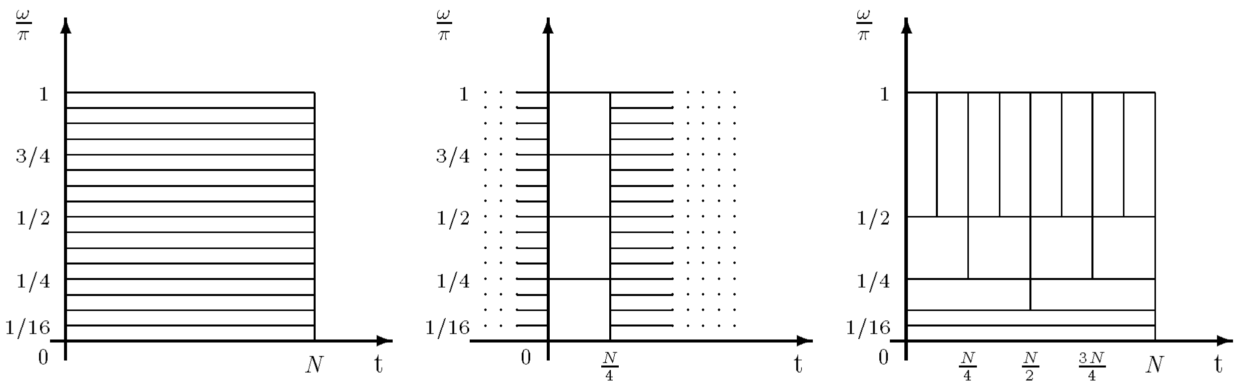

Figure 4.7: Uniform and non-uniform time/frequency division of a real signal. On the left, the bandwidth $[0,\ \pi]$ of the real signal is divided into 16 uniform subbands. On the center, a dynamic switching between two uniform filter banks is illustrated: the uniform filter bank represented on the left, and another uniform filter bank with four times finer time resolution (four times coarser frequency resolution). This configurations allows two time/frequency scales. On the right, a fixed non-uniform filter bank divides the signal bandwidth into five bands organized in four octaves. This configuration allows four time/frequency scales.

to all channels. If a dynamic switching is made between this uniform filter bank and another uniform filter bank with a different time/frequency resolution, a non-uniform analysis/synthesis system is obtained. The example on the center of Figure 4.7 illustrates the switching to a four times finer temporal resolution (which means a four times coarser frequency resolution) uniform filter bank. Thus, the resulting non-uniform analysis/synthesis system based on two uniform filter banks allows two different time/frequency resolutions.

The example on the left side of Figure 4.7 illustrates a fixed non-uniform decomposition of the signal spectrum into five bands whose bandwidths increase in octaves (the lowest two subbands have the same bandwidth). The time resolution varies in the same proportion but inversely to the frequency resolution[10] . It can be seen that this example of a five channel non-uniform decomposition gives rise to four different frequency scales as well as the corresponding time scales.

4.3.4 The Choice of the Multiresolution Scheme

A spectral partition of the audible spectrum according to the Critical Band rate (*i.e.* the *Bark* rate presented in Section 3.3.2) is certainly better approximated by a non-uniform filter bank.

[10]This is a consequence on the *uncertainty principle* which says that if a signal has a *wide support* in one domain (time or frequency), then the representation of that signal in its dual domain (frequency or time) has a *narrow support*. The most obvious example is a single pulse in the time domain (corresponding to a flat spectrum), or a single pulse in the frequency domain (corresponding in time to DC).

It would therefore appear to be a natural tool in perceptual coding of audio. There are, in fact, several examples of audio coding systems based on Wavelets[11] [Sinha93][Rioul91], the Hierarchical Lapped Transform[12] [Vargas93] and QMFs implementing a non-uniform spectral division. In most cases it consists in a tree structure of (fractional) octave-band splitting. However, the implicit assumption when using such a multiresolution approach is that, *most frequently*, low frequency signal components have long durations and high frequency signal components have short durations. We believe this assumption is not entirely applicable to audio signals since low frequency signals can have very short durations, and high frequency signals can have very long durations. This same perspective was presented by Malvar [Malvar92, page 243] to justify the 'disappointing' coding performance of speech signals using wavelets: *'The main problem with speech signals is that we need fine frequency resolution not only at low-frequency subbands, but also at some of the middle-frequency subbands, in order to resolve the formant structure'*.

Furthermore, the *Bark* rate does not indicate that the human auditory system has *permanently* a finer time resolution at high frequencies than at low frequencies, in correspondence to the way critical bandwidths get larger towards higher center frequencies. Instead, it does indicate that the frequency resolution associated to *simple* masking experiments, is not uniform. However, as we indicated in Section 3.3.6, this may not hold exactly for harmonic signals. Thus, the critical bandwidths may not reflect directly the time resolution of the human auditory system. In addition, as we concluded in Section 3.3.2, there are indications that for very short stimuli, *all* critical bands get wider as in a kind of dynamic switching between different time/frequency resolution modes.

As a consequence, we believe a more representative and useful classification of audio in a perceptual sense, is between quasi-stationary or non-stationary regions of the audio signal. The former is associated to a slowly varying acoustic waveform, and the latter is associated to very short (*i.e.* less than about 50 *ms* long) acoustic events, or to sharp transitions between quasi-stationary regions. Both quasi-stationary and non-stationary regions may include low and high frequency components. The emphasis is therefore placed on the comparable time resolution of all signal components. This is the main reason why we prefer to address signal stationarity by using dynamic (signal adaptive) switching between two (or several) uniform filter banks with different time/frequency resolutions. Obviously, the performance of this approach depends directly on the appropriate switching decisions. Ultimately, these carry the responsibility of maximizing the coding gain without degrading the coding quality due to incorrect switching decisions.

[11]Wavelets are basis functions obtained from a single prototype wavelet by dilatation and translation.

[12]The Hierarchical Lapped Transform is a non-uniform PR filter bank based on a hierarchical structure of lapped transforms (*e.g.* the MLT) [Malvar92].

To test this perspective, we used the test items described in Section 3.3.6 to study the occurrence of clear transitions between quasi-stationary and non-stationary regions of audio signals. The results are presented in Table 4.1 and indicate the number of probable transitions. The underlying detection algorithm, which will be detailed in Section 6.2.2, makes a signal analysis on audio segments, approximately 12 *ms* long. The total number of segments in each test item is also indicated in Table 4.1. A review of the detailed description of each test item

Table 4.1: Detection of transitions between quasi-stationary regions and non-stationary regions of a few test items. The number of transitions indicated is relative to the total number of processed segments (each one being about 12 *ms* long) which reflects the duration of each test item.

item	**castanets**	**female vocal**	**harpsichord**	**pitchpipe**
segments	202	443	444	519
transitions	15	16	10	5

(Section 3.3.6) helps to understand these results. The most significant results are those obtained for *castanets* and *harpsichord* as we explain next.

- For *castanets*, the results reflect *exactly* the number of strongly non-stationary occurrences similar to that displayed in Figure 3.8. For *harpsichord*, the results reflect *exactly* the number of played (ascending) notes. Using our interactive simulation environment (Section 2.3.3), we could confirm very easily that the transitions were accurately signaled, within the time resolution of the algorithm, at the onset of each castanet attack, or at the onset of each harpsichord note.

- Not surprisingly, the largest number of transitions was detected for *female vocal* as a result of the frequent voiced/unvoiced transitions.

- The smallest number of transitions was detected for *pitchpipe* because this is the smoothest of all test items: it contains only three different musical notes, each one about 1.6 *sec.* long.

Overall, these results validate our perspective that non-stationary regions represent a very small fraction of the duration of a typical audio signal. In addition, combining these results with those presented in Figure 3.9 (section 3.3.6), we may conclude that a typical audio signal exhibits predominantly quasi-stationary regions, most of which have a well defined harmonic profile, whose analysis coding and post-processing is more appropriately performed through an *uniform* filter bank. These conclusions suggest that a suitable analysis/synthesis for coding and post-processing should be primarily based on a fine frequency resolution uniform filter bank enabling high coding gains. Additionally, a possibility should be provided, when needed, to switch to a fine time resolution filter bank which could be uniform [Johns.92] or non-uniform [Sinha96]. A few recent papers share this perspective, *i.e.* using permanently a fixed non-uniform filter bank to code audio signals, leads to a much lower coding gain than otherwise possible [Wei97][Sinha96][Princen95]. Three final compelling reasons regarding critical aspects associated to non-uniform filter banks, corroborate our interest in (and preference for) uniform filter banks:

1. a detailed harmonic analysis and processing is not possible using non-uniform filter banks,

2. a non-uniform filter bank might be a strong inhibitory factor of high coding gains in a conventional sense [Jayant84] given the predictable necessity of a large amount of side information needed to characterize the diversity of time and frequency scales,

3. non-uniform filter banks are usually implemented using a cascade structure of uniform or hybrid filter banks, which degrades seriously the channel separation by promoting channel crosstalk and by limiting the stop band attenuation. This has been referred to as the *aliasing build-up* problem [Malvar92, page 224]. This problem exists for example in the MPEG-1 Layer 3 audio coding standard [ISO93] and has been attenuated by means of an "aliasing reduction" processing [Sporer92]. On the other hand, a cascade of decomposition stages tends to increase the processing delay beyond that implied by an uniform filter bank achieving the finest frequency resolution of the non-uniform filter bank.

4.4 Coherence Detection Using an Uniform Filter Bank

In Section 3.3.4 we indicated that in the context of frequency domain audio coding, the masker is the input audio signal, whether involving coherent (tone-like) or incoherent (noise-like) components, and the maskee is the quantization noise. The ultimate goal of a perceptual coder is to generate a good estimate for the profile of the quantization noise that does not cause noticeable impairments when actually added to the original signal. In other words, this noise profile, also called Threshold of Masking [Johns.88b] or Just Noticeable Distorsion [Jayant93], should be optimally shaped in frequency, time and space (in the case of stereo or multichannel association of correlated audio signals). Several studies [Hellman72][Moore89] reveal that the Threshold of Masking may vary substantially as a function of the noise-like or tone-like nature (tonality) of the audio signals. This has also been corroborated by our results presented in Figure 3.7. As a consequence, this aspect has a significant influence on the quality and efficiency of the audio encoding process.

The purpose of this section is to focus on methods for tonality classification of audio signals. After reviewing briefly a few algorithms for tonality classification used in known audio coders, we will propose a new method for tonality classification that is adapted to the cosine modulated TDAC filter bank. A few results will illustrate the performance of the technique.

4.4.1 Common Tonality Classification Methods

As we detailed in Section 3.3.4, there are two main results from psychoacoustical studies that relate the amount of masking of one sound (maskee) due to another sound (masker), given their tonality properties. These results are represented by the known expressions for tone-masking-noise and noise-masking-tone thresholds (the latter is normally considered as a good substitute for the case of noise-masking-noise). These expressions differ quite substantially because, in general, a narrow band of noise masks a tone much more effectively than a tone masks a narrow band of noise [Hellman72]. In fact, this difference, which in general depends on the frequency, may be near 10 dB. This means that a particular signal can be quantized using more or less bits according to its tonality properties. This evaluation must therefore be as accurate as possible. Several coders (*e.g.* AC-2) avoid this evaluation by sacrificing performance but gaining in global algorithmic simplicity with obvious advantages for real-time implementation. The solution in this case is to consider some average but eminently conservative masking requirements, in order to prevent worst cases.

The tonality properties of an audio signal are normally evaluated by taking a short segment of audio samples (*e.g.* 512 or 1024 samples) and by making a spectral analysis, using for example a Fast Fourier Transform (FFT). Looking at the power and the phase evolution of each spectral component, it may be possible to infer a tonal behavior of specific regions of the spectrum. This operation depends obviously on the selectivity of the analysis window.

Johnston [Johns.88b] proposes the computation of an average tonality measure for each segment of data using the Spectral Flatness Measure[13] (SFM). After being converted to Decibels (SFM_{dB}), and considering that a reference of $SFM_{dBmax} = -60dB$ denotes a tone-like signal, the SFM_{dB} is converted to a tonality coefficient α whose range of possible values is [0, 1]. The lower value indicates a global noise-like behavior, and the higher value indicates a global tone-like behavior. The tonality coefficient, α, is then used to interpolate between a noise-masking model (implying a lower quality coding) and a tone-masking model (implying a higher coding quality), as in the case of Equation 3.4.

[13] As we have seen in Sub-section 4.3.1.6, the SFM is defined as the ratio of the geometric mean of the power spectrum to the arithmetic mean of the power spectrum [Jayant84].

It was later recognized [Brand.90] that even if the above tonality measure estimates correctly the significant tonal parts of an audio signal, it leads to very conservative Masking Thresholds at non-tonal frequencies, thus requiring an excessive bit rate. Instead of a global tonality coefficient per segment of audio data, a better approach is to identify tonal spectral regions and non-tonal spectral regions within the same segment of audio data. This local tonality measure may be computed on a spectral coefficient basis but, in practice, it is usually computed on a Bark (critical band) related scale. According to this perspective, Brandenburg and Johnston [Brand.90] suggested that the Euclidian distance between the actual and predicted values (based on the previous two occurrences), for the amplitude and phase of each spectral phasor, could be converted into a tonality metric by a logarithmic operation, and used to interpolate between a noise masking model and a tone masking model as above. It was reported [Brand.90] that this solution gives a better performance than the previous method.

This latter solution corresponds to the basic model for tonality classification of the Psychoacoustic Model 2 of the MPEG-Audio Standard [ISO93]. The normalized Euclidian distance is converted to a weighted unpredictability in each coder partition (about 1/3 Bark), and is converted to a tonality index using a simple logarithmic rule. The tonality index is then used as above to interpolate between the two masking models on a quantization-band by quantization-band basis. Psychoacoustic Model 2 is especially recommended for Layer 3 of the MPEG-Audio Standard.

Psychoacoustic Model 1 of the same Standard, which is especially recommended for Layer I and II of the algorithm MPEG-Audio, also makes use of tonality information although it is derived in a different way. After the FFT computation, all local maxima of the spectrum are identified. A local maximum must have at least 7 dB more energy than neighboring spectral lines. Then, these tonal components are decimated: a tonal component at less than 0.5 $Bark$ of a stronger tonal component is eliminated as well as any tonal component with energy below the Threshold in Quiet. The energy of a number of spectral components surrounding each surviving tonal maximum, is summed to represent each main tonal component. The remaining spectral lines are classified as non-tonal (noise) components and their energies are summed within the bandwidth of each critical band.

In Section 3.3.4 we have shown that the two Psychoacoustic Models of the MPEG-Audio algorithm are very similar as far as the Masking Threshold calculation is concerned. However, at the tonality classification level, the main difference is that, while for Psychoacoustic Model 1, the tonality information is binary (each spectral component may be noise-like (0) or tone-like (1)), for Psychoacoustic Model 2, the tonality classification is continuous in the interval [0, 1], *i.e.* a spectral component is allowed to have an intermediary behavior between noise and tone.

4.4.2 Tonality Detection Using the TDAC Filter Bank

Even if a tonality classification method similar to that of the MPEG-Audio Psychoacoustic Model 2 might be considered more flexible and, in average, more likely to give reliable tonality estimations, several problems may result when a frequency domain coder based on the TDAC filter bank must code very different types of audio material. For example, music tends to have long stationary regions, in the order of 100 ms and more, while speech tends to have much shorter stationary regions (a reasonable duration -syllabic duration- is about 10-20 ms [Rabiner78]).

Considering a 50% overlap TDAC filter bank whose time window is 1024 samples long, the phasor predictor of the Psychoacoustic Model 2 implies a delay of 2048 samples to produce tonality information concerning any segment of audio data. For a sampling frequency of 48 *kHz*, this is equivalent to about 43 *ms* of audio data which is an excessive length considering the shorter stationary tonal regions of speech and music signals. This example shows that transform "hops" of about 21.3 *ms* or more may not be appropriate to search for the power and phase coherence of a spectral phasor. In fact, important stationary tonal regions may be simply *missed*.

Instead of an *intersegment* based tonality detection criterion such as the one of the previous example, a probably better approach is to look for an *intrasegment* based tonality detection method, which could include some additional mechanism of validation between segments (*e.g.* an exponential smoothing).

According to this point of view, one possibility is to have a form of spectral power peak detection as considered in the MPEG-Audio Psychoacoustic Model 1 [Brand.92b]. Another possibility is to look for some phase relation (within one segment) indicating the existence of spectral tonal components. This idea is in fact the basis of a new model for tonality detection [Ferre.95b] which we will explain next.

It is assumed that a 50% overlap TDAC filter bank is used for coding and a complex extension, X_{TDAC}, is considered for phase extraction. The oddly stacked version of the TDAC filter bank (which has equally spaced filters of equal bandwidth) [Princen87] has the following expression for the direct transform[14]:

$$X(k) = \sum_{n=0}^{N-1} h(n)x(n)\cos\left[\frac{2\pi}{N}\left(k+\frac{1}{2}\right)\left(n+\frac{1}{2}+\frac{N}{4}\right)\right] , \quad 0 \le k \le N/2-1 , \tag{4.37}$$

which makes use of the analysis basis functions of Equation (4.22), by taking $L = N = 2M$. $x(n)$ is the input signal and $h(n)$ represents the analysis window. This transformation can be

[14] To be accurate, we should include a block (or segment) index $-\infty < m < \infty$ when representing the signal in time, $x(n,m) = x(n + m\frac{N}{2})$, or in frequency, $X(k,m)$, in order to make clear that the analysis/synthesis is made on a segment basis and in order to explicit the 50% overlap between successive segments. Since we are only concerned in this sub-section with the analysis of an arbitrary segment, we will omit this index for simplicity.

expressed as:

$$X(k) = \Re\left\{X_{TDAC}(k)\right\} = \Re\left\{\sum_{n=0}^{N-1} h(n)x(n)e^{-j\frac{2\pi}{N}(k+\frac{1}{2})(n+\frac{1}{2}+\frac{N}{4})}\right\}\ , \quad 0 \le k \le N/2-1\ , \tag{4.38}$$

and further as:

$$X(k) = \Re\left\{e^{-j\frac{2\pi}{N}(k+\frac{1}{2})(\frac{1}{2}+\frac{N}{4})}\sum_{n=0}^{N-1} h(n)x(n)e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n}\right\}\ , \quad 0 \le k \le N/2-1\ . \tag{4.39}$$

$\Re\{C\}$ denotes the real part of the complex number C. We can see that it is possible to extract phase information by computing the Odd-DFT [Bellan.89] of the windowed signal, and then by adding a constant phase term for each spectral phasor k. Our objective is to investigate some particular phase property for pure sinusoidal signals using first the Odd-DFT and then the complex TDAC.

We will first consider a simple case in which the frequency of a sinusoidal input corresponds exactly to a sampled frequency, *i.e.* taking $\Delta\ell = 0.0$:

$$\ell + \Delta\ell = N\frac{f_i}{f_s}\ , \tag{4.40}$$

$$\begin{array}{rcl} & \ell & integer\ , \\ -0.5 \le & \Delta\ell & \le 0.5\ , \\ & f_i & < \dfrac{f_s}{2}\ . \end{array}$$

N represents the length of the transform, f_s is the sampling frequency and f_i is the input frequency. ℓ is the integer index of a transform coefficient. A general expression for the sinusoidal input is given by:

$$x(n) = V\sin\left[\frac{2\pi}{N}(\ell+\Delta\ell)\,n + \phi\right]\ . \tag{4.41}$$

The Odd-DFT coefficients ($X_{ODFT}(k)$) are defined by

$$X_{ODFT}(k) = \sum_{n=0}^{N-1} x(n)h(n)e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n}\ , \tag{4.42}$$

where $h(n)$ represents the analysis sine window (4.23), which we repeat here in a more general form for convenience:

$$h(n) = H\sin\left[\frac{2\pi}{2N}\left(n+\frac{1}{2}\right)\right]\ . \tag{4.43}$$

Recognizing the following identity:

$$\sin(a)\sin(b) = \frac{1}{4}\left[e^{-j(a-b)} + e^{j(a-b)} - e^{-j(a+b)} - e^{j(a+b)}\right]\ ,$$

and making the substitution of Equations (4.41) and (4.43) in Equation (4.42), we obtain

$$\begin{aligned} X_{ODFT}(k) = \sum_{n=0}^{N-1} \frac{VH}{4} & \left[e^{-j\frac{2\pi}{N}(\ell-\frac{1}{2})n} e^{j\left(\frac{2\pi}{4N}-\phi\right)} + e^{j\frac{2\pi}{N}(\ell-\frac{1}{2})n} e^{-j\left(\frac{2\pi}{4N}-\phi\right)} \right. \\ & \left. - e^{-j\frac{2\pi}{N}(\ell+\frac{1}{2})n} e^{-j\left(\frac{2\pi}{4N}+\phi\right)} - e^{j\frac{2\pi}{N}(\ell+\frac{1}{2})n} e^{j\left(\frac{2\pi}{4N}+\phi\right)} \right] e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n} , \end{aligned} \tag{4.44}$$

that may be simplified to

$$\begin{aligned} X_{ODFT}(k) = \frac{NVH}{4} & \left[e^{-j\left(\frac{2\pi}{4N}-\phi\right)} \delta(k-N+\ell) + e^{j\left(\frac{2\pi}{4N}-\phi\right)} \delta(k-\ell+1) \right. \\ & \left. - e^{-j\left(\frac{2\pi}{4N}+\phi\right)} \delta(k-N+\ell+1) - e^{j\left(\frac{2\pi}{4N}+\phi\right)} \delta(k-\ell) \right] . \end{aligned} \tag{4.45}$$

$\delta()$ represents the Kronecker delta function.

Considering only the frequency interval $[0,\ \pi]$, it can be seen that for a single input pure tone, we have two spectral lines, located at indexes $k = \ell$ and $k = \ell - 1$, with the same magnitude. It is thus clear that the sine window $h(n)$ is responsible for a "double side band modulation" (suppressed carrier) effect that maps a pure input sinusoid in at least two consecutive frequency lines of the transform, whose phases are related. In fact, the phase difference between the two side bands of the Odd-DFT is exactly $\pi\left(\frac{1}{N} \pm 1\right)$.

The relevant conclusion at this point is that the initial phase ϕ is canceled which makes the phase difference a constant value. It may also be seen why the sine window given by Equation (4.43) is very popular: if $\ell = 0$ (*i.e.* for a DC signal), two of the above Diracs overlap with the remaining two at $k = 0$ and $k = N - 1$. Considering that the input signal is real, this means that, in practice, the DC component of one signal is entirely compacted on the first spectral line, in the range on unique coefficients ($[0,\ \frac{N}{2} - 1]$) of the Odd-DFT and hence, of the TDAC filter bank.

If the remaining phase components necessary to obtain the complex TDAC coefficients are considered (as in Equation (4.39)), it can be concluded that there is an additional contribution to the phase difference between any two consecutive spectral lines : $-\frac{\pi}{N}\left(1 + \frac{N}{2}\right)$. Summing this value to the previous one, we obtain $-\frac{\pi}{2} \pm \pi$ which leads to:

$$\angle\{X_{TDAC}(\ell)\} - \angle\{X_{TDAC}(\ell-1)\} = \frac{\pi}{2} , \tag{4.46}$$

where $\angle\{X\}$ denotes the phase of X. The obvious question now is: what happens if $\Delta\ell$ is not 0.0 as we have assumed ? In this case, the energy of the single tone input will spread to other frequency lines than ℓ and $\ell - 1$ in the interval $[0,\ \pi]$ of the transform. However, because $-0.5 \leq \Delta\ell \leq 0.5$, the identity (4.46) will be affected by a small error ϵ becoming:

$$\angle\{X_{TDAC}(\ell)\} - \angle\{X_{TDAC}(\ell-1)\} = \frac{\pi}{2} \pm \epsilon . \tag{4.47}$$

The maximum value of $|\epsilon|$ is an issue that can be appropriately evaluated[15] with practical tests for a few representative values of $\Delta\ell$, as we will illustrate later. Assuming that ϵ is bounded and very small, then

$$\cot\left(\frac{\pi}{2} \pm \epsilon\right) \approx \mp\epsilon \tag{4.48}$$

will hold approximately. According to this result, we may define the phase condition for two arbitrary consecutive spectral lines k and $k-1$ as:

$$-\epsilon \leq \cot\left[\angle\left\{X_{TDAC}(k)\right\} - \angle\left\{X_{TDAC}(k-1)\right\}\right] \leq \epsilon\,, \tag{4.49}$$

from which we can derive:

$$-\epsilon \leq \frac{\Re\left\{X_{TDAC}(k)\right\}\Re\left\{X_{TDAC}(k-1)\right\} + \Im\left\{X_{TDAC}(k)\right\}\Im\left\{X_{TDAC}(k-1)\right\}}{\Re\left\{X_{TDAC}(k-1)\right\}\Im\left\{X_{TDAC}(k)\right\} - \Re\left\{X_{TDAC}(k)\right\}\Im\left\{X_{TDAC}(k-1)\right\}} \leq \epsilon\,. \tag{4.50}$$

Taking into consideration the difference in periodicity between the *cotangent* and the *sine* functions, it is necessary to ensure that $\sin\left[\angle\left\{X_{TDAC}(k)\right\} - \angle\left\{X_{TDAC}(k-1)\right\}\right] > 0.0$, which leads to an additional condition:

$$\Re\left\{X_{TDAC}(k-1)\right\}\Im\left\{X_{TDAC}(k)\right\} > \Re\left\{X_{TDAC}(k)\right\}\Im\left\{X_{TDAC}(k-1)\right\}\,. \tag{4.51}$$

$\Im\{C\}$ denotes the imaginary part of the complex number C. Conditions (4.51) and (4.50) have been tested for different values of $\Delta\ell$ and N. Different combinations of pure tones with unequal amplitudes have also been tried. We have concluded that ϵ can be bounded to 0.05 $(\approx \frac{\pi}{63})$ for "large" values of N (≈ 2048) and to 0.1 $(\approx \frac{\pi}{31.5})$ for "small" values of N (≈ 512). We have also noticed that when the input signals consist mainly of low level noise, then condition (4.50) tends to be verified in a rather random way. To avoid spurious indications, a third condition was also included to test the relative magnitudes between the detected spectral component k and the two smaller adjacent spectral lines:

$$10\log\frac{\left[X_{TDAC}(k)\right]^2}{0.4\left[X_{TDAC}(i)\right]^2 + 0.6\left[X_{TDAC}(j)\right]^2} > 10.0dB\,, \tag{4.52}$$

$$\begin{array}{rcl} & |X_{TDAC}(i)| & < |X_{TDAC}(j)|\,, \\ k-2 \leq & i,j & < k\,, \\ k < & i,j & \leq k+2\,. \end{array}$$

The resulting algorithm gives, by its nature, binary answers: every spectral line is tested and the possibilities are that a particular line is considered to have a tonal behavior (index 1.0) or not (index 0.0). In order to avoid rapid changes in tonality causing eventually coding artifacts, an exponential smoothing may be applied to each frequency line, *e.g.* by means of an IIR filter.

[15]A more rigorous analysis of this problem will be presented in Chapter 7. Here, we are mainly interested in an approximate solution.

4.4.3 A Few Results

The model of tonality detection described in the previous sub-section has been tested with both natural a synthesized sounds. We report below two particular examples involving a synthesized signal and a female speech signal.

The synthesized signal comprises four tones whose frequencies are 58.59375 Hz, 2343.75 Hz, 4318.5 Hz, and 7042.96875 Hz. Taking $f_s = 48000$ Hz and $N = 2048$ samples, two tones verify $\Delta\ell = 0.0$ and the remaining two tones verify $\Delta\ell \neq 0.0$. The synthesized signal has a total duration of 130 ms and is smoothed at rise and fall by a sine function during 30 ms. The complete analysis of the signal can be appreciated in Figure 4.8. The left columns refer to the time segments, the center ones correspond to the accumulated spectral energy in each band of 1/3 $Bark$, and the columns on the right-hand side correspond to the tonality evaluation in the same frequency scale. It can be verified that the evolution in time (successive rows in Figure 4.8) has an implicit shift of $N/2$ samples. On the other hand, it is clear that the tonality indication only appears when the observed time segment includes a fairly stationary region of the signal. In fact, in the rise and fall regions, the tonality indication of the tones is not confirmed, only reminiscent values persist during the fall region due to the exponential smoothing. For certain applications, this might correspond to a desirable property of the model.

The illustrated exponential smoothing acts as a corroboration mechanism as can be seen after the third time segment. It can also be seen that tonality indications in the high frequency region rise and fall more rapidly than in the low frequency region. This different sensitivity (time constants) of the exponential smoothing complies with the expected higher spectral dynamics of the high frequency contents of real world signals (speech and music).

A female speech sound (Suzan Vega) has also been used to test the model. Figure 4.9 depicts the transition from a voiced sound to an unvoiced one. This transition corresponds to the phonetic units: "er's" that were taken from the words: "... ther's a ...". It can be seen that there is a consistent indication of tonality according to the observed speech segments. In this case, the model was modified to give a tonality index proportional to the number of tonal spectral lines detected within each band of $1/3Bark$. This is important because the Bark scale is non-linear [Zwicker61], involving a small number of spectral lines (or even one line only) in the low frequency region but a large number of spectral lines in the high frequency region.

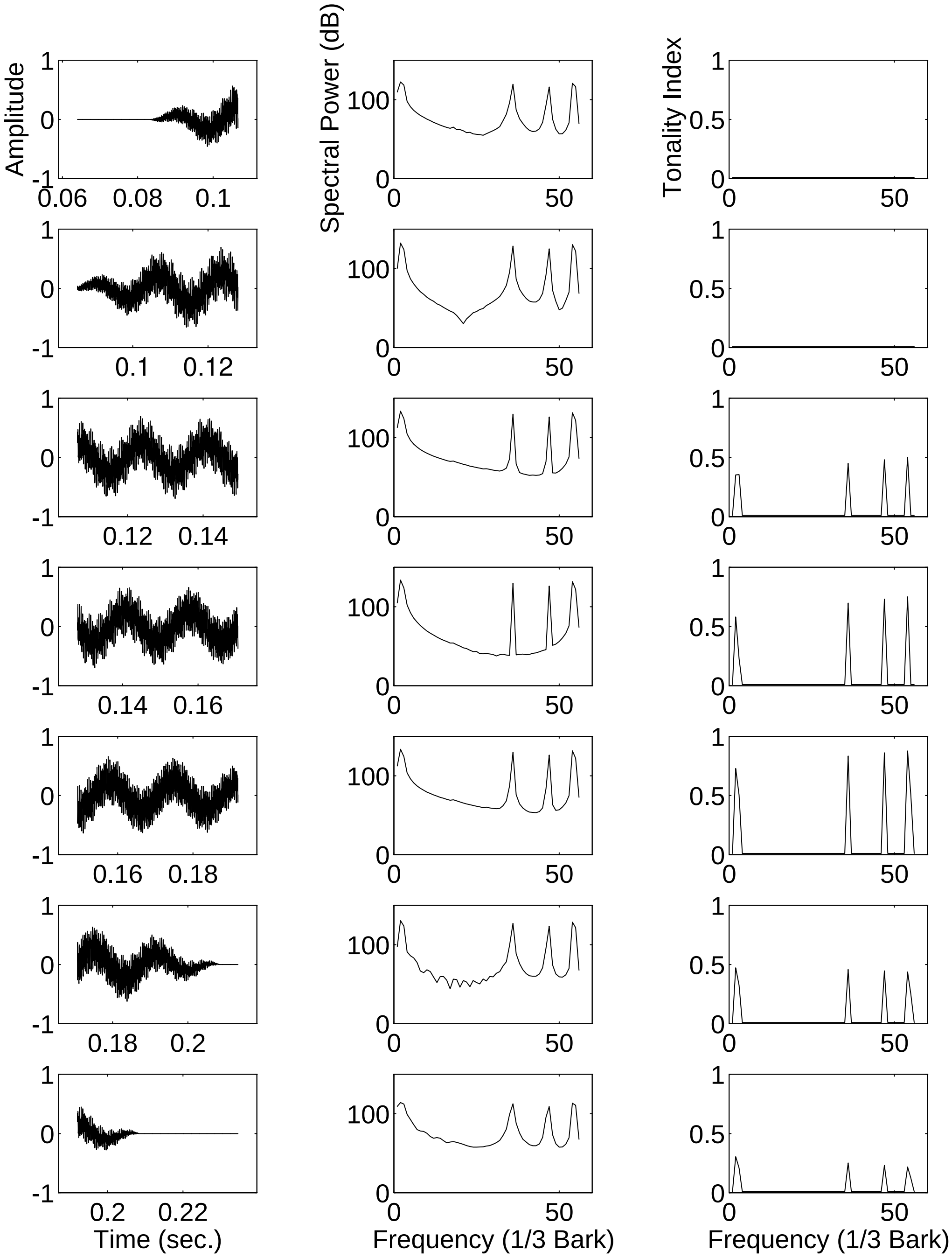

Figure 4.8: Tonality estimation for a synthesized signal comprising four pure tones. The time representation (on the left) denotes the 50% overlap between adjacent segments of data. Tonality is detected only for three segments where the windowed signal is clearly stationary. The reminiscent tonality indexes for the last two segments are due to exponential smoothing.

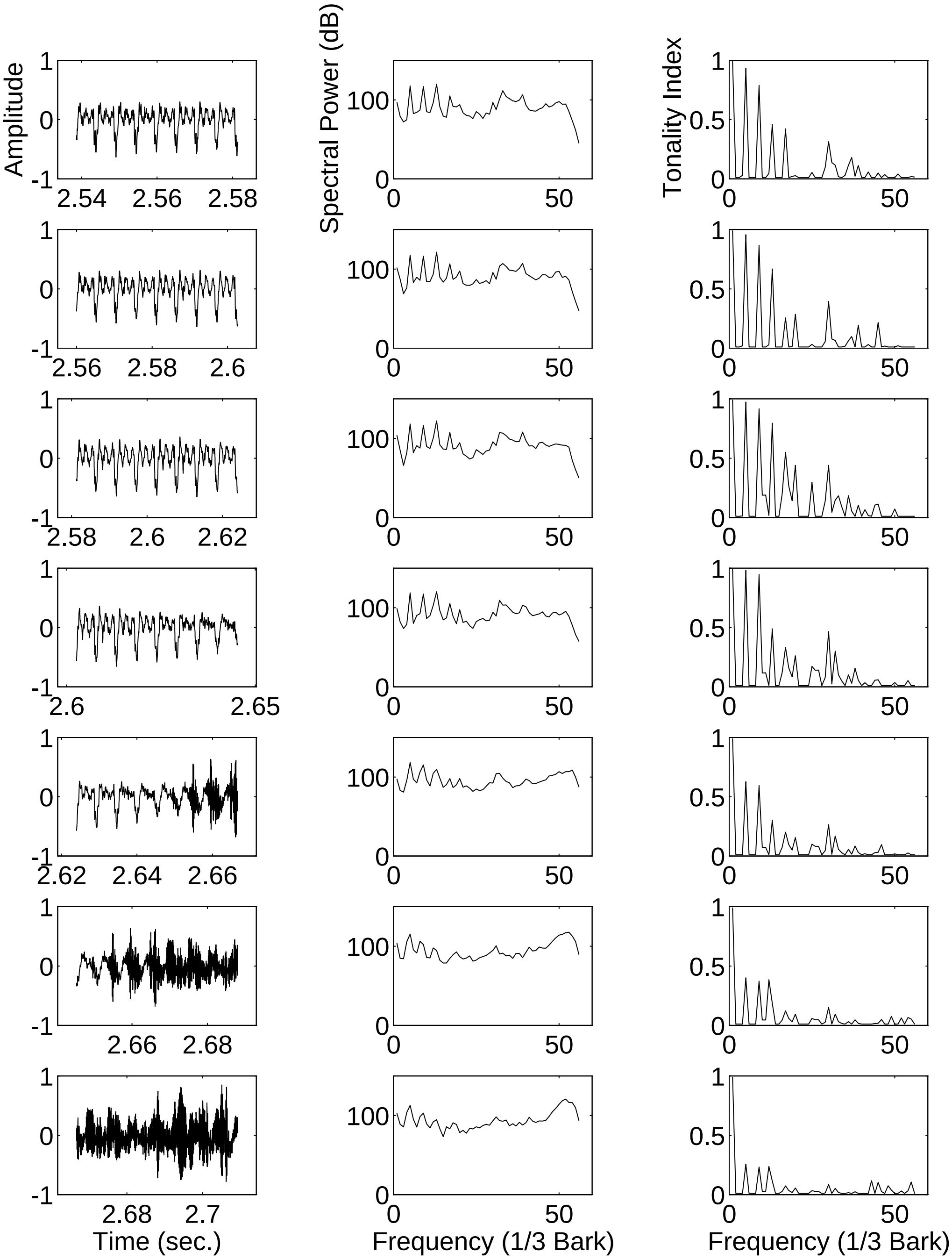

Figure 4.9: Tonality detection for a speech signal (female vocal). A transition from a voiced region to an unvoiced region is clearly reflected on the estimated tonality index.

4.5 Spectral Modification Effects and the Design of the Prototype Filter

The problem of optimum design of PR cosine modulated filter banks is largely simplified by the fact that all analysis and synthesis filters are derived from a prototype filter by modulation. Thus, it suffices to optimize the design of the prototype filter according to a meaningful criterion. In the absence of spectral modification, the most usual criterion compatible with the PR requirement has been the minimization of the stop-band leakage of the prototype filter (Sub-section 4.3.1.5).

In the presence of a spectral modification such as equalization or quantization, the PR property is disrupted because uncancelled aliased terms arise at the output, as well as new error contributions resulting from the multiplicative and additive spectral modifications [Jakob95]. Multiplicative spectral modifications are associated to signal correlated errors which are perceptually disturbing. Additive spectral modifications are associated to signal uncorrelated errors which have lower perceptual effects than the previous ones. Multiplicative errors result typically from explicit spectral filtering or coarse quantization. Additive errors result typically from small interferences or fine quantization.

Several methods have been devised to attenuate the distortions appearing at the output of the synthesis filter bank. In some cases, the PR property is relaxed in order to allow the optimization of synthesis filters and quantizers that minimize a perceptual distortion measure [Gosse97], or in order to insert a compensation filter bank that reduces aliasing while preserving the time invariant part of the original system [Heinle97]. In other cases, the PR property is maintained while the attempt is made to trade-off between signal correlated and uncorrelated errors [Jakob96] or, in the limit, to eliminate the correlated errors at the cost of increasing the uncorrelated errors [Uzun95].

We have taken a different approach to the optimum design of the analysis/synthesis prototype filter, in the sense that it elicits the reduction of undesirable distortions appearing as a result of a spectral modification.

As we are interested in spectral filtering (equalization), and following our observation that under very low bit rate constraints, the quantization operation is rather coarse, which corresponds essentially to a multiplicative spectral modification; we study in detail the associated circular convolution operation in the time domain. We identify the resulting linear filtering term and the resulting aliased terms. Then, we establish a procedure for the optimal design of an analysis/synthesis prototype filter offering a desired trade-off between the reduction of the aliased terms and the reduction of the stop-band leakage of the prototype filter. Throughout this study, we take the sine window (4.23) as a reference. We also refer to the analysis/synthesis prototype filter as the analysis/synthesis *time window*.

4.5.1 A Model of Spectral Modification

A signal compression gain or bit rate reduction can be obtained by means of a suitable quantization of the spectral coefficients $X(k)$, at the output of a maximally decimated filter bank. The quantization operation is a spectral modification of the signal that may, in general, be modeled as the result of a linear operation:

$$\tilde{X}_m(k) = D_m(k)X_m(k) + A_m(k) \; , \tag{4.53}$$

involving a multiplicative term $D(k)$ and an additive term $A(k)$, on a particular time segment m.

When the bit rate is high enough, the quantization process induces only light modifications in the spectrum such that $D_m(k)$ may be assumed to be a constant for any m and k, and $A_m(k)$ may be assumed to have a very small value (relative to $X_m(k)$) and essentially a random behavior. Under these assumptions, the spectral modifications are mainly additive, which leads to a corresponding additive noise component in the time domain [Rabiner78]. This assumption is implicit in the coding gain analysis detailed in Sub-section 4.3.1.6 and in Sub-section 4.3.1.7. However, under very low bit rate coding conditions, the coarse quantization of the spectral coefficients is inevitable [Ferre.93]. Very often, clusters of adjacent coefficients, corresponding in some cases to entire spectral subbands, particularly at high frequencies, are set to zero. This is consistent with the observation that as the bit rate goes down the more the coded signal is band-limited.

Practical results have shown that under very low bit rate constraints, the $D_m(k)$ values vary substantially with k within a particular time segment m, but this variation remains practically unchanged or varies very slowly with m, particularly in stationary regions of the audio signal, during significant periods of time. As $D_m(k)$ has, in general, the dominant effect on the final coding quality, the additive modification $A(k)$ will be ignored. Thus, the nature of the quantization error is essentially multiplicative, which corresponds to a filtering operation. As a consequence, because of the repetitive nature of the sampled spectral representation [Oppenh.75], the reconstructed time segment will reflect the circular convolution between the time domain representations of $D_m(k)$ and $X_m(k)$. If special precautions are not taken, significant amounts of aliased components may seriously degrade the performance of frequency domain coders operating at very low bit rates. In fact, all the efforts devoted to computing the most suitable noise profile, and to shaping it properly through a carefully planned quantization operation, can be made completely unconsequent due to the circular convolution noise terms.

In order to design a suitable analysis/synthesis prototype filter minimizing the aliased components arising due to a spectral multiplicative modification, we identify first their nature and dependencies.

We outline now the structure of this section. In Sub-section 4.5.2 we review the TDAC filter bank and the associated time window constraints for the case of 50% overlap. We show in Sub-section 4.5.3 a few simple examples of windows that meet the TDAC constraints and we present a number of additional desired properties that should be met by the time windows from the perceptual coding point of view. In Sub-section 4.5.4 we address the difficulties in designing the prototype filter. We identify in Sub-section 4.5.5 all time components that result from the circular convolution operation when the Odd-DFT is used in the same *overlap-add* (OA) configuration of the TDAC filter bank. A simple band-pass filter $D_m(k)$ that guarantees the existence of aliased terms will be used. In Sub-section 4.5.6 the unwanted aliased terms are isolated, an optimization procedure is established and implemented to minimize their effects, and the results are discussed. It is shown in Sub-section 4.5.7 that the TDAC conditions include those considered for the Odd-DFT, and that both sets of conditions lead essentially to the same conclusions.

Since the time domain aliased components are the focus of this section, practically all analytical considerations will be referred to this domain.

4.5.2 The TDAC Filter Bank

The direct transformation of the oddly stacked TDAC filter bank is given by Equation (4.37), and the inverse transformation is given[16] by:

$$y(n) = \frac{f(n)}{N} \sum_{n=0}^{N-1} X(k) \cos\left[\frac{2\pi}{N}\left(k+\frac{1}{2}\right)\left(n+\frac{1}{2}+\frac{N}{4}\right)\right] , \quad 0 \le n \le N-1 . \qquad (4.54)$$

The analysis and synthesis time windows are respectively represented by $h(n)$ and $f(n)$. If Equation (4.37) is substituted in Equation (4.54), it can be proved [Ferre.92] that the reconstructed segment, without considering the 50% overlap and add operation, results in

$$y(n) = \frac{1}{2} f(n) \left[h(n)x(n) - h(\frac{N}{2}-1-n)x(\frac{N}{2}-1-n)\right] , \quad 0 \le n \le \frac{N}{2}-1 \qquad (4.55)$$

for the first half of the segment, and results in:

$$y(n) = \frac{1}{2} f(n) \left[h(n)x(n) + h(\frac{3N}{2}-1-n)x(\frac{3N}{2}-1-n)\right] , \quad \frac{N}{2} \le n \le N-1 \qquad (4.56)$$

[16] These expressions indicate only the analysis/synthesis transformation of an arbitrary segment of data $x(n)$ containing N samples. The 50% overlap between successive segments from $-\infty$ to $+\infty$ is therefore not considered. We take this approach to illustrate in a simplified way the mechanism of alias cancellation in the time domain. We will illustrate later on (Figure 4.11) the mechanism of 50% overlap for a representative sub-segment of data that can be generalized for the entire data signal.

for the second half of the same segment. It can be seen that these expressions contain aliased terms that are the symmetrical mirror image of the first half of the segment around $\frac{N}{4} - \frac{1}{2}$, and the mirror image of the second half of the segment around $\frac{3N}{4} - \frac{1}{2}$. The symmetry of the mirrored components are the essence of the aliasing cancellation property of TDAC and hence, of its PR property.

Considering:

1. the overlap-add operation between previous (last half) and current (first half) segments,

2. the absence of quantization of $X(k)$,

3. the linear phase constraint of the prototype filter (Sub-section 4.3.1.5):
$$h(n) = h(N-1-n)\ , \quad 0 \leq n \leq \frac{N}{2} - 1\ , \tag{4.57}$$

4. the analysis-synthesis time window identity:
$$f(n) = h(n)\ , \quad 0 \leq n \leq N-1\ , \tag{4.58}$$

it results that the PR requirement leads to[17]:
$$[h(n)]^2 + \left[h(\frac{N}{2}+n)\right]^2 = 1\ , \quad 0 \leq n \leq \frac{N}{4} - 1\ . \tag{4.59}$$

It should be noted that Equation (4.58) makes the synthesis filter equal to the analysis filter (*orthogonal* filter bank), which represents only a seeming simplification of the prototype filter design. In fact, as we will show later, this reduces the number of degrees of freedom with the consequence that an arbitrary filter design is not possible. A possible alternative to avoid this problem consists in a more general problem where the synthesis filter is different from the analysis filter (*biorthogonal* filter bank), although both windows are constrained to the perfect reconstruction requirement. For example, in the case of 50% overlap ($N = 2M$), $h(n)$ may first be chosen arbitrarily. Then, the synthesis window $f(n)$ may be uniquely determined using [Cheung95]:
$$f(n) = \frac{h(n)}{h^2(n) + h^2(n+M)}\ .$$
The advantage of this alternative design procedure is that it leaves the necessary degrees of freedom for an arbitrary design of one of the time windows.

Cheung argues [Cheung95] that biorthogonality allows the design of the analysis window according to the Masking Patterns (Section 3.3.3), and presents a few results showing essentially

[17]This condition reflects the power complementary requirement (Sub-section 4.3.1.5) and should consider an additional gain factor of 2 in the case of the TDAC filter bank.

a stop-band improvement at the cost of a wider main lobe (pass-band). However, it is clear that an improved analysis window implies a synthesis window with worse filtering characteristics[18] (or *vice-versa*), which makes more difficult the control of the aliased terms or, in Malvar's words *'generates some cross-correlation among the subband signals'* [Malvar92, page 224]. For this reason, we keep the orthogonality condition in order to have a balanced selectivity and control between at the analysis and synthesis stages, over the aliased and over other injected noise components. In this context and as Cheung recognizes [Cheung96, page 74], further optimization beyond the reference sine window (4.23) is very difficult. Eventual improvements may be far from dramatic because the constraints are non-linear, as we will show in the next sub-section.

4.5.3 Transform Windows and Requirements

A few examples of simple windows that verify the constraints (4.57) and (4.59) are:

- the rectangular window:

$$f(n) = h(n) = \sqrt{0.5}\,, \quad 0 \leq n \leq N-1\,, \tag{4.60}$$

  that corresponds to a maximum length maximum overlap window,

- the trapezoidal window:

$$\begin{aligned} f(n) = h(n) = 0.0\,, &\quad 0 \leq n \leq \frac{N}{4} - 1 \\ f(n) = h(n) = 1.0\,, &\quad \frac{N}{4} \leq n \leq \frac{3N}{4} - 1 \\ f(n) = h(n) = 0.0\,, &\quad \frac{3N}{4} \leq n \leq N-1 \end{aligned} \tag{4.61}$$

  that corresponds to a minimum length minimum (null) overlap window,

- a maximum length and maximum overlap sine window:

$$f(n) = h(n) = \sin\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)\right]\,, \quad 0 \leq n \leq N-1\,. \tag{4.62}$$

  This equation (repetition of Equation 4.23) imposes a constant energy time window which means, by the Parseval theorem, a constant energy frequency response.

The impulse response and corresponding frequency response for each one of these examples are depicted in Figure 4.10 for $N = 32$. From the point of view of transform coding and compression efficiency in particular, two specific properties of the transform window are extremely important:

[18] Alternatively, the synthesis window may have better characteristics than the analysis window such as smoother basis functions which reduces further the blocking effects [Malvar97].

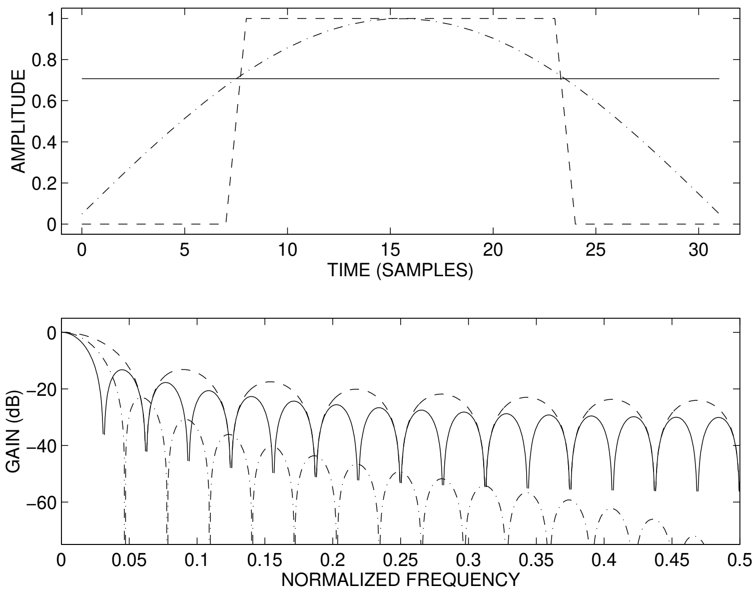

Figure 4.10: Examples of windows that comply with the TDAC perfect reconstruction constraint ($N = 32$).

1. *Selectivity*: the main lobe (pass-band) of the frequency response must be as narrow as possible and the stop-band leakage must be minimized.

2. *Monotony*: the envelope of the stop-band attenuation must be monotonically decreasing or equiripple since it is desirable to confine the spreading of the quantization noise as close as possible to the frequency band where it is injected (*i.e.* not on remote frequency bands). In other words it is desirable to achieve an accurate coloration of the quantization noise across the whole spectral range.

As it can be seen in Figure 4.10, the maximum length rectangular window has the narrowest main lobe but a poor stop-band attenuation, so it fails the selectivity objective. The trapezoidal window is worse in all respects. The sine window offers good stop-band attenuation and provides good attenuation of the block edge effect (smooth and maximum overlap). For these reasons, the sine window is widely used [Johns.92][Brand.91]. In addition, as we mentioned previously (Subsection 4.3.1.5), it offers a useful property (particularly convenient for video signals), when used for instance with the Odd-DFT or the TDAC filter bank, of concentrating the DC component of the signal on a single spectral line [Isabel.90][Ferre.95b], in the set of unique spectral coefficients.

4.5.4 Linear Filtering and Window Design

Without loss of generality, we consider the analysis of the sub-segment spanning fom sample $\frac{N}{2}$ to sample $N-1$, as represented in Figure 4.11. This representation is very convenient because

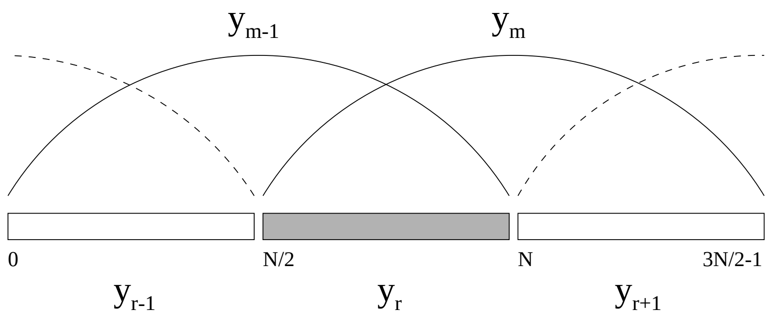

Figure 4.11: The dark sub-segment y_r results from a linear convolution with a zero-phase filter, or from the OA operation between transformed segments y_{m-1} and y_m.

it enables either the interpretation of a linear filtering operation:

$$y_r(n - \frac{N}{2}) = \sum_{p=n-\frac{N}{2}}^{n+\frac{N}{2}-1} x(p)\hat{d}(n-p) \; , \quad \frac{N}{2} \leq n \leq N-1 \; , \tag{4.63}$$

which depends on the zero-delay FIR filter $\hat{d}(n)$ whose impulse response is defined from $-\frac{N}{2}$ to $\frac{N}{2}-1$, or the OA operation between the adjacent transformed and windowed segments $y_{m-1}(n)$ and $y_m(n)$, thus completely defining $y_r(n)$.

Besides the window symmetry condition (4.57), it may also be assumed for practical reasons that the impulse response of the transform window (prototype analysis/synthesis filter) has an even length. This implies, for instance, that its frequency response will have a zero at $\omega = \pi$ (or $z = -1$) [Rabiner73]. Although this is a restriction, it may be proved [McCle.73a][Parks87] that all 4 possible combinations of symmetric or asymmetric impulse response and even or odd filter length, can be reduced to a unified formulation and thus any case may be translated into any other.

The Fourier transform of $h(n)$ is given by

$$H(\omega) = \sum_{n=0}^{N-1} h(n) e^{-j\omega n} \tag{4.64}$$

which, given the even length assumption, may be simplified to

$$H(\omega) = e^{-j\frac{N-1}{2}\omega} \sum_{n=0}^{\frac{N}{2}-1} 2h(n) \cos\left(\frac{N-1}{2} - n\right)\omega \; . \tag{4.65}$$

This expression shows the linear phase property of the filter. If there are no additional constraints besides symmetry and the even length condition of the impulse response of the filter $h(n)$, several design methods are available to obtain an arbitrary frequency response of the prototype filter.

For example, it may be proved [McCle.73a] that Equation (4.65) can be converted to:

$$H(\omega) = e^{-j\frac{N-1}{2}\omega}\cos(\frac{\omega}{2})\sum_{n=0}^{\frac{N}{2}-1} b(n)\cos(n)\omega \ , \tag{4.66}$$

which is convenient for the design of *optimal* filters in the minimax sense (minimization of the maximum error over a set of frequencies), using the Chebyshev approximation theory on compact sets. These filters have an approximation error function with the equiripple property and can approximate any arbitrary frequency response characteristic with a specified weighing error. Since there are $\frac{N}{2}$ degrees of freedom [Parks87] in the impulse response (which correspond to a maximum of $\frac{N}{2}$ unique pairs of zeros locations in the Z plane), the alternation theorem gives the necessary and sufficient conditions for the optimal solution [Rabiner73][Parks87]. The optimal solution can be easily and efficiently reached using the Remez multiple exchange algorithm [McCle.73b].

Linear programming techniques, based for example on the method of frequency sampling, have also been successful in the design of FIR filters [Rabiner72]. Besides the design of different types of filters like minimax filters (although using a slower -single exchange- convergence than that allowed by the Remez -multiple exchange- algorithm), they also offer the important possibility of including time domain constraints.

More, recently, a flexible method has been proposed by Adams [Adams91] for the design of optimal windows having properties of Dolph-Chebyshev filters, Prolate-Spheroidal filters or a combination of both.

In our context, if condition (4.59) is included in Equation (4.65), we obtain

$$H(\omega) = e^{-j\frac{N-1}{2}\omega}\sum_{n=0}^{\frac{N}{4}-1} 2\left[h(n)\cos\left(\frac{N-1}{2}-n\right)\omega + \sqrt{1-h(n)^2}\cos(n+\frac{1}{2})\omega\right] \ , \tag{4.67}$$

which makes clear that the above methods are not directly applicable because the number of degrees of freedom in the impulse response is $\frac{N}{4}$, which is less than required ($\frac{N}{2}$ as N is even) for an arbitrary design. In addition, the frequency response is not a linear function of the impulse coefficients. On the other hand, as stated in [Adams91] and [Parks87], it is known that even mathematically optimal windows like the Dolph-Chebyshev or the Prolate-Spheroidal have performances that are sometimes far from optimal in a practical sense. The desired window properties stated in Sub-section 4.5.3 will be major guidelines to the design of an optimal window in a mathematical *and* practical sense.

4.5.5 Convolution and Time Aliasing Using the Odd-DFT Filter Bank

To evaluate the nature and magnitude of the unwanted aliased terms that result from the multiplicative modification of a spectral representation as considered in Section 4.5.1, a simple but representative strategy will be used.

Each segment is multiplied by an analysis window $h(n)$ and transformed using the direct Odd-DFT. Then, the spectral coefficients are multiplied by a band-pass filter $D(k)$ and the inverse Odd-DFT calculated. Finally, successive segments are multiplied by a synthesis window and overlapped and added as illustrated in Figure 4.11. As these operations correspond to a circular convolution in the time domain, the equivalent expression will be found for the entire sub-segment ($y_r(n)$ in Figure 4.11) that results from the influence of adjacent segments ($y_m(n)$ and $y_{m-1}(n)$ in Figure 4.11). The terms that represent the implicit linear filtering will be isolated from the aliased terms. The former terms indicate the expression for the equivalent linear convolution operation in the time domain, as well as the equivalent filter. The latter terms will be subjected to minimization in an optimization procedure defined and explained in the next sub-section. This analysis is valid for all other sub-segments and hence for the entire processed signal.

The direct Odd-DFT is given by:

$$X(k) = \sum_{n=0}^{N-1} h(n)x(n)e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n} , \tag{4.68}$$

and the inverse Odd-DFT is given by:

$$y(n) = \frac{f(n)}{N} \sum_{k=0}^{N-1} X(k)e^{j\frac{2\pi}{N}(k+\frac{1}{2})n} . \tag{4.69}$$

To simplify the analytical development, we will consider a rectangular and zero phase low-pass filter whose discrete frequency response is defined as:

$$\begin{aligned} D(k) &= 1.0\,, & 0 \le k \le M-1\,; \\ D(k) &= 0.0\,, & M \le k < \frac{N}{2}\,; \\ D(N-1-k) &= D(k)^*, & 0 \le k \le \frac{N}{2}-1\,. \end{aligned} \tag{4.70}$$

The corresponding impulse response is given by its inverse Odd-DFT:

$$d(n) = \frac{1}{N}\sum_{k=0}^{N-1} D(k)e^{j\frac{2\pi}{N}(k+\frac{1}{2})n} = \frac{1}{N}\frac{\sin(\frac{2\pi}{N}Mn)}{\sin(\frac{\pi}{N}n)} , \quad -\frac{N}{2} \le n \le \frac{N}{2}-1\,. \tag{4.71}$$

For convenience, $d(n)$ will be associated to a single period of the filter comprising the range $-\frac{N}{2} \le n \le \frac{N}{2} - 1$, although it is clear that Equation (4.71) represents, in fact, a periodic signal. Considering the analysis and synthesis windows satisfying (4.57), (4.59) and (4.58), and using Equations (4.71), (4.68) and (4.69), the inverse Odd-DFT of the filtered spectrum $\tilde{X}(k) = X(k)D(k)$ can be easily shown to be:

$$y(n) = f(n) \sum_{p=0}^{N-1} x(p)h(p) \sum_{\ell=-\infty}^{+\infty} d(n-p-\ell N) . \tag{4.72}$$

Referring to Figure 4.11, Equation (4.72) is converted to:

$$y_{m-1}(n) = f(n) \sum_{p=0}^{N-1} x(p)h(p) \sum_{\ell_1=-\infty}^{+\infty} d(n-p-\ell_1 N) , \quad 0 \leq n \leq N-1 , \tag{4.73}$$

for the time segment $m-1$, and to:

$$y_m(n-\frac{N}{2}) = f(n-\frac{N}{2}) \sum_{p=\frac{N}{2}}^{\frac{3N}{2}-1} x(p)h(p-\frac{N}{2}) \sum_{\ell_2=-\infty}^{+\infty} d(n-p-\ell_2 N) , \qquad \frac{N}{2} \leq n \leq \frac{3N}{2}-1 , \tag{4.74}$$

for the time segment m. The time windows $h(n)$ and $f(n)$ are referred to the range $0 \leq n \leq N-1$. These segments are overlapped and added as illustrated in Figure 4.11. The sub-segment of interest is represented by $y_r(n-\frac{N}{2})$ and results only from the above segments, *i.e.* from $y_{m-1}(n)$ and $y_m(n-\frac{N}{2})$ with $\frac{N}{2} \leq n \leq N-1$. If the result of this operation is split in convenient components and the possible values of ℓ_1 and ℓ_2 are accounted for in each case, we obtain:

$$\begin{aligned} y_r(n-\frac{N}{2}) = y_{m-1}(n) + y_m(n-\frac{N}{2}) = {} & f(n) \sum_{p=0}^{n-\frac{N}{2}-1} x(p)h(p) \sum_{\ell_1=-\infty}^{+\infty} d(n-p-\ell_1 N) + \\ & f(n) \sum_{p=n-\frac{N}{2}}^{\frac{N}{2}-1} x(p)h(p) \sum_{\ell_1=-\infty}^{+\infty} d(n-p-\ell_1 N) + \\ & \sum_{p=\frac{N}{2}}^{N-1} x(p) \left[f(n)h(p) + f(n-\frac{N}{2})h(p-\frac{N}{2}) \right] d(n-p) + \\ & f(n-\frac{N}{2}) \sum_{p=N}^{\frac{N}{2}+n-1} x(p)h(p-\frac{N}{2}) \sum_{\ell_2=-\infty}^{+\infty} d(n-p-\ell_2 N) + \\ & f(n-\frac{N}{2}) \sum_{p=n+\frac{N}{2}}^{\frac{3N}{2}-1} x(p)h(p-\frac{N}{2}) \sum_{\ell_2=-\infty}^{+\infty} d(n-p-\ell_2 N) . \end{aligned}$$

At this point, if suitable terms are added and subtracted, we obtain

$$y_r(n-\frac{N}{2}) = y_{filter}(n) + y_{alias1}(n) + y_{alias2}(n) , \quad \frac{N}{2} \leq n \leq N-1 , \tag{4.75}$$

where $y_{filter}(n)$, $y_{alias1}(n)$ and $y_{alias2}(n)$ are respectively given by Equations (4.76), (4.77) and (4.78).

$$y_{filter}(n) = \sum_{p=n-\frac{N}{2}}^{n+\frac{N}{2}-1} x(p)\left[f(n)h(\lfloor p\rfloor_N) + f(n-\frac{N}{2})h(\lfloor p-\frac{N}{2}\rfloor_N)\right]d(n-p) \tag{4.76}$$

$$y_{alias1}(n) = f(n-\frac{N}{2})\sum_{p=n-\frac{N}{2}}^{\frac{N}{2}-1}\left[x(p+N)-x(p)\right]h(p+\frac{N}{2})d(n-p) \tag{4.77}$$

$$y_{alias2}(n) = f(n)\sum_{p=N}^{n+\frac{N}{2}-1}\left[x(p-N)-x(p)\right]h(p-N)d(n-p) \tag{4.78}$$

The notation $\lfloor p\rfloor_N$ denotes p *modulo* N.

Comparing Equation (4.76) to Equation (4.63), it is clear that the filtering in the frequency domain using overlapped transforms as illustrated in Figure 4.11 corresponds mainly to a linear filtering operation whose equivalent filter, for the 50% overlap configuration, is implicitly given by Equation (4.76). Although the equivalent filter is in general time varying, it can be easily proved that when an alternative form[19] of Equation (4.76) is considered together with the time window given by Equation (4.62), a time invariant filter is obtained and Equation (4.76) reduces to:

$$y_{filter}(n) = \sum_{p=n-\frac{N}{2}}^{n+\frac{N}{2}-1} x(p)\cos\frac{\pi}{N}(n-p)d(n-p)\ . \tag{4.79}$$

This result makes clear that the analysis/synthesis time window (prototype filter) introduces a smoothing effect on the filter $d(n)$. This will be exploited in Chapter 7 to improve the characteristics of filters corresponding to a desired equalization curve.

In addition there are, as expected, circularly aliased terms whose magnitudes are given by Equations (4.77) and (4.78). Their sum, $y_{alias}(n) = y_{alias1}(n) + y_{alias2}(n)$, will be the main objective function subjected to minimization:

$$y_{alias}(n) = f(n-\frac{N}{2})\sum_{p=n-\frac{N}{2}}^{\frac{N}{2}-1} 2x_{max}h(p+\frac{N}{2})d(n-p) + f(n)\sum_{p=N}^{n+\frac{N}{2}-1} 2x_{max}h(p-N)d(n-p). \tag{4.80}$$

As nothing is postulated about the statistical properties of $x(n)$, it will be assumed that $[x(p+N)-x(p)]$ and $[x(p-N)-x(p)]$ are worst case constant values equal to $2x_{max}$ ($= 2^{16}$ if 16 bit linearly quantized audio samples are considered).

[19] By multiplying by -1 those terms depending on $\lfloor p\rfloor_N$ when ℓ is outside the range $[0,\ N-1]$.

4.5.6 Optimization Criteria and Results

It is not surprising that the direct minimization of (4.80) leads to a solution corresponding to the trapezoidal window (4.61) for any band-pass filter $d(n)$. This solution corresponds in fact to a trivial solution where the time aliased terms are eliminated by brute force, with negative implications to the quality of the linear filter resulting from Equation (4.76). In other words, with the exception of the final multiplication by window $f(n)$, this case is identical to the classical overlap-add method [Oppenh.75] of linear convolution using transforms and zero padding. In this case, $d(n)$ is supposed to be a zero phase and symmetrical filter with a maximum length of $\frac{N}{2}+1$ taps. This trivial solution suffers from blocking noise, and has a poor frequency response as illustrated in Figure 4.10.

In order to formulate a reasonable optimization problem, a constraint was added on the frequency response of $h(n)$, and hence on that of $f(n)$. As the sine window (4.62) is commonly used in audio coding, it was decided to consider it as the reference window (abbreviated to *refwin*). The optimal window (abbreviated to *optwin*) should have a measure of the frequency leakage (*i.e.* the stop-band leakage) above the first zero defining the main lobe (pass-band) of the reference window, which occurs at $\frac{3\pi}{N}$, preferably better than the same measure obtained for the reference window, *e.g.* it should be

$$\int_{\omega=\frac{3\pi}{N}}^{\pi} |H_{optwin}(\omega)|^2 d\omega \quad \leq \quad \int_{\omega=\frac{3\pi}{N}}^{\pi} |H_{refwin}(\omega)|^2 d\omega \ . \tag{4.81}$$

This criterion ensures that the main lobe of the prototype filter is not allowed to be so large as to relax the selectivity of the filter.

We have considered (4.82) as the metric for the stop-band leakage and (4.83) as the metric for the time aliasing.

$$\int_{\omega=\frac{3\pi}{N}}^{\pi} |H_{optwin}(\omega)|^2 d\omega \tag{4.82}$$

$$\sum_{n=\frac{N}{2}}^{N-1} y_{alias}(n)^2 \tag{4.83}$$

The optimization problem can finally be formulated as follows: minimize (4.83) and (4.82), subject to Equations (4.57), (4.59), (4.58), and to

$$0.0 \leq h(n) \leq 1.0 \ , \quad 0 \leq n \leq \frac{N}{4} - 1 \ , \tag{4.84}$$

and and taking into account the selectivity and monotony requirements presented in Sub-section 4.5.3.

As shown in Sub-section 4.5.4, $H(\omega)$ is given by Equation (4.67). Any band-pass filter $d(n)$ with demanding transition band requirements will be appropriate (it should give rise to significant amounts of time aliasing). For simplicity, the low-pass filter specified in Equation (4.71), with $M = 1$ and $-\frac{N}{2} \leq n \leq \frac{N}{2} - 1$, was considered for $d(n)$.

Given the non-linear nature of this optimization problem, in order to detect and avoid convergence difficulties specific to a particular optimization procedure, we used two different optimization algorithms to assess our results: our own algorithm based on a simple single exchange approach, and the Nelder-Meade simplex algorithm used within a commercial numeric software package. All the results presented here can be considered to be obtained from either technique, since the only relevant difference has been noticed in the number of iterations needed before reaching a stable solution.

We used the impulse coefficients of the reference window to set the initial conditions.

We have concluded that if condition (4.82) is ignored, we reach the undesirable trivial solution corresponding to window (4.61) as mentioned above. On the other hand, if we ignore condition (4.83) and minimize (4.82) instead, we obtain an optimal solution as illustrated in Figure 4.12 for N=32. This optimal window corresponds to a reduction of the stop-band leakage by about

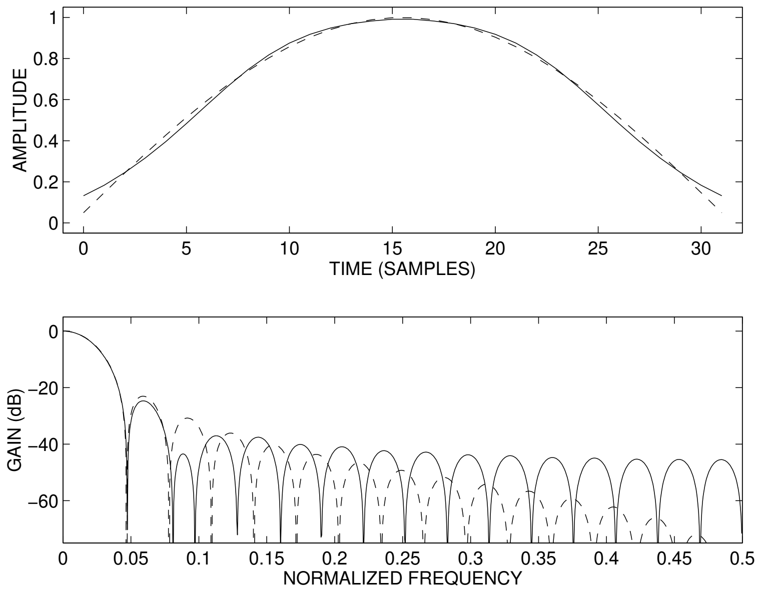

Figure 4.12: A window with minimized stop-band leakage (solid line) relative to the reference window (dashed line) when $N = 32$. Top: impulse response, bottom: frequency response.

30.5% relative to the reference window. It can be seen that this gain is the consequence of a concentration of attenuation effort (of zeros in the Z plane) near the stronger stop-band lobes. However, this window represents an increase of the time aliasing, according to Equation (4.83) and relative to the reference window, by about 13.0%. A reduction of this disadvantage can only be reached by giving up some fraction of the stop-band leakage gain. Therefore, in general, any optimal window resulting from our optimization problem represents a tradeoff between the ability to reduce stop-band leakage and the ability to reduce time aliasing. Different tradeoffs are possible by means of an appropriate weighting of each aspect in the optimization procedure.

According to our results, it is possible to have gains in both aspects relative to the reference window which means the reference window is *not* optimal. As an example, Figure 4.13 illustrates an optimal window corresponding to a balance of 33.6% gain in time aliasing and 11.8% gain in stop-band leakage relative to the reference window. In the limit, additional gains in time

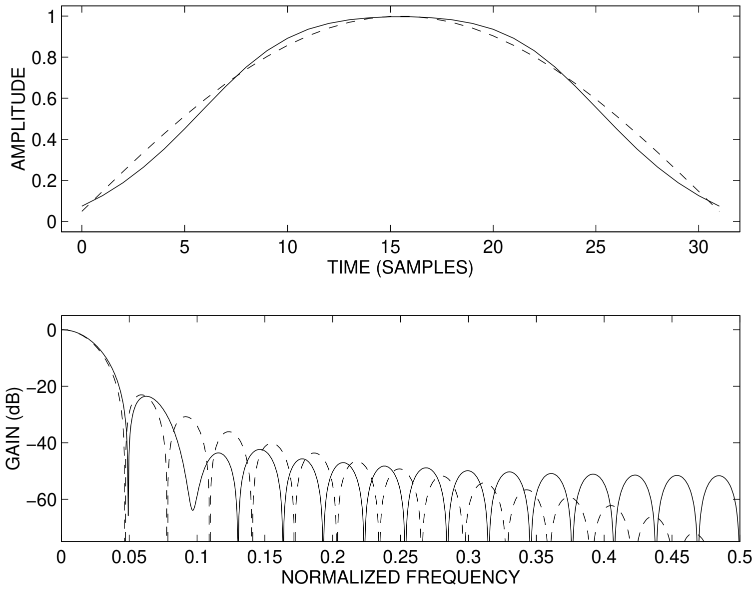

Figure 4.13: An optimal window (solid line) allowing gains in stop-band leakage and time aliasing relative to the reference window (dashed line) when $N = 32$. Top: impulse response, bottom: frequency response.

aliasing can be obtained only at the cost of negative gains in stop-band leakage, as we will show later.

The examples of Figure 4.12 and Figure 4.13 correspond to window lengths of 32 taps, which represent, by the considerations presented in Sub-section 4.5.4, an optimization problem involving 8 independent parameters. For longer window lengths, such as 1024 taps, the convergence of the optimization process becomes a problem. In the limit, it may not be possible to reach a satisfactory solution within an acceptable number of iterations. For this reason, and as suggested by the smooth shape of the optimal windows illustrated previously, we investigated a model based on a closed form expression, and involving a small number of parameters, to approximate optimal windows.

Several optimal window designs, for different values of N (up to 128), lead us to investigate the possibility of using for the closed-form expression, the reference window expression modified by the addition of a non-linear phase term. Taking the optimum window represented in Figure 4.13 as an example, the normalized phase deviation from the phase of the reference window in the interval $\left[0,\ \frac{N}{4}-1\right]$ is depicted in Figure 4.14. Also represented in this figure is the best

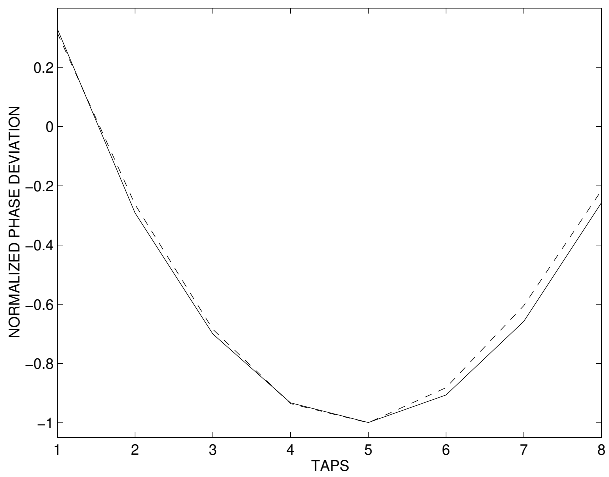

Figure 4.14: A modeled approximation (solid line) of the phase deviation (dashed line) of an optimal window from the phase of the reference window ($N = 32$).

approximation, in a least square sense, of a deformed parabola involving only two parameters. The denormalization adds a third parameter which leads to the final model:

$$h(n) = \sin\left\{\frac{\pi}{N}\left(n+\frac{1}{2}\right) + \frac{4\pi\beta}{\left(1-\delta^2\right)}\left[\left(\frac{4n}{N-2}\right)^{\alpha} - \delta\right]\left[\left(\frac{4n}{N-2}\right)^{\alpha} - 1\right]\right\}, \tag{4.85}$$

$$h(N-1-n) = h(n)\,,$$

$$h(\frac{N}{2}+n) = h(\frac{N}{2}-1-n) = \sqrt{1-h(n)^2}\,,$$

$$\begin{aligned} 0 \leq \quad & n & \leq \frac{N}{4}-1\,, \\ 0.0 \leq \quad & \alpha,\beta & , \\ -1.0 < \quad & \delta & < 1.0\,. \end{aligned}$$

As a consequence of this simplification, our optimization problem has been reformulated as a function only of the parameters of Equation (4.85). Our new optimal solution for the same length and tradeoff of time aliasing and stop-band leakage of the optimal window of Figure 4.13 differs less than 0.05% from the results obtained previously. The difference between the ideal optimal solution and the approximated solution given by (4.85) for other window lengths, is of the same order which is acceptable for practical purposes. To complete our example, it can be confirmed in Figure 4.15 that the impulse response and frequency response of the approximated optimal window are practically indistinguishable from those of the optimal window represented in Figure 4.13.

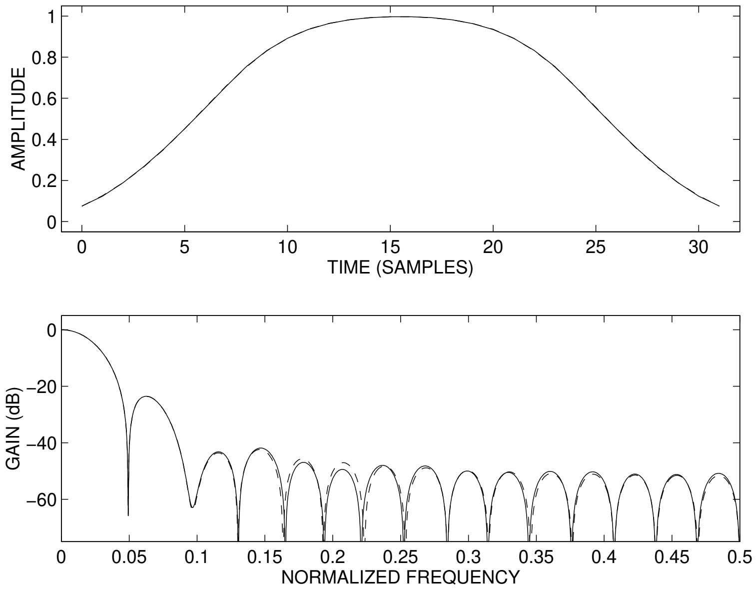

Figure 4.15: An approximation given by a closed form expression (solid line) to the ideal optimal window (dashed line) when $N = 32$.

As expected, the main advantage of the simplified optimization procedure using Equation (4.85) is a faster convergence to the optimal solution, regardless of the value of N. It is therefore possible to evaluate rapidly different optimal tradeoffs between time aliasing and stop-band leakage, *e.g.* for N in the order of 512 and 1024, as it is usually considered in high quality audio coding.

Taking $N = 1024$, we obtained, in particular, three optimal tradeoffs as specified in Table 4.2. We concluded that maintaining approximately the same stop-band leakage of the reference

Table 4.2: Reduction of the (Odd-DFT) time aliasing and stop-band leakage when $N = 1024$, relative to the reference window.

			Reduction (%)	
β	α	δ	**Time Aliasing**	**Stop-Band Leakage**
0.02320	0.937	0.145	30.0	10.1
0.02630	0.937	0.111	37.9	0.7
0.03125	0.920	0.000	53.8	-31.0

window, an optimal window would allow a reduction of about 38% in time aliasing. This window is represented in Figure 4.16. It can be seen that the attenuation effort, as already observed

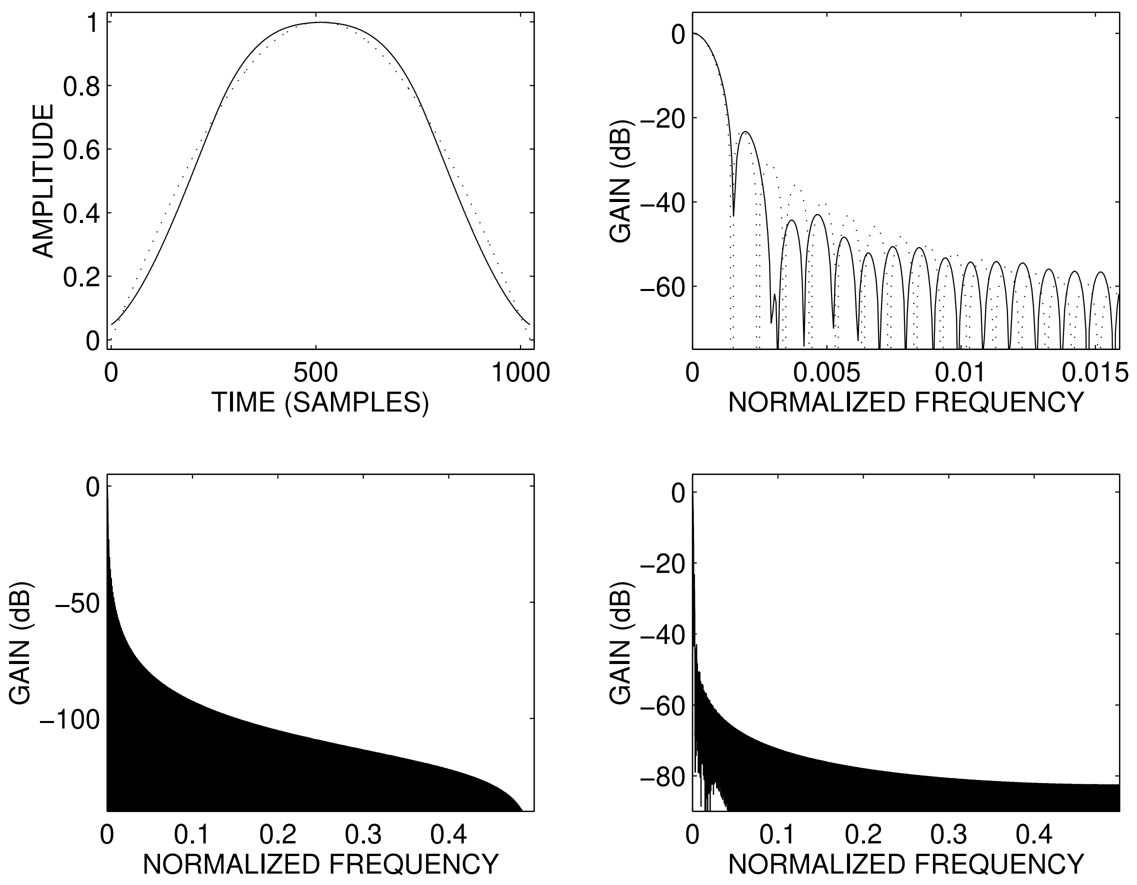

Figure 4.16: An optimal window (solid line in upper figures and in lower right figure) having reduced time aliasing and stop-band leakage relative to the reference window (dotted line in upper figures and solid line in lower left figure) when $N = 1024$.

in Figure 4.12 for $N = 32$, is mainly concentrated between the first and second lobes of the stop-band region. This is so because the constraint (4.59) of our optimization problem does not allow an effective control over the first stop-band lobe which is primarily responsible for the stop-band leakage. If a further reduction in time aliasing is desired, an increase of the stop-band leakage relative to the reference window must be allowed. An example is specified in Table 4.2 and illustrated in Figure 4.17. The increase in stop-band leakage relative to the reference window is mainly due to a small increase of the first stop-band lobe. However, above this lobe, a better channel isolation is achieved. This window exhibits a monotonic increase of the attenuation towards high frequencies in a rather curious way. In fact, clusters of three lobes, approximately equiripple, alternate with an isolated lobe and their attenuation difference progressively decreases, until the floor of the remaining stop-band is equiripple. This example is suitable for digital audio applications since it exhibits a well-behaved attenuation of more than 110 dB in a wide range of frequencies. Our tests with real speech and music signals in

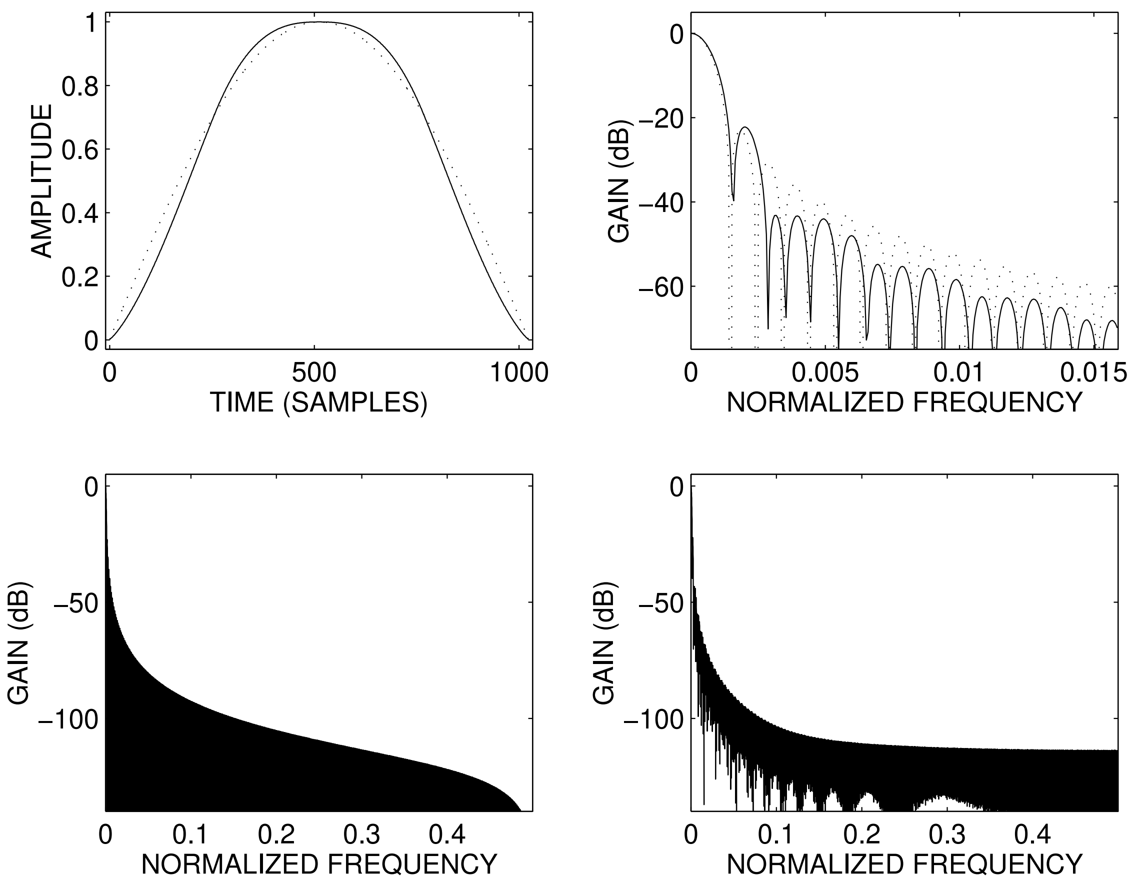

Figure 4.17: An optimal window (solid line in upper figures and in lower right figure) having less time aliasing but more stop-band leakage than the reference window (dotted line in upper figures and solid line in lower left figure) when $N = 1024$. The higher stop-band leakage is primarily due to the higher first lobe in the stop-band.

audio coding [Ferre.95a] and frequency domain filtering [Ferre.95c] (with different values for M in (4.71) as well as other more complex filter responses) revealed convincing subjective quality improvements relatively to those obtained with the reference window.

Although we did not consider any special frequency weighting of the frequency response of the optimal window, it is possible to include this weighting in Equation (4.82).

As a final remark, it is interesting to note a duality property between the reference window and an optimal window according to Equation (4.85). In fact, it is apparent that the inflection of the envelope of the reference window frequency response at $\frac{2\pi}{4}$ (digital frequency) has a counterpart, in the time domain, corresponding to an inflection of the impulse response of any optimal window at $\frac{N}{4}$.

4.5.7 Convolution and Time Aliasing Using the TDAC Filter Bank

The analysis of the circular convolution resulting from the multiplication of the TDAC spectrum of a signal $x(n)$ by a filter $D_{TDAC}(k)$, follows identical steps to those presented in Sub-section 4.5.5 for the Odd-DFT transform. In fact, recognizing first that Equation (4.37) can be expressed as:

$$X(k) = \frac{1}{2}\sum_{n=0}^{N-1} h(n)x(n)\left[e^{-j\frac{2\pi}{N}(k+\frac{1}{2})(n+\frac{1}{2}+\frac{N}{4})} + e^{j\frac{2\pi}{N}(k+\frac{1}{2})(n+\frac{1}{2}+\frac{N}{4})}\right] , \tag{4.86}$$

and that Equation (4.54) can be expressed as:

$$y(n) = \frac{f(n)}{2N}\sum_{n=0}^{N-1} X(k)\left[e^{j\frac{2\pi}{N}(k+\frac{1}{2})(n+\frac{1}{2}+\frac{N}{4})} + e^{-j\frac{2\pi}{N}(k+\frac{1}{2})(n+\frac{1}{2}+\frac{N}{4})}\right] , \tag{4.87}$$

and redefining $d(n)$ as the inverse TDAC transformation of $D_{TDAC}(k)$, it can be concluded that the inverse TDAC transformation of $\tilde{X}(k) = X(k)D_{TDAC}(k)$ can be expressed as a function of two basic components given by $y_{TDAC1}(n)$ and $y_{TDAC2}(n)$:

$$y_{TDAC1}(n) = \frac{1}{2}f(n)\sum_{p=0}^{N-1} x(p)h(p)\sum_{\ell=-\infty}^{+\infty} d(n-p-\ell N) , \qquad 0 \le n \le N-1 ; \tag{4.88}$$

$$y_{TDAC2}(n) = \frac{1}{2}f(n)\sum_{p=0}^{N-1} x(p)h(p)\sum_{\ell=-\infty}^{+\infty} d(n+p+1+\frac{N}{2}+\ell N) , \qquad 0 \le n \le N-1 . \tag{4.89}$$

Given that

$$X_{TDAC}(N-1-k) = -X_{TDAC}(k) , \qquad 0 \le k \le \frac{N}{2}-1 , \tag{4.90}$$

and taking for $D_{TDAC}(k)$ the simplified low-pass filter introduced in Sub-section 4.5.5, the corresponding impulse response results as:

$$d(n) = \frac{1}{N}\frac{\sin\left[\frac{2\pi}{N}M\left(n+0.5+\frac{N}{4}\right)\right]}{\sin\left[\frac{\pi}{N}\left(n+0.5+\frac{N}{4}\right)\right]} . \tag{4.91}$$

This result is simply a time shifted version of Equation (4.71).

As Equation (4.88) is identical to Equation (4.72) except for a gain factor, the results obtained previously in Equations (4.75), (4.77) and (4.78), do also apply to this case.

On the other hand, Equation (4.89) may be further analyzed by considering the reconstructed time segment $m-1$, *i.e.* $y_{TDAC2(m-1)}(n)$:

$$y_{TDAC2(m-1)}(n) = \frac{1}{2}f(n)\sum_{p=0}^{N-1} x(p)h(p)\sum_{\ell_1=-\infty}^{+\infty} d(n+p+1+\frac{N}{2}+\ell_1 N) , \qquad 0 \le n \le N-1 , \tag{4.92}$$

and by considering the reconstructed time segment m, *i.e.* $y_{TDAC2(m)}(n)$:

$$y_{TDAC2(m)}(n-\frac{N}{2}) = \frac{1}{2}f(n-\frac{N}{2})\sum_{p=\frac{N}{2}}^{\frac{3N}{2}-1} x(p)h(p-\frac{N}{2})\sum_{\ell_2=-\infty}^{+\infty} d(n+p+1+\frac{N}{2}+\ell_2 N)\ , \quad \frac{N}{2}\leq n\leq \frac{3N}{2}-1\ . \tag{4.93}$$

Following the same development described in Sub-section 4.5.5, we obtain:

$$\begin{aligned} y_{TDAC2(r)}(n-\frac{N}{2}) = y_{TDAC2(m-1)}(n) + y_{TDAC2(m)}(n-\frac{N}{2}) = \\ y_{filterTDAC2}(n) + y_{alias1TDAC2}(n) + y_{alias2TDAC2}(n)\ , \\ \frac{N}{2}\leq n\leq \frac{3N}{2}-1\ , \end{aligned} \tag{4.94}$$

where $y_{filterTDAC2}(n)$ is given by:

$$y_{filterTDAC2}(n) = \frac{1}{2}\sum_{p=N-n-1}^{2N-2-n} x(p)\left[f(n)h(\lfloor p\rfloor_N) + f(n-\frac{N}{2})h(\lfloor p-\frac{N}{2}\rfloor_N)\right] d(n+p+1-\frac{3N}{2})\ , \tag{4.95}$$

$y_{alias1TDAC2}(n)$ is given by:

$$y_{alias1TDAC2}(n) = \frac{f(n)}{2}\sum_{p=0}^{N-2-n}\left[x(p)-x(p+N)\right]h(p)d(n+p+1-\frac{N}{2})\ , \tag{4.96}$$

and $y_{alias2TDAC2}(n)$ is given by:

$$y_{alias2TDAC2}(n) = \frac{f(n-\frac{N}{2})}{2}\sum_{p=N-n+1}^{\frac{N}{2}-1}\left[x(p+N)-x(p)\right]h(p+\frac{N}{2})d(n+p+1-\frac{3N}{2})\ . \tag{4.97}$$

$y_{filterTDAC2}(n)$ is the result of a time varying filter that "filters" in a reversed (aliased) fashion. In fact, when the equivalent filter window is centered around n the result of the operation is placed in the symmetrical position $\frac{N}{2}-1-n$ within $y_{TDAC2(r)}(n)$. As such, it corresponds to an unvoidable aliasing component in the time domain that must be considered in the objective function (4.83) together with other components given by Equation (4.96), Equation (4.97), as well as the aliasing components resulting from Equation (4.88). Despite this fact, the nature of our optimization problem is maintained. For simplicity, we set $x(p)$ in Equation (4.95) to a constant value. Our simulations lead to the following conclusions:

1. If the objective function (4.82) is ignored in the optimization, and if we consider in the objective function (4.83) the aliasing terms resulting from (4.88) only, (4.96) and (4.97) only, those resulting from (4.95) only, or a combination of all, we obtain again the trivial

solution corresponding to window (4.61). The trivial solution reduces to zero the aliasing terms resulting from (4.88), as verified previously for the case of the Odd-DFT, but not the remaining aliased terms.

2. If we take any combination of the time aliasing terms identified previously and consider the two objective functions in our optimization problem, we obtain the exact same solutions obtained in Sub-section 4.5.6, except for a significant difference in the values of the time aliasing reduction. This is a consequence of the fact that Equation (4.95) represents the main contribution to the time energy to be minimized. Since the magnitude of this filter approaches the magnitude of the main filter resulting from Equation (4.88), the time aliasing reduction for a particular tradeoff of stop-band leakage reduction / time aliasing reduction is rather modest, relative to the results obtained previously for the Odd-DFT.

   Table 4.3 summarizes the new results with the TDAC filter bank for the same optimal windows and conditions of Table 4.2.

Table 4.3: Reduction of the (TDAC) time aliasing and stop-band leakage when $N = 1024$, relative to the reference window.

			Reduction (%)	
β	α	δ	**Time Aliasing**	**Stop-Band Leakage**
0.02320	0.937	0.145	6.3	10.1
0.02630	0.937	0.111	8.3	0.7
0.03125	0.920	0.000	13.3	-31.0

We conclude from these results that our model of an optimal solution as given by Equation (4.85), has been corroborated by the results of the new context involving the TDAC filter bank. The main new result of this context is that the reduction of the time aliasing terms achieved with an optimal window of practical interest is not as relevant as the results obtained in Sub-section 4.5.6 are.

4.6 Summary

This Chapter has reviewed the theory of multirate filter banks and has discussed the importance in audio coding of specific forms of signal decomposition.

Two other important aspects in audio coding have been addressed: the classification of audio signals according to their spectral coherence, and the minimization of aliased distortions and block edge effects resulting from a spectral modification of the audio signal.

The conclusions of the research reported in this Chapter have determined the analysis/synhesis filter bank of ASC.

Chapter 5

Coding Techniques and the State-of-The-Art

History repeats itself.

common sense.

5.1 Introduction

The digital representation of audio signals is extremely attractive from the point of view of signal quality and system design. In fact, signal processing techniques can be used, for example, to preserve the signal quality, integrity and privacy, regardless of the storage or transmission *medium*; to detect and correct errors introduced by poor transmission channels, to multiplex different signals allowing dynamic reconfiguration of the transmission system, to compress audio-visual data allowing an efficient share of transmission channels suffering from low bandwidth, or storage devices characterized by a low capacity.

The simplest digital representation of audio signals consists in the PCM format which corresponds to an uniform discretization in time (sampling) and magnitude (quantization) of the analog audio signal. The sampling frequency (in Hertz) and the resolution of the quantization (in bits per sample) define the quality of the digital representation. Table 5.1 illustrates typical values for three categories of quality: speech, wideband speech, and high quality audio. The sampling frequencies indicated for high quality audio correspond to broadcast applications (32 *kHz*), consumer applications (44.1 *kHz*), and professional applications (48 *kHz*). As introduced in Chapter 1, the main topic of this Dissertation is the bit rate reduction of the PCM representation of an audio signal, without degrading significantly its subjective quality, and while minimizing other variables of the performance equation (2.1), namely the system delay and

Table 5.1: Typical quality parameters of different types of PCM audio.

	Signal Bandwidth (kHz)	**Sampling Frequency** (kHz)	**Resolution** (bit/sample)	**PCM Bit Rate** (Kbit/sec)
Speech	3.4	8.0	8	64
Wideband Speech	7.0	16.0	16	256
High Quality Audio	15.0, 20.0	32.0, 44.1, 48.0	≥ 16	512, 706, 768

algorithmic complexity.

In this Chapter we will address briefly the quality measures used to rate the subjective quality of a coding process, and we will review the most common schemes of bit rate reduction for each one of the above signal categories. The terms *compression* or *coding* will be used with the same meaning as bit rate reduction.

High quality audio coding techniques were derived from or are largely based on speech coding techniques. Historically, these coding techniques were the first to be investigated (also due to technological reasons), implemented and standardized. For this reason, we will review in Section 5.3 the main concepts and techniques commonly used in the speech coding area, focusing essentially on the fact that aspects of perception have been progressively given more importance in the design of a coder.

In Section 5.4 we will give a few examples of wideband speech coders which may involve extensions of typical speech coders, as well as adaptations of wideband audio coders. In Section 5.5 we will show how the new coding paradigm based on models of human perception and known as *perceptual coding*, dominates in general the underlying structure of most high quality audio coders. Both proprietary and standardized algorithms will be illustrated.

5.2 Quality Measures

As introduced in Section 4.3.1.7, traditional objective measures of signal fidelity such as the SNR and the Total Harmonic Distortion (THD), are not appropriate to inform about signal quality as perceived by the human auditory system. To evaluate the impact of a coding process on the quality of the audio signal, it is necessary to take into account aspects of human perception.

Automatic evaluation procedures exist that are based on measures of perceptual similarity between the coded signal and the (delayed) original signal [Brand.92a][Pailla.92]. However, such objective perceptual measurements are based on practical (and thus inevitably incomplete) psychoacoustic models. Although some of their results correlate with formal subjective assessments, their validity has not yet being widely recognized [Bosi97].

A more reliable evaluation procedure, even if much more expensive and time consuming, involves formal listening tests where subjects are asked to quantify their impression regarding the *quality* or the *impairment* of the coded signal, relative to the original signal. These tests are known as the *Mean Opinion Score* tests [Jayant84]. Concerning quality, a 5-point rating scale (MOS scale) is frequently used that assigns 5.0 to the subjective impression of *excellent quality*, and 1.0 to the subjective impression of *bad quality*. Intermediate levels are detailed in Table 5.2. Concerning impairment, the 5-point ITU-R rating scale [Bosi97] is frequently used to characterize subtle differences between original and coded signals, ranging from *imperceptible* (5.0) to *very annoying* (1.0), as detailed in Table 5.2. The statistical analysis of the subjective

Table 5.2: 5-point grading scale of subjective impression regarding quality (MOS) and impairment (ITU-R scale) of coded material.

Grade	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not Annoying
3	Fair	Slightly Annoying
2	Poor	Annoying
1	Bad	Very Annoying

answers provided by subjects (trained or naïve), yields an average judgement that rates the quality of the coding process, and serves to compare objectively different coding methods, on the basis of the same test conditions and listener panel [Noll95].

5.3 Speech Coding

5.3.1 Speech: Properties, Quality and Bit-Rate Targets

Most frequently, speech signals exhibit quasi-stationary over durations in the order of 10 to 20 *ms*, and essentially two types of structure, which explains the traditional classification between *voiced* and *unvoiced* speech. Voiced speech is produced by exciting the vocal tract with quasi-stationary glottal air pulses generated by the vibrating vocal chords. The frequency of the periodic pulses is referred to as the fundamental frequency or *pitch* [Spanias94]. The short-term spectrum of a voiced signal is characterized by a fine harmonic structure, and a spectral envelope (evident from the spectral peaks) defining the formant structure. The formant structure results from the transfer characteristics (*i.e.* the resonant modes) of the vocal tract and the spectral "tilt" due to the glottal pulse [Spanias94]. Unvoiced speech is produced by forcing air through a constriction in the vocal tract and corresponds approximately to a broadband random-like signal.

Speech is usually sampled at 8 *kHz* and quantized uniformly to 13 or 14 bits per sample (uniform PCM). The white quantization noise introduced by this resolution is sufficiently low to preserve approximately the quality of the unquantized signal. A non-uniform PCM representation is obtained by companding the uniform PCM according to a logarithmic law. Two specific laws, the A-law (mainly used in Europe) and the μ-law (mainly used in North America), have been standardized [ITU72]. These laws requantize each speech sample of 13 bits (A-law) or of 14 bits (μ-law) to 8 bits. This memoriless operation provides a very modest compression because it consists of a scalar quantization whose errors are made proportional to the magnitude of the original samples, and because it does not exploit signal redundancies, *i.e.* the correlation between adjacent samples. Its rationale is to employ a fine quantization step for frequently occurring amplitudes (mostly of low level) and a coarse quantization step for rarely occurring amplitudes (mostly of high level), according to an average *probability density function* (PDF). As a consequence, the SNR is reasonably maintained over a wide range of amplitudes. This 8-bit non-uniform PCM representation sets the highest grade of telephone speech quality: *network* or *toll* quality. The following list identifies the most common grades of speech quality, the associated expectation of subjective quality, and the bit rates for which they can be achieved with current technology [Spanias94]:

Broadcast ($\geq$ 64 $Kbit/s$) wideband speech and high quality commentary speech,

Network or Toll ($\geq$ 16 $Kbit/s$) quality quite comparable to the analog speech,

Communications ($\geq$ 4.8 $Kbit/s$) somehow degraded speech but preserving the naturalness and intelligibility of the original speech,

Synthetic ($<$ 4.0 $Kbit/s$) intelligible speech but not entirely natural, with a possible loss of speaker recognizability.

The bit rate of the non-uniform PCM representation (64 $Kbit/s$) represents an upper limit reference for the target bit rate of a speech coder. Coders targeting specific bit rates are classified as follows [Spanias94]:

high rate coders (16 – 64 $Kbit/s$),

medium rate coders (8 – 16 $Kbit/s$),

low rate coders (2.4 – 8 $Kbit/s$),

very low rate coders ($<$ 2.4 $Kbit/s$).

5.3.2 Coding Schemes

The general function of a speech coder is to analyze the signal, to remove the redundancies, and to efficiently code the non-redundant part (*i.e.* the innovation) in a perceptually suitable manner [Kondoz94, page 120]. Taking into account the target bit rate, and the target subjective quality, three main classes of coding schemes can be devised: waveform coders, vocoders, and hybrid coders. We will first introduce the concepts of pitch prediction and LPC analysis since they are closely related to the concepts of long-term redundancies of speech, *i.e.* related to the periodicity or pitch of the speaker's voice, and the concept of short-term redundancies of speech introduced by the shape of the speaker vocal tract [Gibson84]. We will then address and illustrate each one of the above classes of coding schemes.

5.3.2.1 LPC Analysis and Pitch Prediction

Time and frequency are both important domains of signal analysis, although the frequency-domain has the potential to reveal features of the speech signal (such as phase) not obvious from its time-domain representation. Both time-domain techniques and frequency-domain techniques are used to extract parametric information of the speech signal such as voiced/unvoiced behavior, pitch period, and formant structure. Given their importance, we will review two important tools present in most speech coders: *linear predictive coding* (LPC) and *pitch prediction*.

LPC analysis The objective of the LPC analysis is to model the short-term correlation between speech samples (formants), in order to reduce the existing redundancy through an efficient decorrelating short-order filter [Kondoz94]. As we will explain, this is equivalent to the removal of the non-flat spectral envelope existing in the signal spectrum. The LPC analysis is closely related to a source model of speech production, built upon the simplified assumption of a two-state (voicing/unvoicing) system. According to this model, the combined contributions of the glottal flow and of the vocal tract, are represented by a time-varying digital filter containing zeros and poles. This filter is usually approximated by an all-pole filter (several poles approximate the effect of a zero). Voiced speech is approximated by exciting this filter with an impulse train reflecting the pitch period, and unvoiced speech is approximated by exciting the filter with random noise. The LPC difference equation of the complete model is given by:

$$s(n) = Gx(n) + \sum_{j=1}^{P} a_j s(n-j) \ , \tag{5.1}$$

where $x(n)$ is the excitation signal, G is the gain, $s(n)$ is the output speech, and a_j are the LPC coefficients. The corresponding system function results as:

$$H(z) = \frac{S(z)}{X(z)} = \frac{G}{1 - \sum_{j=1}^{P} a_j z^{-j}} = \frac{G}{1 - A(z)}\,. \tag{5.2}$$

Naturally, the stability of this system is guaranteed as long as $1 - A(z)$ has the minimum phase property, *i.e.* all of its zeros should be inside the unity circle. The coefficients a_j are usually estimated by minimizing the mean square of the error $e(n)$ [Kondoz94]:

$$e(n) = s(n) - \sum_{j=1}^{P} a_j s(n-j)\,. \tag{5.3}$$

This minimization problem results in a Toeplits matrix that can be solved using the Levinson-Durbin recursion algorithm [Jayant84].

The prediction residual $e(n)$ consists in the excitation $Gx(n)$ signal that exhibits a much lower variance than the input signal $s(n)$. In addition and as expected (given the objective of reducing the correlation between speech samples), the spectrum of $e(n)$ tends to be flat. Another relevant consideration is that the LPC envelope matches the signal spectrum more faithfully in the spectral peaks than in the spectral valleys. This is explained by the absence of zeros in the transfer function $H(z)$ [Kondoz94]. It is therefore important to realize that the residual excitation signal carries all the information that has not been captured by LP analysis such as phase and pitch information, and zeros due for example to nasal sounds [Spanias94].

The core of the LPC framework consists in the computation and coding of the coefficients of the LPC filter. The quality of the encoded speech will be determined by the coding approach concerning the prediction residual:

- a faithful waveform coding delivers high quality but at the cost of an increased bit rate;
- a pure parametric modeling (assuming the two-state model) comprising the gain, the pitch period and the binary voicing/unvoicing decision, simply discards the prediction residual and, as a consequence, delivers synthetic quality but reduces substantially the bit rate.

Pitch prediction The prediction residual obtained after LPC inverse filtering exhibits a periodic structure related to the pitch period. This redundancy can be reduced by a second stage of prediction. The objective of pitch prediction is therefore to model the long-term correlation of speech samples, *i.e.* to exploit the correlation between the speech samples separated by an integer multiple of the pitch period or, in other words, to further spectrally flatten the prediction residual (innovation). The transfer function of the long-term predictor is usually specified as [Kondoz94]:

$$P(z) = \frac{1}{1 - \sum_{j=-L}^{L} b_j z^{-(j+T)}} \ , \tag{5.4}$$

where T is an estimate of the pitch period, b_j are the coefficients of the pitch predictor, and $2L+1$ is the number of taps of the pitch predictor. Methods for the estimation of the pitch period T include time-domain techniques such as the auto-correlation pitch detector, and frequency-domain techniques such as the harmonic peak detector[1] [Kondoz94, page 60]. The coefficients b_j are found by minimizing the mean square error in the long-term periodicity prediction.

5.3.2.2 Waveform Coders

Waveform coders attempt to minimize an objective measure of the quantization noise based on the fidelity of the coded signal to the original signal. As a consequence, waveform coders are not speech specific which makes them particularly suitable to deal with both speech and voice-band data (such as modem and facsimile data). Waveform coders are typically high rate coders ($> 16\ Kbit/s$) and provide toll quality.

The ITU-T[2] G.711 64 $Kbit/s$ coder [ITU72] with logarithmic companding (A-law or μ-law) is a simple waveform coder. Another example very popular is the G.721 adaptive differential PCM (ADPCM) 32 $Kbit/s$ standard [ITU88a]. The general structure of this predictive coder is depicted in Figure 5.1. The algorithm uses an adaptive pole-zero predictor (six poles introduced by $A(z)$, two zeros introduced by $B(z)$), and an adaptive quantizer to code each prediction residual using a 4 bit word. The two zeros lead to a better performance in the modeling of signals exhibiting zeros such as in the case of nasalized sounds [Spanias94]. They also contribute to improve the prediction performance in the presence of transmission errors[3] [Gibson84].

This standard illustrates that the predictor and the quantizer are two important components of a predictive coder, the former being responsible for redundancy reduction and the latter being responsible for entropy reduction (since the difference between the input speech sample and the predicted value is quantized to a limited number of levels [Gibson84]).

Both the step-size of the quantizer and the predictor of the G.721 standard are adaptive so as to track the time-varying statistics of both speech and voice-band data signals. The adaptation is based on the previously quantized residuals and decoded samples (backwards adaptivity),

[1] In Chapter 6, we will present a new frequency-domain pitch estimator.

[2] The *International Telecommunication Union*, Telecommunications Sector (ITU-T) is the new designation of the former *International Consultative Committee for Telephone and Telegraph* (CCITT).

[3] Given that the introduction of zeros in a predictor allows the reduction of the number of poles, the memory effects due to an error are reduced.

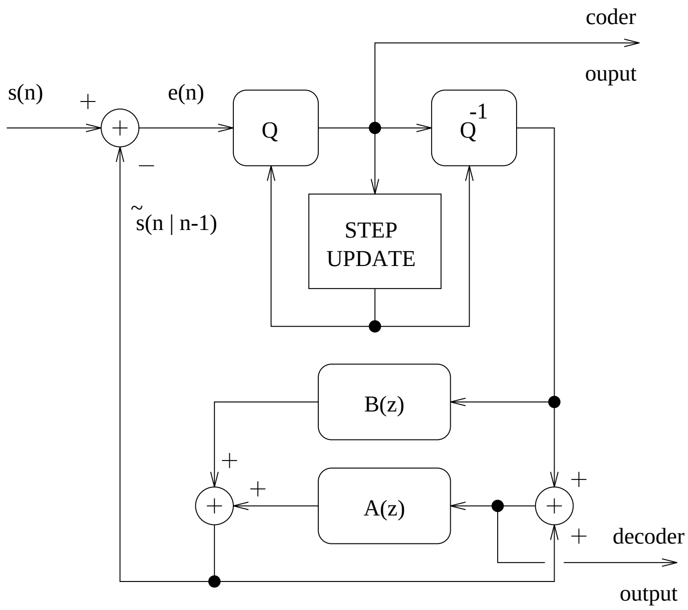

Figure 5.1: Simplified structure of the ITU G.726 ADPCM standard.

which makes the algorithm amenable to low delay coding.

The G.723 standard [ITU88c] extends the operation of the G.721 algorithm to 24 $Kbit/s$ and 40 $Kbit/s$ coding. Both G.721 and G.723 standards have been recently replaced[4] by the ITU-T G.726 standard which, besides maintaining backwards compatibility, has also extended the operation of the ADPCM coder down to 16 $Kbit/s$.

It should be noted that it is possible to take into account auditory masking effects using these types of predictive coders by modifying the (white) spectral profile of the quantization noise. In fact, the spectrum of the quantization noise can be colored as desired by feeding the difference between the input and the output of the quantizer, back to the speech signal, by means of an appropriate noise-shaping filter. This technique is known as *noise feedback coding* [Jayant84], and provides improvements in quality as long as the quantization noise is not too coarse.

5.3.2.3 Vocoders

Contrary to waveform coders, vocoders (*vo*ice *coders*) are speech specific since they synthesize speech using a compact set of parameters related to a model of speech production (*e.g.* the two-state voicing/unvoicing model), with no attempts to preserve the waveform of the original speech. Vocoders operate predominantly in the region below 4.8 $Kbit/s$ and deliver synthetic quality.

[4] For this reason, the ITU-T G.723.1 now refers to a different standard for a 5.3 $Kbit/s$ and 6.4 $Kbit/s$ dual rate speech coder using a CELP based coding approach.

Therefore, they are not evaluated objectively such as through the SNR distortion measure, but rather through subjective measures such as the *Mean Opinion Score* (MOS) [Jayant84].

Linear predictive vocoders are strongly based on the LPC speech analysis but their major difference to a DPCM coder is that the quantized prediction error is not transmitted. Only the analytical parameters such as the pitch period, the voicing/unvoicing decision, the gain (or the energy), and the LPC coefficients are transmitted. Using this information, the receiver generates an excitation consisting in random noise for unvoiced frames, or in a train of impulses based on the transmitted pitch period for voiced frames. The LPC coefficients are transmitted to the decoder in the form of *Reflection Coefficients*, also called *par*tial *cor*relation (PARCOR) coefficients, or *Log Area Ratios* (LAR's), or *Line Spectrum Pairs* (LSP); in order to avoid sensitivity and instability (in the worst case) due to quantization and transmission errors [Kondoz94].

5.3.2.4 Hybrid Coders

Hybrid coders combine features from waveform coders and vocoders and therefore benefit from the advantages of both schemes. Hybrid coders may be further classified into two sub-categories: time-domain hybrid coders, and frequency-domain hybrid coders.

Time-Domain Hybrid Coders Time-domain hybrid coders rely on LPC analysis (short-term prediction) and pitch prediction (long-term prediction) to derive the parameters of the filter-based source production model of speech. These parameters are encoded and transmitted to the decoder in the case of forward adaptation. The excitation signal (*i.e.* the prediction residual) is encoded and transmitted to the decoder, with an accuracy depending on the algorithm. For example, the algorithm known as *residual excited linear predictive coding* (RELP), transmits only the low frequency region of the excitation, where most of the residual information is assumed to be concentrated [Kondoz94]. More efficient algorithms assume that the parameter extraction, the quantization and coding procedures are not separable if a better control of the whole encoding process is to be achieved. This approach, which leads in general to a minimized distortion and thus to an improved quality, is referred to as *analysis-by-synthesis* (AbS).

AbS algorithms use a local decoder in the encoder such that the synthesized speech is compared to the input speech and a related distortion measure is minimized by searching the optimal parameters and excitation in an iterative, closed-loop procedure (instead of an open-loop procedure as in the previous algorithms). One such example is the *multi-pulse* LPC algorithm (MPLPC) which is being used for aeronautical telecommunications (*Skyphone Service*) [Gersho94], using a bit rate of 9.6 $Kbit/s$. This algorithm parametrizes the excitation signal as a set of pulses whose locations and magnitudes are found using an AbS procedure.

Another example of AbS LPC predictive coder is the regular-pulse excited LPC algorithm (RPELPC) used in the *Pan-European Digital Cellular Mobile Radio System* (GSM), which operates at 13 $Kbit/s$ and whose overall delay exceeds 60 ms. The algorithm combines features from RELP and MPLPC. The main relevant aspect is that the optimal excitation is constrained to consist of equally spaced pulses of different amplitudes and only the baseband of the excitation with highest energy is effectively encoded.

The most relevant AbS LPC predictive coder is perhaps the Code-Excited LPC coder (CELP) [Schroe.85]. CELP based coders are medium to low-rate coders and their simplified structure is represented in Figure 5.2. The coefficients of the long-term predictor, $A_L(z)$, and of the

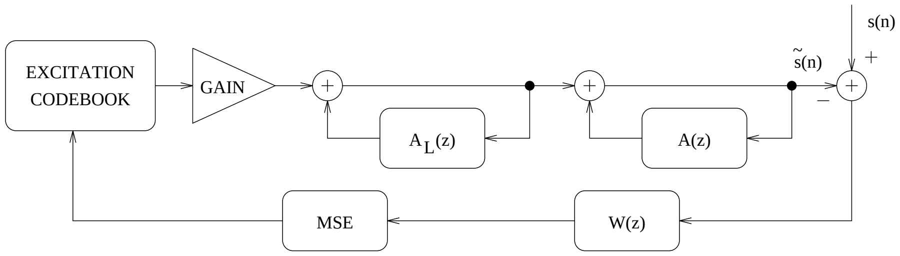

Figure 5.2: Simplified structure of a CELP coder.

short-term predictor, $A(z)$, are computed and coded using conventional (open-loop) methods. However, the coefficients of the long-term predictor can be computed in a closed-loop manner [Kondoz94, page 179], after those of the short-term predictor are calculated in open-loop. Due to the speech dynamics, the long-term (pitch) predictor is updated more often than the short-term (formant) predictor. The building block in Figure 5.2 represented as the excitation codebook, selects by closed-loop an excitation sequence that minimizes a measure (typically the mean square) of the approximation error (between coded and original speech) weighted by the perceptual filter $W(z)$. Although not explicitly indicated in Figure 5.2, the coded information sent to a CELP decoder comprises: the LPC and pitch prediction filter coefficients, the excitation gain and the excitation code-vectors. The perceptual weighting filter $W(z)$ exploits the properties of the human auditory system (although in a very modest way) by reducing the quantization noise in the spectral valleys of the speech signal, and by increasing it near the spectral peaks. This is achieved by de-emphasizing the error energy in the formant regions through a filter that is based on the short-term predictor, and that broadens the bandwidth of the formants. It is frequently implemented using the following form:

$$W(z) = \frac{1 - A(z)}{1 - A\left(\frac{z}{\gamma}\right)}\,, \quad 0 \leq \gamma \leq 1\,. \tag{5.5}$$

The underlying rationale is that in the formant regions the quantization noise is partially *masked* by the speech signal.

CELP coders deliver very good coding quality, even at low bit rates[5], as a consequence of the AbS procedure, and of the benefits gained by using efficient coding techniques such as vector quantization (VQ) [Gray92] to represent the excitation signal. The computational penalty due to the exhaustive search of the optimal excitation using VQ may be attenuated by using multiple codebooks. This strategy offers reduced search and storage complexity, as well as greater robustness to channel errors [Gersho94]. CELP coders differ significantly in the organization and utilization of multiple excitation codebooks. To a certain extent, there exists a trend in using two types of codebooks: adaptive and stochastic [Noll95]. The adaptive codebook handles the pitch periodicity (which ultimately may eliminate the need for a long-term synthesis filter) and the stochastic codebook handles the non-periodic (stochastic) excitation component [Kroon92]. The coding quality is further improved, particularly in the case of tandeming, by using post-filters which emphasize[6] the formant and pitch structure of pitch. The coding delay of a typical CELP coder is in the order of 60 ms.

A well known CELP algorithm is the 16 $Kbit/s$ low-delay CELP coder (LD-CELP), which has been adopted by the ITU-T, in 1992, as an international standard. Its most outstanding feature is the extremely low system delay (less than 5 ms), which is achieved by using backward estimation of the filter coefficients and of the gain. The coding quality at 16 $Kbit/s$ has been reported to be equivalent or better than that of the G.721 ADPCM standard at 32 $Kbit/s$ [Kondoz94].

Other standardized CELP based coders are the ITU G.729 and its simplified version, the G.729 Annex A. Both algorithms are based on conjugate-structure algebraic[7] code excited linear prediction (CS-ACELP), and were designed to code speech and data at 8 $Kbit/s$ with a quality equivalent to that of G.726 at 32 $Kbit/s$ [Salami97].

Frequency-Domain Hybrid Coders Frequency-domain hybrid coders use techniques that take advantage of the short-time spectral structure of the speech signal in such a way as to:

- perform a convenient signal analysis which may include the derivation of parametric information related to a source model,

[5] In this sense, CELP coders establish a bridge between waveform coders and vocoders.

[6] This approach may however introduce a slight distortion in the speech signal because the bandwidths of the spectral peaks are reduced.

[7] This designation stems from the fact that the algorithm uses a deterministic (algebraic) codebook whereby the excitation code-vector is derived from the transmitted codebook index. Therefore, there is no need for codebook storage.

- code different spectral regions in an independent way,
- implement an accurate shaping of the quantization noise according to perceptual rules.

The spectral division into several subbands by means of a filter bank with good stop-band characteristics leads to subband components (subband samples or transform coefficients) nearly uncorrelated. This represents an excellent opportunity for redundancy reduction since the energy of a *structured* signal is likely to be compacted in a small number of subband components. In a typical speech or audio coder, the spectral decomposition building block reflects the theory presented in Chapter 4. Therefore, in this paragraph, we focus on coding techniques related to spectral modeling, coding, and noise shaping.

Three classic references, all published in the late seventies [Zelins.77], [Schroe.79], [Tribol.79], stimulated the interest in (and paved the way to) the extension of frequency-domain techniques developed in the context of speech coding, to the coding of wideband audio signals, eliciting the importance of perceptual criteria. After we review these references, we will address briefly sinusoidal coding, where speech is constrained to a harmonic or sinusoidal representation.

The FIRST CLASSIC REFERENCE, the *Adaptive Transform Coder* (ATC) developed by Zelinsky and Noll [Zelins.77], uses a 128 point DCT and a coding criterion based on the maximization of the *objective* coding gain over PCM (Section 4.3.1.6). It is recognized that the coding process must be adapted to the *local* statistical structure of speech sounds, in order to reach high coding performances. Accordingly, the optimal bit assignment is recalculated for each block of data, which implies that a spectrum descriptor must be sent to the decoder as side information. In order to reduce the amount of side information, a sub-sampled version of the log-averaged spectrum is quantized and transmitted. Both encoder and decoder use this information to estimate the spectrum (which is assumed to be smooth), and to derive the optimal bit assignment which determines the quantization of all spectral coefficients. A schematic representation of the encoder is presented in Figure 5.3. A decoder is obtained by reversing the data paths and by eliminating the building block of "smooth spectral modeling".

Since the bit assignment as well as the quantizer step-size are a function of the time varying statistics of speech, the overall coding noise is far from white (*i.e.* locally it is white but its level varies over time). The authors have noted that since this fact takes advantage of (perceptual) masking effects, the quality of speech coded at 12 $Kbit/s$ is very good, exceeding that of other coding schemes using fixed bit assignment (*i.e.* based on the long-term statistics of speech), under similar operating conditions.

In a paper that represents the SECOND CLASSIC REFERENCE, Tribolet and Crochiere proposed [Tribol.79] an ATC based algorithm modified in such a way as to point out that '*the objective*

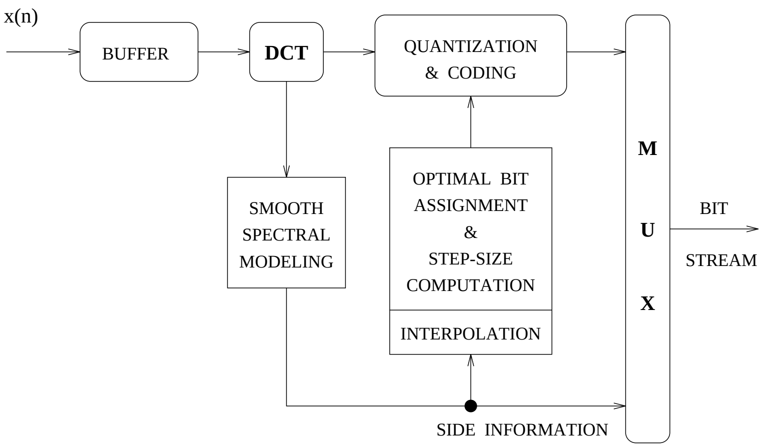

Figure 5.3: Simplified block diagram of the Zelinsky and Noll Adaptive Transform Coder.

of frequency domain coding is to isolate perceptually important components of the short-time spectrum for encoding', and as to recognize that '*the mean square criterion is not necessarily the most appropriate criterion in terms of speech perception*'.

The algorithm implements a spectral parametrization that is more speech specific than the Zelinsky and Noll ATC algorithm, for several reasons, as can be concluded from Figure 5.4.

First of all, the sub-sampled description of the spectral envelope is made non-uniform according to the articulation scale [Jayant84] that has a good correspondence to the critical bands at low frequencies. Although the motivation has been the observation that at low frequencies, formants are narrower than at high frequencies; this can also be seen as the inclusion of perception aspects in the encoder.

The spectral envelope is found by recognizing first that the DCT spectral envelope is bounded by the spectral envelope of the DFT[8], and thus it also reveals the existing formant structure and pitch structure. Using the *Wiener Khinchine* theorem [Jayant84], the DCT spectrum is squared and inverse DFT transformed to obtain a pseudo-autocorrelation function. The resulting first few values are used to obtain an LPC filter which is coded as LAR coefficients. These are transmitted to the decoder as side information. The inverse spectrum of this filter finally yields an estimate of the formant structure (*i.e.* the spectral envelope) of the DCT spectrum.

Second, a spectral pitch pattern is also estimated such that when it is multiplied by the spectral envelope estimate previously computed, a global spectral estimate of the speech signal is obtained. The gain and the pitch period are estimated from the pseudo-autocorrelation function and are transmitted to the decoder as side information. The global spectral estimate is then

[8]In Chapter 6 we will show a similar relation between the oddly stacked TDAC filter bank and the Odd-DFT.

used for the bit allocation and step-size adaptation algorithm. The bit allocation algorithm uses Equation (4.28) where a frequency weighting is included (as we pointed out in Section 4.3.1.7), such that the quantization noise is effectively masked by the speech signal [Tribol.79]. The net effect of this frequency weighting is illustrated if Figure 5.4, and depends only on the spectral envelope estimate (and not on the estimate of the spectral pitch pattern). Its purpose is to shape the quantization noise between two limits:

1. **flat** - which means the bit allocation is such that the SNR has the shape of the signal spectrum,

2. **paralell to the input spectrum** - which means the bit allocation is constant and therefore the SNR is constant as a function of the frequency.

The chosen degree of frequency weighting, as reported in the paper, makes that more bits are assigned to those regions of the spectrum containing formants and pitch harmonics [Crochi.79, page 79]. Using this perceptual noise shaping and a 256 point DCT, the quality delivered by

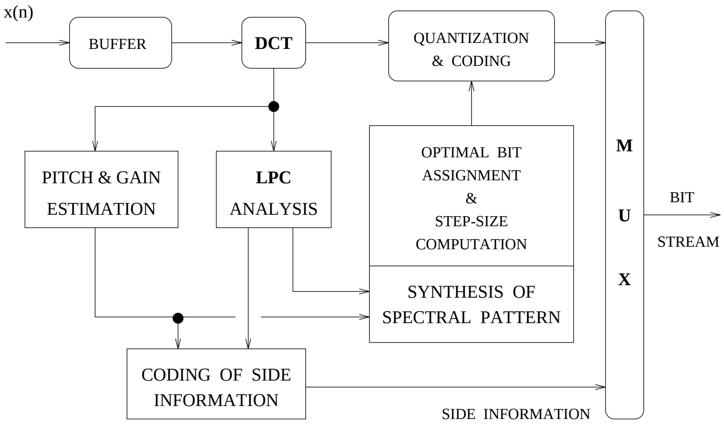

Figure 5.4: Simplified block diagram of the Tribolet and Crochiere speech specific ATC.

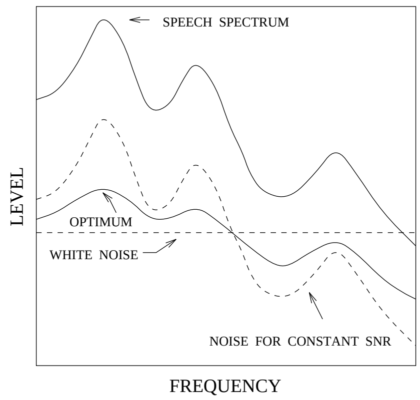

Figure 5.5: Flexible shaping in frequency of the quantization noise between two limits: white noise and constant SNR in frequency.

the algorithm exceeded that of the Zelinsky and Noll coder for similar working conditions, and proved to be usable down to 9.6 $Kbit/s$. The coding quality at this particular bit rate has been improved in a enhanced version of the algorithm [Tribol.80] that includes three major modifications.

1. DYNAMIC PREEMPHASIS: the speech spectrum is flattened prior to analysis/synthesis by means of an adaptive LPC filter, in order to reduce the frequency domain aliasing due

to the quantization of the speech short-term spectrum, and in order to use normalized quantizers. This operation is equivalent to '*a division of the DCT spectral coefficients*[9] *by the estimated speech spectrum*' [Tribol.80].

2. RESTORATION OF HARMONICITY : harmonics are artificially synthesized where spectral holes appear as a consequence of insufficient (zero) bit assignment. The restoration is achieved by means of a suitable shift of low frequency harmonics, which also helps to attenuate the loss of signal bandwidth in low bit-rate coding.

3. POST-PROCESSING: after decoding, a signal analysis and enhancement is applied so as to reduce the aliasing noise existing in bands with no coded speech signal.

As a consequence of these modifications, significants improvements in quality were reported, indicating that the enhanced coder was usable down to 7.2 $Kbit/s$.

The Crochiere and Tribolet paper of 1979 also discusses other insightful aspects[10] such as ways to control the effects due to frequency domain aliasing, the additive or multiplicative nature of the quantization as a result of the large or low number of bits available to code the quantized coefficients, and the reduction of block boundary errors by means of overlapping analysis/synthesis windows.

The THIRD CLASSIC REFERENCE [Schroe.79] does not propose a specific coder but presents an entirely new approach to '*optimizing digital speech coders by exploiting masking properties of the human ear*'. In fact, Schroeder, Atal and Hall argue that speech quality does not really depend on the absolute level of the noise but rather on its loudness (Section 3.3) relative to the loudness of the main speech signal. Therefore, instead of minimizing an objective measure (such as the least squares) of the quantization noise, a coder should minimize an objective measure of its subjective loudness (*i.e.* relative to the loudness of the speech signal) that, in turn, depends on perceptual masking effects (Section 3.3.3). The degree of partial masking, (Sections 3.3.3 and 3.3.4), *i.e.* the reduction of the loudness of the quantization noise due to a masking signal (speech) in the same critical band, is assumed to be analytically given by:

$$N' = \frac{N}{1+\left(\frac{S}{N}\right)^2} \,, \tag{5.6}$$

where N and S are the noise and signal powers, on a critical band basis, and N' is the power of the equivalent unmasked noise matched to have the same loudness. The total loudness L_N of

[9] We shall see later that even high-quality audio coders use this estimate to flatten the spectrum of the signal prior to the quantization and coding.

[10] It is the opinion of the author of this Dissertation that this paper has represented a major milestone in the area of frequency domain coding of speech, inspiring and paving the way to subsequent work in wideband speech and high-quality audio coding.

the noise in the presence of a speech signal is obtained as:

$$L_N = \int_{0.5}^{24.5} \left(N''\right)^{0.25} dx \;, \tag{5.7}$$

where x is the critical band rate and N'' is obtained from N' by considering that $L_N = 0$ when the noise falls below the Masked Threshold or below the Absolute Threshold of Hearing. $L_N = 0$ denotes total masking, *i.e.* the noise is inaudible. The objective measure of the subjective loudness of the quantization noise is then taken as[11]:

$$D = \frac{L_N}{L_S} \;, \tag{5.8}$$

where L_S is the loudness of the signal.

Schroeder, Atal and Hall propose the use of this measure in a predictive coder or transform coder, in order to *hide* (mask) the quantization noise under the signal or, if the bit rate is unsufficient, in order to reduce its loudness as much as possible. This new approach to speech coding based uniquely on perceptual criteria, laid down the foundations to an increasing interest concerning the application of perceptual models to audio coding. In fact, the merit of this approach is particularly obvious in those cases where there is a lack of a reliable source model, and classical assumptions like Gaussianity, quasi-stationarity, and spectral smoothness, can not be assumed. For example, the Psychoacoustic Model II of the MPEG standard (Section 3.3.4) estimates the masked threshold using the procedure presented in this paper of 1979, with the exception for tonality aspects.

Sinusoidal Coding An important sub-class of frequency domain hybrid coders offering good coding performances below 9.6 $Kbit/s$ down to 2 $Kbit/s$, is known as sinusoidal coding. The sinusoidal model assumes speech can be expressed as a combination of a time varying number of L sinusoids with time-varying amplitudes, frequencies, and phases:

$$s(n) = \sum_{k=1}^{L} A_k \cos\left(\omega_k n + \phi_k\right) \;, \quad n = 0, \ldots, N-1 \;. \tag{5.9}$$

The opportunity to reduce the bit rate using this model lies in the fact that voiced speech is typically highly periodic and thus it can be represented by a constrained set of sinusoids. On the other hand, the statistical structure (reflected in the short-time spectrum) of unvoiced speech can be preserved by a sinusoidal model with appropriately defined random phases [Spanias94, page 1554]. Two particular implementations of sinusoidal modeling are known as the *Harmonic Coder* and the *Multiband Excited Vocoder* (MBEV).

[11] It should be noted that this approach conflicts with our conclusions concerning the perceptual coding gain (Section 4.3.1.7), except in the case of overcoding.

The harmonic coder identifies spectral peaks harmonically related in each successive frame of the speech, and encodes and transmits their frequencies, magnitudes, and in some cases, their phases. The receiver synthesizes speech with similar time-varying spectra as the original speech, by controlling the frequencies, the magnitudes and phases of the harmonic sine waves. However, a realistic reproduction of unvoiced speech using harmonic modeling is difficult because of the tonal nature of the coding noise. The problem can be alleviated by removing the constraint on harmonic frequencies, and by forcing a randomization of the phases, or by using an alternative coding method for unvoiced regions such as the ATC [Tranco.88].

The MBEV is illustrated in Figure 5.6 and takes advantage of the fact that the spectra of mixed sounds or noisy speech contain both voiced (harmonic) and unvoiced (random-like) regions. The spectrum is divided into subbands which are centered around the pitch-harmonic

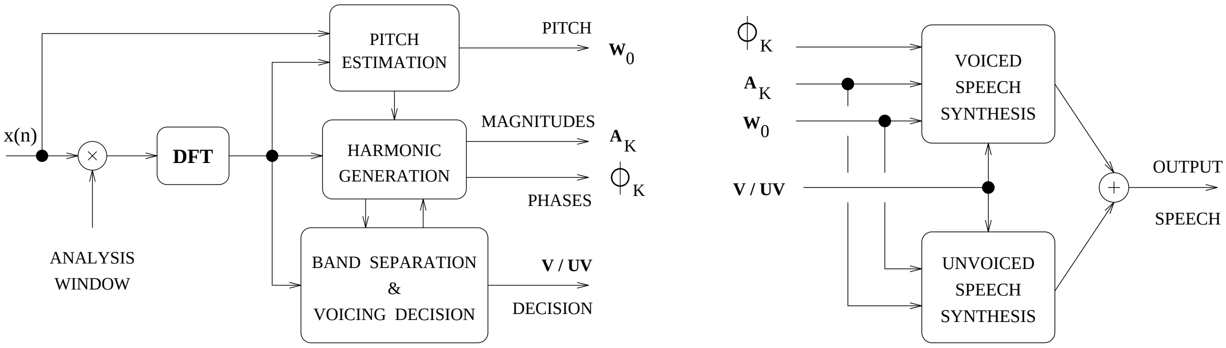

Figure 5.6: Simplified schematic representation of a Multiband Excited encoder (on the left) and decoder (on the right).

frequencies, and each subband is declared voiced or unvoiced using a measure of the difference between original and synthetic spectra in each band. This partition represents the excitation spectrum. The short-time speech spectrum is seen as the product of an excitation spectrum and a vocal tract envelope. The vocal tract envelope is a smoothed version of the short-time speech spectrum and can be obtained by interpolating between the harmonic peaks of the spectrum, or by LPC analysis.

The estimation of the excitation and envelope parameters may be computed simultaneously using an analysis-by-synthesis approach. The information transmitted to the decoder consists of: the fundamental frequency (the pitch), the spectral envelope, the voicing information, and the phase information for voiced harmonics. The decoder synthesizes separately the voiced and unvoiced components and then combines them in the time-domain to produce speech (Figure 5.6).

Two aspects of particular relevance in the MBEV concern pitch estimation and phase synthesis. The pitch is usually estimated using the autocorrelation function and then is refined in the frequency domain to get fractional accuracy. In the case of very low bit rate coding, the phase is not transmitted to the decoder. Phase information is therefore predicted from the current and previous pitch frequencies by considering that at harmonics, the phase should be such that all harmonics add up so as to produce the harmonic peak [Kondoz94, page 131].

5.4 Wideband Speech Coding

In a large number of applications involving the communication of voice signals, such as in the case of audio-conferencing, high-quality speech is often desired. High quality speech is obtained by extending the low frequency content of the basic speech bandwidth down to about 50 *Hz*, and by extending the high frequency content up to about 7 *kHz*. The low frequency content improves the sensation of presence of the speaker, and the impression of naturalness of the speech sounds. The high frequency content improves the intelligibility, and the impression of crispness of the speech sounds [Jayant93]. The uniform PCM representation of wideband speech is typically obtained using 16 *kHz* sampling frequency and a sample resolution of 14 or 16 bit. In the following, we will describe briefly two similar wideband speech coding algorithms: the ITU-T G.722 standard and the APT-X100. Other wideband speech coders exist that are simple extensions of CELP coders[12], or simplifications of perceptual transform coders[13]. We will also describe briefly one example of each one of these approaches, as well as one example combining both approaches.

The G.722 ITU-T Standard

A well-known and widely used wideband speech coder is specified by the recommendation '*7 kHz audio coding within 64 Kbit/s*' [ITU88b], which was standardized by the former CCITT in 1986. The algorithm compresses 16 *kHz* sampled and 14 bits uniformly quantized speech signals, representing a data rate of 224 $Kbit/s$, into a digital representation of 64 $Kbit/s$.

G.722 is based on a two-band division of the spectrum by means of a QMF, followed by ADPCM coding (SB-ADPCM) of each decimated subband signal. The structure of the algorithm is depicted in Figure 5.7. The QMF is based on two 24-tap linear phase filters with the high-pass filter related to the low-pass filter by $h_H(n) = (-1)^n h_L(n)$. In the frequency domain, this is equivalent to $H_H(z) = H_L(-z)$. In the absence of coefficient quantization, this relation insures aliasing cancellation but not perfect reconstruction (Chapter 4).

[12]Typically, CELP coders operate at or below 16 $Kbit/s$.

[13]Typically, perceptual audio coders operate at or above 64 $Kbit/s$.

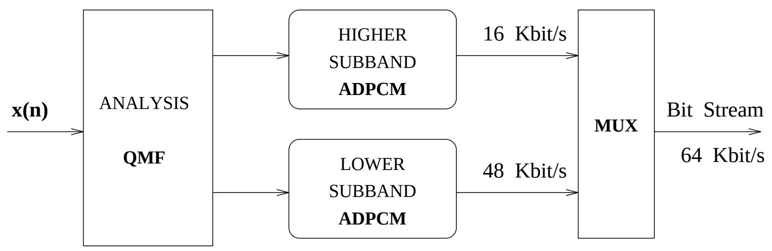

Figure 5.7: Simplified structure of the ITU G.722 SB-ADPCM standard.

The ADPCM encoders are structurally similar to that illustrated in Figure 5.1 for the G.726 standard. The main difference is that a non-linear 6-bit quantizer is used for the lower band, while a non-linear 2-bit quantizer is used for the upper band.

An auxiliary data channel may be accommodated by using one or the two least significant bits of the quantized values pertaining to the lower subband. Accordingly, three modes of operation are possible corresponding to a data channel rate of 0, 8, or 16 $Kbit/s$. The mode of operation must be signaled to the decoder, externally. Mistracking problems in the decoder are avoided by insuring identical feedback operation at the encoder and decoder. This is achieved by using only the 4 most significant bits of the lower subband quantized samples in the quantizer step-size adaptation, as well as in the inverse adaptive quantizer. The system delay of the G.722 coding standard is about 3 ms.

The G.722 coding standard is a well established reference for the assessment of new algorithms targeting 7 kHz bandwidth speech coding, at bit rates below 64 $Kbit/s$, while preserving that subjective quality of G.722 coded signals. In fact, in recent years, a number of proposals has been made to reduce that bit rate down to 16 $Kbit/s$, which is in line with recent ITU-T standardization objectives [Chen97].

APT-X100

APT-X100 [Wylie96] is a wideband speech coder that has gained a large acceptance in many application fields because of its low system delay (7.6 ms for 16 kHz sampling frequency) and robustness to tandeming. It achieves 4:1 compression independently of the sampling frequency. Its structure is essentially equivalent to that of the G.722 standard, with the main difference that the QMF splits the incoming signal into four bands, and that the subband signals are then ADPCM encoded to 7 (lowest subband), 4, 3 and 2 (highest subband) bits/sample.

Wideband CELP

Ordentlich and Shoham [Ordent.91] have adapted a Low Delay-CELP coder (LD-CELP), originally designed to code the standard telephony band at 16 $Kbit/s$, to the coding of 7 kHz wideband speech at 32 $Kbit/s$. The authors present an interesting conclusion that justifies the adaptation: 'The incorporation of a noise weighting filter, $W(z)$, in a CELP coder, decreases the *Mean Square Efficiency* of the coder. Therefore, the efforts should be concentrated on the perceptually efficient spectral shaping of the increased overall noise level, taking advantage of the fact that noise is less audible if it shares the same spectral band of a high level signal.'

In order to implement an efficient shaping of the coding noise, the authors have developed a perceptual weighting filter comprising two reasonably decoupled components. One of them has a more general form than Equation (5.5), and is mainly devoted to the control of the formant modeling (although it also affects the *spectral tilt* [Ordent.91]). The other component consists of an all-pole filter devoted to the control of the spectral tilt. The authors have reported [Ordent.91] that this improvement on the 32 $Kbit/s$ CELP coder, lead to a subjective quality of the coded material, identical to that delivered by the G.722 standard at 64 $Kbit/s$ coding.

Wideband Transform Coder

Quackenbush [Quacke.91] describes a two channel wideband speech coder, as well as its real-time implementation for transmission via a single channel B of an ISDN access. The structure of the coder reflects the perceptual coding paradigm (Figure 1.3) since it is based on a transform audio coder [Johns.88b]. The coder uses dynamic bit allocation in the frequency domain according to a perceptual noise weighting criterion, and uses Huffman coding of quantized data. The psychoacoustic model is based on the Schroeder, Atal and Hall model [Schroe.79]. The total system delay is about 75 ms.

Transform Predictive Coder

Chen [Chen97] describes a hybrid approach to wideband speech coding. It combines CELP coding techniques with transform coding techniques and consists in one of the candidate coders for the ITU-T new wideband speech coding standard at 16, 24 and 32 $Kbit/s$. The coder is a variant of the "Transform Predictive Coder" (TPC) [Chen96]. It performs LPC filtering, a closed-loop pitch prediction on perceptually weighted speech, and then quantizes the prediction residual using perceptually based transform coding techniques[14]. At the lower bit rate, the

[14]A similar but entirely open-loop approach (Transform Coded eXcitation-TCX) has been proposed by LeRoux *et al.* [LeRoux95].

coder allocates bits only to frequencies below 4 *kHz*. At higher bit rates, all frequencies till 7 *kHz* are allowed to receive bits. An interesting aspect is that all zero quantized frequency components are synthesized at the receiver using a randomized phase and a magnitude derived from the Signal-to-Mask-Ratio (SMR) values. In turn, these are derived from the quantized LPC spectrum. The decoder duplicates many of the encoder functions since the SMR values are not transmitted, and this obviously increases its complexity.

5.5 High Quality Audio Coding

As we have noted in previous Chapters, high quality audio coders are, in their core structure, almost invariably subband/transform coders. This is explained by the fact that the opportunity to reduce the bit rate of the digital representation of an incoming audio signal, lies in the structure of its short-term power spectrum, and also in the perceptual properties of the human auditory system, that are better exploited in the frequency domain (namely simultaneous masking). In this section, we will describe briefly a few examples of both standardized and proprietary coders that share this perceptual coding approach.

However, initial efforts to compress a high quality digital audio signal have used simple time-domain non-uniform quantization techniques. We will describe first two currently used examples of such an approach for the digital transmission or storage of 15 *kHz* bandwidth audio signals sampled at 32 *kHz*. The technique is known as *companding*.

5.5.1 Companding

Companding consists in the requantization of linear PCM to non-uniform (*e.g.* logarithmic) PCM. This procedure is accompanied by a reduction in the number of bits per sample. Two variants of this procedure exist that are in use today: *instantaneous companding* and *near-instantaneous companding*.

5.5.1.1 Instantaneous Companding

Consists in a conversion, sample-by-sample, using a non-uniform rule. For example, the ITU-T J.41 [ITU88d] standard[15] converts 14 bit linear PCM samples into 11 bit logarithmic PCM samples according to the A-law. Including overhead bits, the J41 standard codes a stereo audio signal into a bit rate of 768 $Kbit/s$. This standard is used, for example, in international digital links.

[15] Instantaneous companding variant described in §4 of the Recommendation J.41.

5.5.1.2 Near-Instantaneous Companding

Consists in the coding of a group of PCM samples into a normalizing scale-factor and an associated group of PCM samples using a coarser resolution than the input. For example, the ITU-T J.41 standard[16] uses blocks of 32 samples, and each block is assigned one scale-factor (out of 5 possible "ranges"), in an adaptive way, according to the highest sample value in the block. The samples are then normalized by the scale-factor and quantized to 10 bits. The resulting companding characteristic is shown in Figure 5.8.

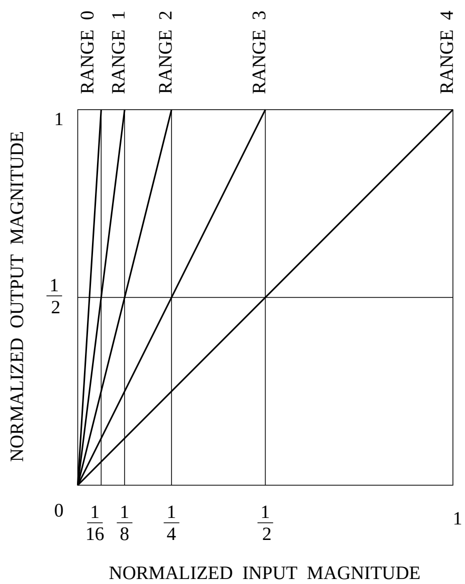

Figure 5.8: Near-Instantaneous companding characteristic of the ITU-T J.41 standard.

A well known example of this technique is the *Nearly-Instantaneous Companding Audio Multiplex* (NICAM) that is widely used for the transmission of sound in broadcast television networks. Including overhead bits, the NICAM algorithm codes a stereo signal using a bit rate of 728 $Kbit/s$. Other more complex variants of this technique are the NICAM-3 and the ITU-T J.43 standard [ITU88e]. NICAM-3 was developed to link studios to FM stereo radio transmitters. It uses protection information and scale-factors for three groups of 32 samples rather than for a single group. This increases the system delay but reduces the bit rate to 338 $Kbit/s$ for each audio channel. The ITU-T J.43 standard compresses further the PCM digital representation by using differential coding techniques in addition to companding. The bit rate of each compressed audio channel is 320 $Kbit/s$.

The main difference between the instantaneous and the near-instantaneous techniques is

[16]Near-instantaneous companding variant described in §5 of the Recommendation J.41.

that, in the first case, the quantization error is made almost proportional to the magnitude of each individual sample, while the second technique generates a quantization error having a fixed level for each sample within the same block. Both schemes deliver an identical subjective quality [Soumag.86]. However, neither of these time-domain schemes reduces significantly the bit rate of the digital audio representation (less than 2:1 compression). It fact, neither statistical dependencies between audio samples nor other auditory masking effects than the sensitivity as a function of the intensity are exploited.

5.5.2 Perceptual Audio Coders

Most frequency-domain audio coding schemes reflect the structure of the perceptual coding paradigm as presented in Section 1.1.3.2. Both objective and perceptual contributions to the coding gain can be made significant because, as we have concluded in Section 4.3.4, a typical audio signal exhibits a quasi-stationary behavior for the most part of its duration. This possibility is realistic for the following main reasons:

- an adaptive decomposition of the audio signal into nearly uncorrelated components can be made effective,
- an accurate signal-adaptive shaping of large amounts of quantization noise in frequency can be made effective in a perceptual sense,
- additional coding gains provided by complex quantization and coding techniques such as Vector Quantization and Entropy Coding can be made significant.

Given the expectation of very high quality in the presentation of audio signals, normally, the goal in audio coding is to achieve transparent coding, *i.e.* the quality of the decoded signal should be *subjectively* indistinguishable from that of the original signal. As the differences between original and coded material are rather subtle, the impairment scale is frequently used to rate the quality of the coding process (Section 5.2).

Many independent coding schemes, including coding standards and proprietary audio coders, achieve transparent coding at compression ratios (ratio between the bit rate of the PCM representation of the original signal and the bit rate of the compressed representation) ranging from 4:1 to 8:1. In general, while maintaining coding transparency, a higher compression ratio is reached at the cost of a higher coding delay and algorithmic complexity.

Taking the Compact Disc (CD) quality as a reference, current technology allows easily the transparent coding of audio signals at $128 Kbit/s$ per channel. In fact, at compression ratios in the order of 6:1, there still exists a large margin of statistical redundancy and irrelevancy unexploited. However, at higher compression ratios, in the order of 10:1, transparent coding implies high encoding efficiency. This is synonymous with the need to maximize of the global encoding performance of the algorithm. As we have discussed in Chapter 2, coding performance depends obviously on the efficiency of each individual building block of the coder, but depends essentially on the joint efficiency of critical building blocks of the coder.

To give a first picture of what current technology is capable of, we may cite specific examples of coding solutions. For example, the Dolby's AC-2 [Fieldi.92] coder, the Sony's ATRAC coder [Tsutsui92] (used in the magnetic-optical MiniDisc) and the Philips's PASC coder [Lokhoff91], [Hoogen.94] (used in the magnetic Digital Compact Cassette) that corresponds to a simplified version of the MPEG-1 Audio Layer I standard, are not highly optimized coding algorithms. However, they have the advantage of enabling very simple (*i.e.* cheap) implementations and achieving very good coding qualities for compression ratios in the order of $4:1$ or $5:1$. On the other hand, the MPEG-1 Audio Layer II and Layer III standard, the MPEG-2 Advanced Audio Coder (AAC) [Bosi96], the AT&T's PAC coder [Johns.92], and INESC's ASC coder [Ferre.96a], are a few examples of moderate to highly optimized algorithms capable of reaching very good coding qualities at compression ratios higher than $8:1$.

In this section, we will illustrate the structure of some of these coding solutions (except ASC which will be detailed in the next Chapter) that we believe are fairly representative of the current state-of-the-art. As most perceptual audio coders make use of techniques already discussed in the context of speech and wideband speech coding, we will give only a high level illustration. We will address in particular algorithms that are MPEG coding standards or that somehow contributed to the MPEG standardization activities. This relation between proprietary coding solutions and MPEG standards can be understood by the fact that the MPEG-Audio is the result of a collaborative effort receiving technological contributions from many companies, research laboratories and universities, as explained in Section 1.1.4. For example, two proprietary coders: MUSICAM[17] [Dehery91] and ASPEC[18] [Brand.91], played the most important role in the definition of the MPEG-1 Audio standard. Other algorithms such as ATRAC have also been submitted but as they were poorly ranked in the MPEG-1 competitive tests, they did not contribute substantially to the MPEG-1 Audio standard. ATRAC found application in consumer electronics [Tsutsui92]. Other proprietary algorithms such as the multichannel PAC and the Dolby AC-3 [Davis93] (which also found specific areas of application such as home and cinema surround sound) were involved in a special standardization phase, within the MPEG-2

[17] MUSICAM has been developed within a European project known as Eureka 147 [Dehery91].
[18] ASPEC gathers coding techniques imported from various proprietary coding schemes [Brand.91].

standardization activities, that lead to the non-backwards compatible AAC standard [Bosi96].

For an improved efficiency, perceptual coders often consider the joint coding of a combination of correlated audio channels. In fact, further gains can be obtained by exploiting interchannel statistical redundancies as well as interchannel perceptual *irrelevancies* [Blauert83][Johns.92]. At the same time, an appropriate control is made of spatial artifacts that would be otherwise generated by coding independently each channel of a multichannel configuration. As we have explained in Section 1.1.4, we do not address multichannel coding in this Dissertation. Therefore, unless stated otherwise, when referring to compression ratios we assume implicitly the coding of a single audio channel.

5.5.2.1 The MPEG-1 Audio Standard

As mentioned in Section 1.1.4, the MPEG-1 Audio coding algorithm [ISO93] is structured in three layers, each one offering a specific trade of bit rate, complexity, delay, and subjective quality. The MPEG-1 Audio standardizes the coding of one or two channels at bit rates ranging from 32 $Kbit/s$ (mono) till 448 $Kbit/s$ (stereo), and admits broadcast (32 kHz), consumer (44.1 kHz) and professional (48 kHz) sampling frequencies. Approximate bit rates for the transparent coding of a single channel are as follows: 192 $Kbit/s$ for Layer I (that has the lowest system complexity and delay of all layers), 128 $Kbit/s$ for Layer II, and 96 $Kbit/s$ for Layer III (that has the highest system complexity and delay of all layers).

The structure of the MPEG-Audio encoder is represented in Figures 5.9, 5.10 and 5.11. The structure of the MPEG-Audio decoder [Brand.92b] can be easily derived from the previous figures with the exception of the psychoacoustic model and associated DFT which are not necessary to the decoding operation.

The structure of the MPEG-1 Audio algorithm reflects features it has inheritted from two algorithms: MUSICAM [Dehery91] and ASPEC [Brand.91]. It is important to note that only the bit stream structure (syntax) and the decoding process are specified in the normative part of the standard. The whole encoding process is not normative and therefore the standard describes, in its informative part, a possible but not necessarily optimal implementation.

The basic section of the MPEG-1 Audio encoder that is common to all three layers is shown in Figure 5.9, and the complementary section concerning Layer I or Layer II encoding is shown in Figure 5.10. The former implements a spectral division of the audio spectrum into 32 uniform subbands by means a Pseudo-QMF, and computes the SMRs. As recommended in the informative part of the standard, SMRs may be computed using a 512-point (for Layer I) or a 1024-point DFT (for Layer II or Layer III) and using one of two psychoacoustic models described in the

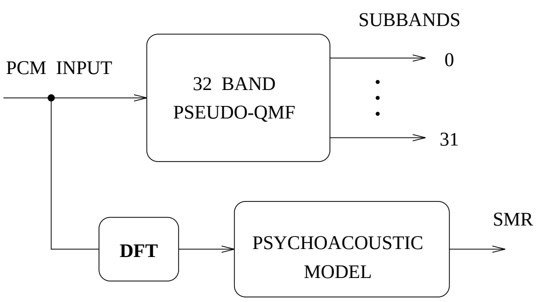

Figure 5.9: Basic section of the MPEG-1 Audio encoder that is common to all layers.

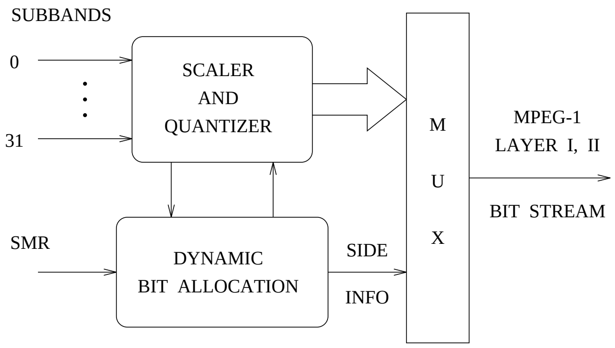

Figure 5.10: Complementary section of the MPEG-1 Audio encoder for Layer I or Layer II.

standard. In Section 3.3.4 we have discussed the main differences between these two models. In Section 4.4.1 we have seen how each model takes into account the global tonality properties of the audio signal and concluded that, in practice, Model 1 is used for Layer I and Layer II, while Model 2 is used for Layer III [Brand.92b]. It is worthwhile noting that either model estimates the maximum amount of quantization noise (denoted by the SMR) allowed within each quantization band or quantization partition, but does *not* specify *how* that amount of noise should be shaped *within* each quantization partition. For Layer I or Layer II a quantization partition corresponds to a subband. For Layer III, a quantization partition corresponds to a group of spectral coefficients in such a manner that the associated spectral partitioning is related to the *Bark* division of the spectrum (Section 3.3.2).

Layer I codes fixed blocks of 384 samples (32×12) using block companding. Layer II codes fixed blocks of 1152 samples ($3 \times 32 \times 12$) using block companding and implements a more effective coding of the side information than Layer I. Layer III has a more complex structure that is based on a hybrid filter bank as shown in Figure 5.11. It implements adaptive segmentation of each Pseudo-QMF output into 12 or 36 samples with 50% overlap, in order to control pre-echo[19] coding artifacts [Brand.92b]. The segmentation is adapted according to a local non-stationarity measure based on the fluctuation of the PE estimates [ISO93]. A 12-point or 36-point MDCT[20] is then applied to each subband to improve the frequency resolution of the analysis filter bank. Each MDCT delivers 6 or 18 coefficients that are processed by an aliasing reduction block for the reasons presented in Section 4.3.4.

It is important to note that the adaptive segmentation gives rise to two possible time/frequency

[19] Pre-echo results when the spreading in time of the quantization noise is such that it is not masked by the audio signal. This may occur when transform encoding non-stationary signals because in the reconstruction, noise will spread over the length of the time synthesis window. We have illustrated the pre-echo effect in Section 3.3.5.

[20] This designation has been frequently used [Noll95] to denote the TDAC Filter Bank (Section 4.3.1.5).

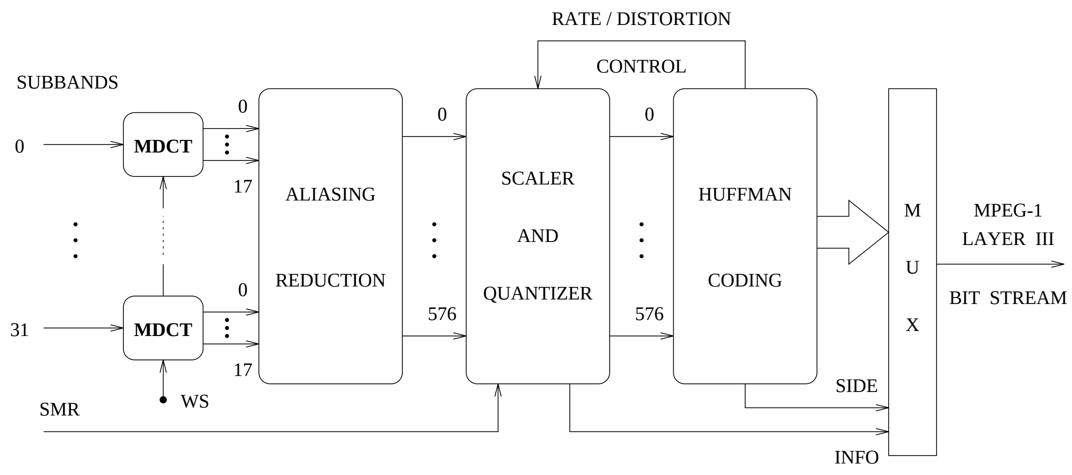

Figure 5.11: Complementary section of the MPEG-1 Audio encoder for Layer III. The diagram represents the finest spectral resolution of the hybrid filter bank. However, the resolution is adaptively controlled by means of a window switching mechanism (WS) that modifies the segmentation on each subband (see text).

resolutions of the hybrid filter bank. In order to make it possible to switch between the two modes, a special combination of windows is used. Specifically, "long", "short", "start" and "stop" windows are used as illustrated in Figure 5.12, where the 50% overlap between adjacent

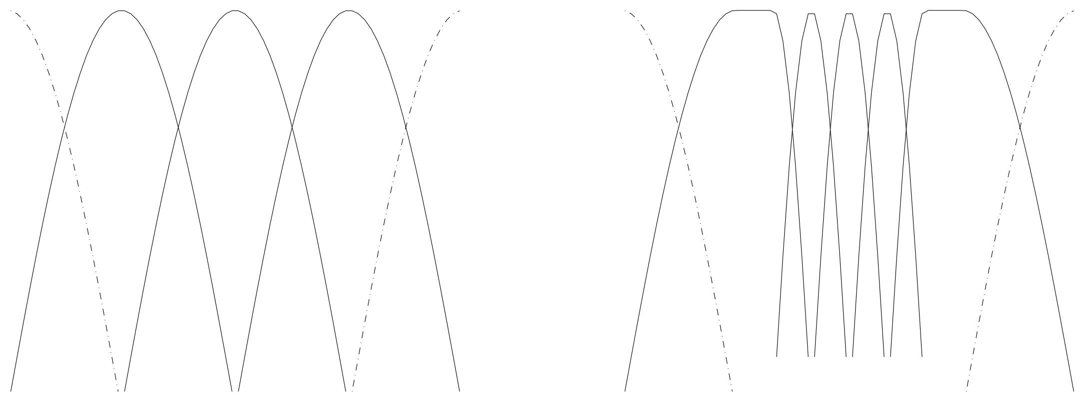

Figure 5.12: Adaptive segmentation within the Layer III hybrid filter bank: monotonic segmentation (on the left) using 36-point "long" windows, dynamic segmentation (on the right) using "long" and 12-point "short" windows, as well as two special "start" and "stop" windows necessary to comply with transition requirements.

windows is implicit.

If the signal is stationary, Layer III codes two granules, each one comprising 576 frequency coefficients, using non-uniform (exponential) quantization and Huffman coding. It is important to note that the non-uniform quantization represents, by itself, a specific shaping of the quantization noise *within* each coded spectral band (quantization partition), according to the magnitudes of the spectral components. In this sense, it may be seen as a companding technique applied in the frequency domain.

While Layer I or Layer II code audio signals at constant bit rate, Layer III codes audio signals at variable bit rate. Thus, in order to code at constant bit rate, Layer III uses an iterative procedure to control/adjust the rate/distortion of the coding process[21].

When coding both channels of a stereo pair, any Layer of the MPEG-Audio standard may use one out of two techniques that exploit inter-channel redundancies and irrelevancies: Mean-Side Stereo (MS) or Intensity Stereo (IS). MS denotes the coding of the sum and difference of the Right and Left audio channels. The underlying rationale is that the MS representation basis is more uncorrelated than the basis comprising the Left and Right channels. In addition, the MS basis allows a perceptually more efficient spatial distribution of the quantization noise relative to the virtual image of the stereophonic audio signal.

IS is mainly used to control the bit rate within Layer I or Layer II. IS forces the coding of both Left and Right channels, at frequencies above 2 kHz, on the basis of a single signal and a coefficient characterizing the difference in intensity between the two channels. The underlying rationale is that above 2 kHz, the spatial information lies *only* on the intensity difference between the energy envelope of the two channels, and not on their relative fine temporal structure. We do not share entirely this point of view since the IS destroys the phase information of the stereo signal, which may still be important for the localization of frequency components above 2 kHz and well as for very short (*i.e.* impulsive like) signals. This point of view is shared by Johnston [Johns.96, page 3] '*IS coding is not suited when signals have a time delay component*'. The importance of the spatial information is emphasized by Noll [Noll97, page 60]: '*In the case of stereophonic and multichannel signals, spatial integrity is an additional dimension of quality*'.

5.5.2.2 The MPEG-2 Audio Standards

MPEG-2 Audio comprises two multichannel coding standards: the MPEG-2 Audio BC and the MPEG-2 Audio AAC. The former is backwards compatible (BC) with the MPEG-1 Audio[22], and the latter is not.

The MPEG-2 Audio BC The MPEG-2 BC Audio coding standard [ISO94] consists in an extension of the capabilities of the MPEG-1 Audio standard by allowing the coding of additional

[21] This is implemented using an *approximate* analysis-by-synthesis control of the quantization error since time-domain aliasing effects are not taken into account.

[22] This means that an MPEG-1 stereo or mono decoder can deliver a correct (down-mix) audio signal when decoding an MPEG-2 multichannel bit stream [Noll95].

audio channels, and by admitting lower sampling frequencies: 16 *kHz*, 22.05 *kHz* and 24 *kHz*. A complete description of the MPEG-2 BC Audio algorithm can be found in [Noll95].

Backwards compatibility implies the use of "matrixing" so as to obtain a compatible down-mix of sterophonic signals. When dematrixing the coded down-mix, signal artifacts may appear as a result of unmasked quantization noise [Noll95]. This limitation lead to the development of a new multichannel MPEG-2 Non-Backwards Compatible (NBC) coder that was completed and approved by the ISO in April 1997, under the designation of ISO/IEC 13818-7 MPEG-2 Advanced Audio Coder [Noll97].

The MPEG-2 Audio AAC The AAC algorithm started to be developed in 1994 [Bosi96], having at its origin two major contributions: the multichannel AC-3 coding algorithm from Dolby [Davis93], and the multichannel PAC coding algorithm from AT&T [Johns.92]. Later, other companies joined the collaborative effort bringing other technological contributions [Bosi96].

The structure of the AAC algorithm can be expressed as a collection of self-contained processing modules called *tools*. This term is now commonly used within the MPEG-4 standardization activities. The AAC tools are arranged to build an AAC encoder as represented in Figure 5.13. In the following we describe briefly each tool.

Pre-processing Consists in a 4-band polyphase quadrature filter bank (Pseudo-QMF), gain detectors and gain modifiers. The gain in each filter output is independently controlled by the gain detectors and modifiers in order to reduce pre-echo effects.

Filter Bank Implements a time-frequency transformation by means of an MDCT (TDAC filter bank), using 50% overlap. The length of the MDCT is 2048 samples or 256 samples in case the pre-processing tool is used (in this case one 256-point MDCT is applied at each subband of the 4-band Pseudo-QMF). Dynamic window switching is also allowed in order to minimize pre-echo effects. The length of the "short" window is 1/8 the length of the "long" window. The shape of the analysis/synthesis time window can also be changed in order to make it possible to benefit from a better far stop-band rejection of the transfer function of each subband channel[23] (at a price of a wider pass-band and worse near stop-band rejection) [Bosi96].

[23] The new window proposed by Dolby (Kaiser-Bessel Derived -KBD) did not have other design concerns than far stop-band attenuation. Therefore, it can not be compared with the results of our optimal window design (Section 4.5).

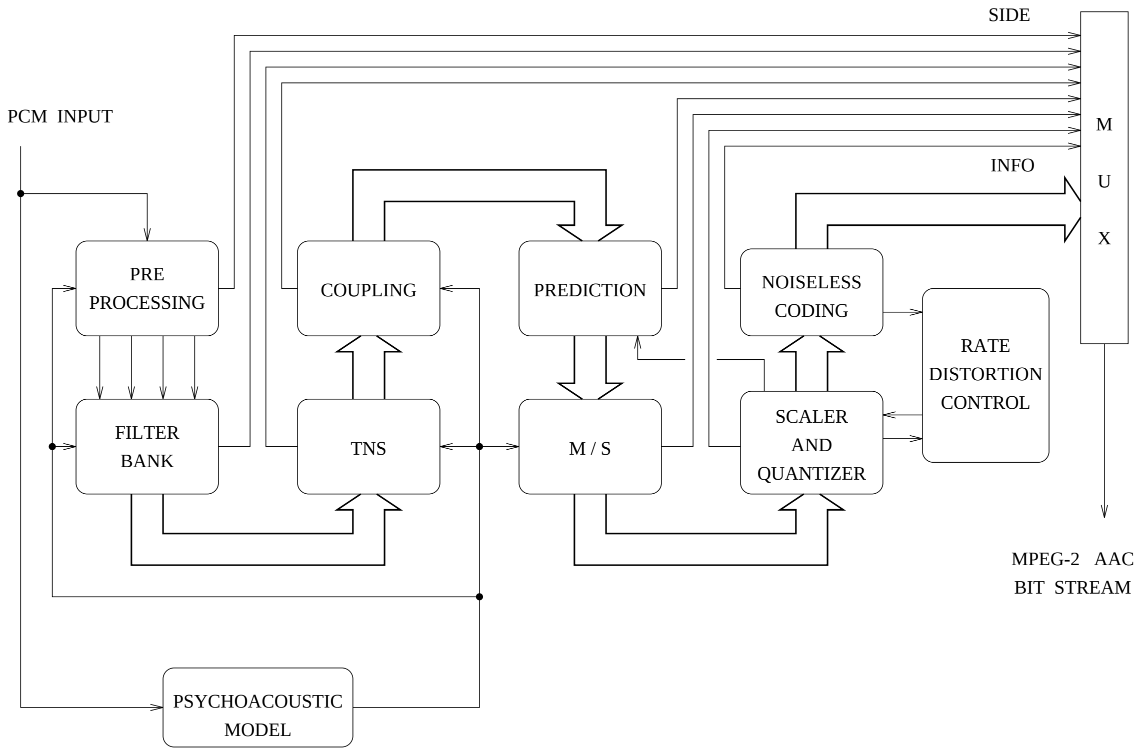

Figure 5.13: Structure of the MPEG-2 Advanced Audio Coder (AAC).

Temporal Noise Shaping (TNS) Consists in a predictive coding technique (Section 5.3.2) applied to spectral coefficients [Herre96]. As the dual of time-domain predictive coding techniques (that model the envelope of the Power Spectral Density of the signal), this tool attempts to model the envelope of the time signal over the lenght of a time window, providing a means to modulate the quantization noise, in time, with a similar shape. Thus, this tool represents an alternative way to control pre-echo effects. Up to three all-pole filters may be applied in each audio frame. The highest permitted order of any filter is 20. Although coding gains have been reported [Herre96] for non-stationary and "pitch-based" signals, the implications of applying DPCM to spectral data obtained from a TDAC filter bank, have not been thoroughly addressed.

Coupling Consists in the application of the Intensity Stereo Technique to a selection of channel pairs, in order to reduce bit rate or in order to allow the down-mix (replication) of a sound object (*e.g.* a commentary channel), into the stereo imaging of the multichannel program [Johns.96].

Prediction This tool consists of several second-order backwards-adaptive lattice predictors. Each predictor is applied to a group of subsequent spectral components in a given subband. The objective is to obtain a coding (prediction) gain in stationary parts of the audio signal (*i.e.* prediction is only used for "long" windows). In order to define entry points in the bit stream (and also in order to increase stability), a cyclic reset mechanism is applied in the encoder and decoder [Bosi96].

Mean-Side Stereo (M/S) Codes pairs of channels that are arranged symmetrically on the left/right listener axis, using the M/S technique. M/S can be applied in time and frequency, in a flexible manner [Johns.92].

Scaler and Quantizer This tool quantizes the spectral data according to the noise-shaping requirements estimated by the psychoacoustic model, and according to the data rate limitations. An approximate analysis-by-synthesis procedure is used that is similar to that existing in the MPEG-1 Audio Layer III algorithm.

Noiseless Coding This toll comprises a collection of Huffman CodeBooks that are used to code *sections* of quantized spectral coefficients (1024 total per transform), a specific CodeBook being assigned to each individual section. Sectioning is performed using a *greedy merging algorithm* [Bosi96] that the author of this Dissertation developed for the AT&T PAC coder [Johns.92], and that is also used in the INESC ASC coder (described in the next Chapter). Quantization step-sizes are normalized by a global gain, are quantized in steps of 1.5 dB, are differentially encoded (relative to the previous scale factor), and finally are Huffman encoded using a dedicated Huffman CodeBook.

Multiplexer Assembles all coded data and delivers a flexible bit stream syntax including transport information (*e.g.* synchronization).

The psychoacoustic model of the AAC algorithm is taken from the MPEG-1 Audio standard (Psychoacoustic Model 2) [Noll97].

In order to allow a tradeoff between quality and complexity, AAC admits several configurations, each configuration comprising a specific collection of tools. The standard specifies three configurations (or *profiles*): Main Profile, Low Complexity Profile, and Sample Rate Scalable Profile [Bosi96][Noll97].

Subjective tests have revealed that, in average, the 5-channel AAC coding at 320 $Kbit/s$

scores better than the 5-channel MPEG-2 Layer II coding at 640 $Kbit/s$ [Bosi96].

As it can be concluded from Figure 5.13, most tools imply the transmission of some side information to the decoder. According to the profile, some tools may also require additional side information to inform the decoder about their activation or not (*e.g.* TNS). The descriptions of the AAC algorithm we had access to, do not make clear how the efficiency of a specific tool depends on the right activation decisions, nor how its influence might be rather negative due to a wrong activation decision. A global appreciation of the algorithm is therefore very difficult because there is, in addition, some functional overlap among different tools. For example, Pre-processing, TNS and the Window Switching mechanism, all are meant to solve the pre-echo coding problem, but it is not too clear how to chose a specific one, nor what the contribution of each one is if all are used at the same time.

The MPEG-2 AAC algorithm will constitute the kernel of the forthcoming MPEG-4 multimedia standard, for the range of bit rates above about 32 $Kbit/s$ per channel.

5.5.2.3 Proprietary Audio Coding Schemes

In this sub-section we will describe several proprietary audio coding algorithms we believe are important references to explain the current state-of-the-art, as well as recent standards. Specifically, we will briefly describe AC-2 and AC-3 (Dolby), PAC (AT&T), ATRAC (Sony), OCF (FhG), ATC-ABS (NEC) and Twin-VQ (NTT).

AC-2 The Dolby AC-2 coding algorithm is based on a evenly-stacked TDAC filter bank [Princen86], consisting of alternating MDCTs and Modified Discrete Sine Transforms (MDST) whose size is 512 samples (for moderate coding delay) or 128 samples (for low delay coding). The log-spectral envelope of each subband (grouping a number of adjacent coefficients according to the Bark rate) is used to derive the dynamic (signal adaptive) component of the global bit allocation. This processing is performed by a simplified bit allocation routine that is implemented in both encoder and decoder. The dynamic bit allocation component represents 20% of the total bit allocation. The remaining 80% is fixed and determined by conservative masking rules. AC-2 achieves 4:1 to 6:1 compression with a moderate ($\approx 40\ ms$) to low delay coding ($\approx 8\ ms$).

While AC-2 focuses on low complexity and low delay coding of two independent audio signals, AC-3 addresses the efficient coding of a multichannel combination of audio channels[24]. AC-3 does not use channel matrixing to avoid unmasked noise, but uses the Intensity Stereo technique,

[24] Davis [Davis93] indicates that '*in general, the average bit rate demand of multiple channels appears to be roughly proportional to the square root of the number of channels*'.

the Mean-Side Stereo technique, and all channels are allowed to share a common bit pool.

The basic structure of the AC-3 encoder comprises: time-frequency transformation by means of a TDAC filter bank, transient detection by means of a high frequency band-pass filter (which activates a window switching mechanism), conversion of spectral coefficients to a floating point format, bit allocation computation from the TDAC exponent information, dithering, and scalar quantization. Quantized coefficients are combined into triplets or into pairs which are represented by a single code-word. Unlike the AC-2 coder, AC-3 uses only a small fixed bit allocation component. Some intermediate information of the bit allocation computation is directly transmitted to the decoder. This reduces the complexity of the decoder, *i.e.* increases the encoder-decoder asymmetry which is a desirable feature in many applications. The coding delay of AC-3 is in the order of 100 ms.

One of the merits of the AC-3 coder is the underlying concern of a practical implementation and operation. For example, the possibility is provided for the decoder to follow explicit bit allocation indications conveyed by the encoder, instead of using the locally derived bit allocation. As another example, the coded data format allows editing at the transform block level, despite the fact that both AC-2 and AC-3 assemble six transform blocks into a transmitting frame.

A rather interesting technological aspect is that the AC-3 bit stream as recorded in film soundtracks, is placed in data blocks between the perforations along one edge of the film [Davis93].

AC-3 has been selected as the audio coding sub-system for the US Grand Alliance High Definition Television System (the new television transmission standard of the United States) [Noll95].

PAC The AT&T PAC coder[25] [Johns.92] is an improved version of the PXFM coder [Johns.88b], shares a few technical approaches with ASPEC [Brand.91], and exploits the concept of Mean-Side Stereo coding introduced by Johnston [Johns.89]. PAC uses an MDCT filter bank, a (1024/256-point) window switching mechanism to control pre-echo effects, signal adaptive Left/Right (L/R) or M/S coding (in time or frequency), scalar quantization, and Huffman coding of quantized coefficients and most side information.

The window switching mechanism is entirely driven by the psychoacoustic model using Perceptual Entropy estimates for "long" and "short" transform blocks of audio data [Ferre.92]. The L/R or M/S coding modes are specified on a quantization-partition by quantization-partition

[25]The description given here is only based on Johnston and Ferreira [Johns.92], and therefore does not include other more recent versions of the algorithm, namely the extension to multichannel PAC (MPAC).

basis, using thresholds based on Binaural Masking Level Differences (MLD, Section 3.4.3.3) to prevent stereo unmasking of the quantization noise.

Rate control is achieved by increasing the global Threshold of Masking in case of undercoding, or by interpolating between the Threshold of Masking and the Absolute Threshold of Hearing, in the case of overcoding.

Huffman coding involves spectral partitioning and assignment of CodeBooks. The technique is also used in ASC and has been adopted by the MPEG-2 AAC standard.

Subjective tests revealed that for a total bit rate of 128 $Kbit/s$ (two audio channels) the quality advantage by using stereo coding over dual mono coding is about 0.6 on the 5-point impairment scale [Johns.92]. The tests also revealed that stereo coding at 96 $Kbit/s$ is preferable to dual mono coding at 128 $Kbit/s$ total.

ATRAC This algorithm is used in the MiniDisc system [Yoshida94]. ATRAC uses a hybrid filter bank, window switching, adaptive bit allocation and scalar quantization. The hybrid filter bank includes a cascade of two QMFs that divides the signal spectrum into three bands: $0 - 5.5$ kHz, $5.5 - 11.0$ kHz, and $11.00 - 22.0$ kHz. An MDCT is then applied to each subband. The length of the MDCT (128-point or 32-point), is adaptively selected in each subband (window switching), in an independent way, to control pre-echo effects.

Spectral coefficients are grouped into Block Floating Units (BFU), in a non-uniform spectral organization approximating the *Bark* rate. Grouped spectral coefficients are used to derive bit allocation information and subsequently are quantized using a companding-like approach in the frequency domain (similarly to that used in AC-2). All coefficients within a BFU share the same scale-factor and are quantized to the same number of bits.

Bit allocation is derived as in AC-2 but is implemented differently. In fact, the weight of the fixed bit allocation component and the complementary weight of the dynamic bit allocation component, depend on a measure of the local tonality of the audio signal. Under/over coding is controlled by lowering/raising the global bit allocation by the same number of bits (*i.e.* by raising/lowering the Threshold of Masking by the same offset on a logarithimc scale). Unlike in AC-2, in ATRAC all bit allocation information is parametrized and conveyed to the decoder.

Although very simple in its structure, which is a strong requirement for portable consumer products, ATRAC achieves a very good subjective quality at a compression ratio of 5:1.

OCF Optimum Coding in the Frequency domain (OCF) is a very important reference coder because it had a clear impact on more recent coders such as ASPEC, MPEG-1 Audio Layer

III, and MPEG-2 AAC. In fact, some of the building blocks of these coders were taken (with adaptations) from OCF. For example, the rate distortion control, the non-uniform quantization and the entropy coding building blocks of Layer III were "borrowed" from OCF. To illustrate this, we will describe briefly a few important processing functions existing in OCF, as taken from two of the earliest OCF publications [Brand.87][Brand.88].

Transformation Time-frequency transformation is achieved by means of a DCT [Brand.87], or by means of an MDCT [Brand.88].

Rate Control Using the fact that according to optimum ATC coding (Section 5.3.2) the quantization step-size for all coefficients is identical (in the case of scalar uniform quantization), OCF uses a single step-size in the quantization of all coefficients. These are Huffman encoded and the total bit count is adjusted to the desired data rate by iterating the procedure with a coarser of finer step-size. This consists in the so-called inner iteration loop. The spectral flatness measure is used to set the initial step-size.

Distortion Control Using a simple psychoacoustic rule indicating that in order to be masked, '*the noise has to be 6 dB below the wanted signal*' [Brand.87, page 143], or using a more complex Psychoacoustic Model [Brand.88]; an outer iteration loop monitors the quantization process (by means of an approximate analysis-by-synthesis procedure), to detect an excessive quantization noise in at least one critical band. In this case, the quantization process is perturbed by lowering the quantization step-size in that particular band or, equivalently, by amplifying the spectral coefficients in that band. The objective is to force the quantization noise below the allowed limit that corresponds to the Threshold of Masking. After this, the quantization and entropy coding processes start over again. For a given bit rate, the whole process stops when noise is masked for all bands, or when the overall Noise-to-Mask-Ratio (NMR) is minimized [Brand.88].

Regarding quantization, a non-uniform (exponential) quantizer is used '*to get some natural way of noise shaping*' [Brand.88, page 143].

Regarding the coding of quantized coefficients, OCF *forces* a organization of spectral coefficients as follows: the region with "zeroes" in the highest frequency region is run-length encoded. The next spectral region towards low frequencies containing quantized spectral coefficients ranging from -1 to +1, is coded with short codes. The remaining quantized spectral coefficients are Huffman encoded.

Coding results of very high quality were reported to be achieved with 3 bit/sample (*i.e.* about 5:1 compression) [Brand.87], and later with 1.5 bit/sample [Brand.88].

Since we could not find any published conclusions concerning the interaction between the inner and the outer loop, we are convinced these nested loops may interact in such a way as to make the convergence process very slow and possibly far from smooth. The OCF modifications reported by Brandenburg [Brand.88] and addressing the reduction of the number of inner and outer iterations (by means of a coarser increasing step-size in the inner loop and "proper" selection of the allowed distortion in the outer loop), somehow denote that this is a real concern in OCF, and therefore corroborate our hypothesis.

ATC-ABS This coder is a very important reference regarding dynamic window switching in ATC based perceptual audio coders. Sugiyama *et al.* proposed to adaptively change the size of the time-frequency transformation so as to avoid coding artifacts in non-stationary regions of the audio signal, referred to as pre-echoes [Sugiya.90].

The ATC-ABS encoder is based on the Zelinsky and Noll ATC coder [Zelins.77] and uses a local decoder in order to compare the SNR resulting from the difference between original and reconstructed signal, for a selection of block sizes. Only a limited number of candidate block sizes are allowed that define several possible segmentations. The segmentation delivering the highest overall SNR is chosen for coding.

Besides the above "feedback" approach, Sugiyama [Sugiya.90] and Iwadare [Iwadar.92] propose also a "feed-forward" approach that does not require a local decoder in the encoder. This approach is based on inter-block differences of a measure based on the variance of normalized time-samples. While leading to a slightly lower coding performance than the "feedback" proposal, the advantage over fixed block coding is clear which explains that the idea has been retained and is used today in most perceptual audio coders.

Twin-VQ This recently developed coder, targets high quality audio coding at less than 64 $Kbit/s$, and has been one of the successful candidates to the MPEG-4 Audio competitive tests. The coder has been substantially modified since its first publication [Iwakam.95][Moriya97]. For example, a window switching mechanism has been recently added to the algorithm. These algorithmic improvements are also a consequence of a strong involvement of Twin-VQ in the current MPEG-4 standardization activities.

The basic structure of Twin-VQ comprises: MDCT transformation, coefficient flattening (in the frequency domain), power spectral normalization, and weighted interleave vector quantization of the residual MDCT coefficients. The first stage of spectral flattening is achieved by actually dividing the MDCT coefficients, in the frequency domain, by the spectral envelope of the signal as estimated by means of an LPC filter. This technique obviously reminds the DYNAMIC PREEMPHASIS technique developed by Tribolet and Crochiere (Section 5.3.2.4).

The second stage of flattening is achieved by dividing the spectral residual of the first stage by a backwards predicted fine-structure spectral envelope (*i.e.* a pitch like spectral profile). The residual coefficients are then interleaved and are weighted vector quantized with a two-channel conjugate structure VQ (hence Twin-VQ). The weighted distortion measure is derived from the envelope of the signal spectrum.

The subjective quality of Twin-VQ encoded signals at rates below 64 $Kbit/s$ have been reported to be superior to that of MPEG-1 Layer II coded signals at the same bit rate [Iwakam.95]. More recent results report [Moriya97] on 8 $Kbit/s$ and 16 $Kbit/s$ coding of audio.

The coding efficiency of the algorithm depends essentially on how much spectral flattening can be achieved since this maximizes the gain of the weighted interleave vector quantization. In fact, in its original version, Twin-VQ does not use neither a psychoacoustic model nor adaptive bit allocation. In this sense, the algorithm does not fit too well in the category of perceptual audio coders.

5.6 Summary

This Chapter has reviewed speech, wideband speech, and high quality audio coding techniques.

The emphasis has been placed on the evolution of perceptual coding paradigm over time, and on its actual implementation by coding schemes that represent the current state-of-the-art.

Proprietary and standardized coding algorithms for high quality audio have been described. This not only gives a snapshot of the international research activity in this area, but also facilitates a relative evaluation of the ASC algorithm that is described in the next Chapter.

Chapter 6

ASC - Audio Spectral Coder

Failure is not an option.

Apollo XIII, Universal Studios, 1995.

6.1 Introduction

In previous Chapters we have discussed the concepts, the methods, and the techniques we believe are extremely relevant in the area of high quality audio coding.

Regarding specific coding issues such as psychoacoustic modeling (Chapter 3) and time/frequency transformation for analysis and coding (Chapter 4), we presented our perspective on specific topics and proposed our own solutions as a result of intense research.

In this Chapter, we describe how we have substantiated those discussions and solutions in a perceptual audio coder that we have developed from 1993 till 1995. We decided to name it Audio Spectral Coder (ASC). In its basic structure, ASC is an enhanced Adaptive Transform Coder that allows the coding of very distinct audio material, from basic speech to high quality (monophonic) audio. As we will detail, ASC allows a flexible choice of coding conditions and parameters.

The benefits of the interactive simulation and evaluation environment that assisted the development of ASC, as described in Chapter 2, will be implicit throughout this Chapter. In order to constrain this Chapter to a reasonable size and in order to not distract from the main subject in focus, we just mention here a few topics showing how that interactive environment was of paramount importance to identify a number of critical aspects that influence the global coding performance of ASC. For example, the problem of extensive spectral zero bit allocation over consecutive encoded audio frames, which gives rise to multiplicative aliasing effects in addition to additive aliasing effects, has been easily identified and studied as reported in Section 4.5, and elsewhere [Ferre.96b]. Other critical factors were also found that will not be detailed but were carefully addressed in our research :

- the suitable spectral resolution (relative to the Bark rate) of the encoding process since a very fine (sub-Bark) resolution can expose too much the inaccuracies of the psychoacoustical model and imply a significant increase in side information, whereas a poor resolution can lead to a sub-optimal bit allocation but allow a better tradeoff between side information (mainly quantization step-sizes) and main information (quantized spectral coefficients),
- the criterion to raise the Just Noticeable Distortion (JND) noise profile [Jayant93] for a "graceful" quality degradation of the undercoded audio signal (*i.e.* coded with less bits than those required for transparent coding),
- the adjustment, correction or compensation of the psychoacoustic model considering the idiosyncrasies of the analysis/synthesis filter bank,
- the existence of other concurrent factors besides extensive spectral zero bit allocation that expose the "aliasing signature" of the analysis/synthesis filter bank,
- the update of Huffman tables according to the bit rate, and the associated criterion to collect statistics which depends on the overcoding or undercoding expectation,
- the type of quantizer (*e.g.* linear, logarithmic) used to quantize spectral coefficients and quantization step-sizes, as well as the width of the "zero" region of the quantizer,
- the use of an output buffer to accommodate different bit rate requests among adjacent encoded audio frames,
- the judicious exploitation of inter-frame correlations (*e.g.* correlations existing between adjacent short windows that are coded when the time resolution of the encoding process is improved),
- the characterization and efficient encoding of all side information.

The remaining of this Chapter is organized as follows: In Section 6.2 we describe the main building blocks of the ASC encoder. Because the decoder reflects the structure of the encoder to a large extent, the ASC decoder is only briefly described in Section 6.3. Section 6.4 illustrates the ASC performance taking the MPEG-1 Audio standard as a reference. Several aspects of the participation of ASC in the MPEG-4 Audio standardization activities, namely the results of the competitive evaluation tests, are presented in Section 6.5. Section 6.6 concludes this Chapter.

6.2 The Encoder

6.2.1 General Structure

A simplified diagram of the ASC encoder structure is shown in Figure 6.1. Besides including the main building blocks of a typical perceptual audio coder, namely the time-frequency transformation, the psychoacoustic model, the quantization, and the entropy coding, other building blocks, pertaining typically to source coders, are also included. In fact, ASC analyzes the harmonic structure of the audio signal (only in stationary regions of the signal), selects the most relevant "pitch" and harmonic pattern, and uses this information as a weighting factor in the quantization operation. The theory underlying this approach has been presented in Chapter 3, Section 3.3.6. The practical motivation will be presented in Section 6.2.4.

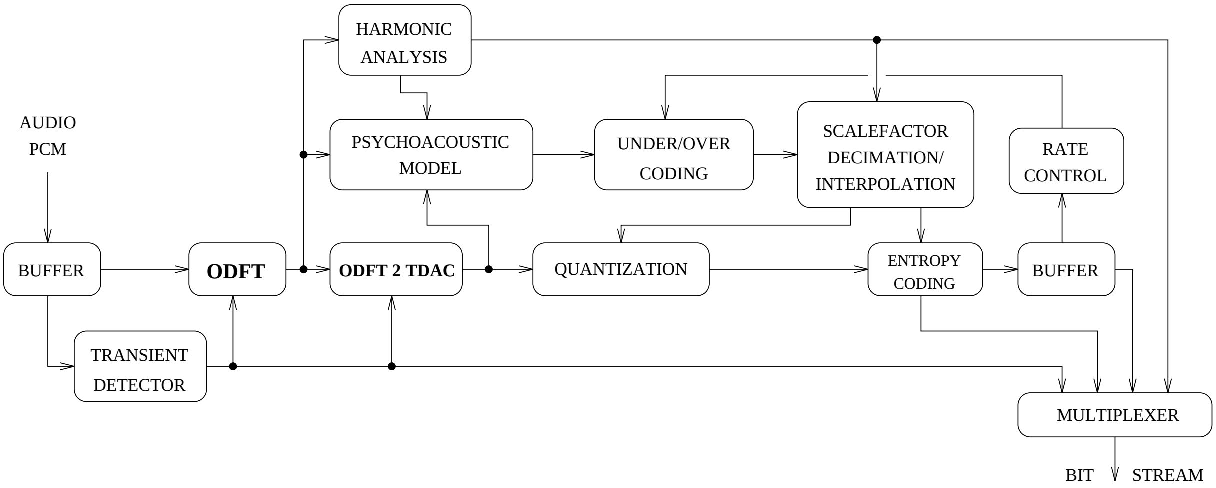

Figure 6.1: A simplified diagram of the ASC encoder.

A particularly relevant aspect which permits a reduction of the complexity of the encoding algorithm, as will be shown later, is that the most significant effort involved in the computation of the spectral coefficients for quantization can be reused to derive important information for the psychoacoustic model, as well as for the harmonic analysis. Therefore, contrary to what is recommended in the MPEG-Audio standard, a parallel DFT is avoided, as well as the associated overhead in computation.

ASC was designed so as to allow a flexible selection of the most important coding parameters: the sampling frequency, the length of the basic audio frame, one out of three possible improved time resolution coding modes, the selectivity of the time analysis/synthesis filter, the average bit rate, and the bit rate coding mode (constant bit rate (CBR) or variable bit rate (VBR)). In the following we describe further the associated possibilities.

- Any sampling frequency is accepted for simulation purposes although only a limited set has representation in the output bit stream. These sampling frequencies range from 8 kHz till 48 kHz and include the main sub-multiples of the professional and industry standards.

- The average bit rate is lower limited to 16 $Kbit/s$ and upper limited to 128 $Kbit/s$. The ASC bit stream presumes the constant bit rate coding mode and allows the specification of one out of 16 possible coding bit rates.

- The length of the audio frame, which defines the overall coding delay, may vary from 32 samples till 1024 samples, in steps that correspond to power of two numbers. The length of the time window filter that precedes the time/frequency transformation is two times the frame length. The overall system delay including all input and output buffers is four times the frame length.

- In order to deliver the best subjective coding quality according to local (non-)stationarities of the audio signal, a window switching scheme is provided that improves the time resolution of the encoding process in order to comply with masking requirements in the time domain (Section 3.3.2). The degree of the time resolution improvement may be selected from several possibilities, as will be illustrated later.

- The spectral resolution and spectral selectivity are important aspects that determine the coding performance [Ferre.96c]. The former depends mainly on the length of the time analysis window and the latter is mainly a function of the shape of this window. ASC allows the selection of a specific window shape satisfying the perfect reconstruction property [Malvar92][Ferre.96b], thus permitting an easy evaluation of its impact on the coding performance.

In the following, we will describe the main processing functions of the ASC encoder. Each processing function may involve more than one of the building blocks shown in Figure 6.1.

6.2.2 Time Resolution Coding Mode

The improved time resolution coding mode is activated when strong local non-stationarities in the audio signal are detected. This is needed for two reasons. First, as ASC is a transform based coder, its coding gains are effective only with quasi-stationary signals (Section 4.3.1.6). Second, as the spread of the quantization error is weighted by the time synthesis window across its entire length, it may happen that the level of the coding noise exceeds that of the audio signal for a sufficient period of time such that it is not masked by the signal. In other words, pre-masking or post-masking may not be successfully exploited. A practical and efficient way to overcome this problem is to permit dynamic segmentation, *i.e.* to code the audio signal with a shorter time analysis/synthesis window, when required, such that the spread of quantization noise in time is within the limits tolerated by time masking requirements. As we argue in Section 4.3.4, this window switching occurs only when non-stationary conditions are detected.

ASC permits several possibilities for the dynamic segmentation as illustrated in Figure 6.2. The three switching possibilities correspond to an improvement of the time resolution by a factor

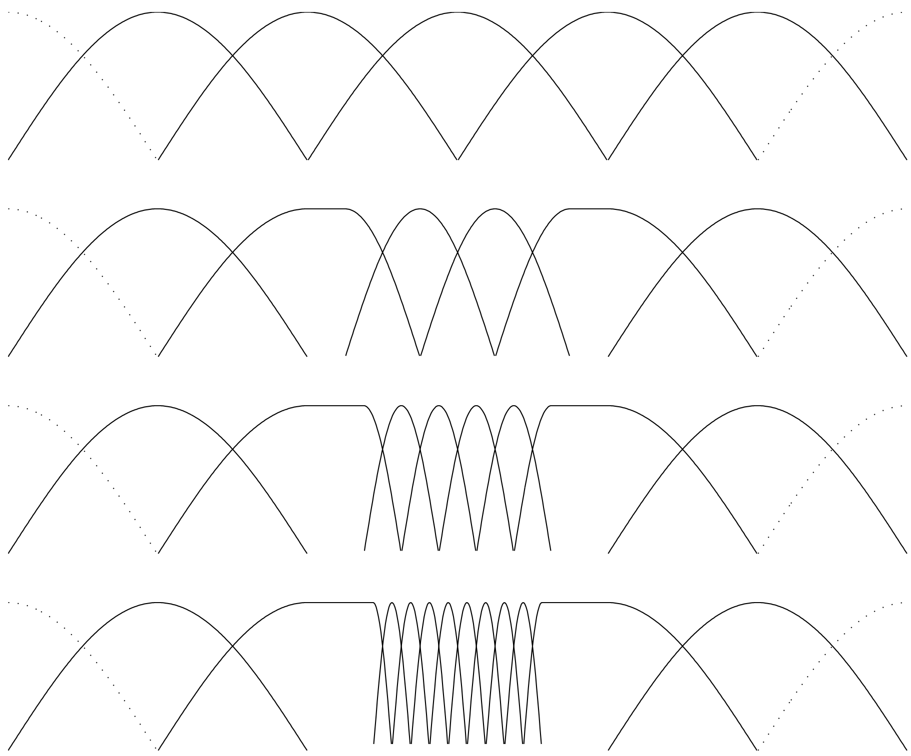

Figure 6.2: Possible window switching modes allowing an improvement of the coding resolution in time by factors 2, 4, and 8 (from top to bottom).

of 2, 4, and 8. Although only one frame is illustrated in the improved time resolution coding mode, several adjacent frames might use the same fine time resolution if coding conditions determine so. As illustrated, special windows (satisfying the perfect reconstruction property) are used to perform the transition between coding modes. This requirement is similar to that illustrated in Section 5.5.2.1 for the MPEG-1 Audio Layer 3 coder.

ASC allows the user to specify the switching factor of the dynamic segmentation for each "run" of the encoder. This is important in order to take into account the selected frame size and sampling frequency, and to make it possible to study time masking effects. The local signal non-stationarities are monitored by means of a transient detector. Its current implementation is based on a 5^{th} order all-pole linear predictor whose spectral envelope is illustrated in Figure 6.3.

This model was derived from quasi-stationary regions of typical high quality audio signals. As

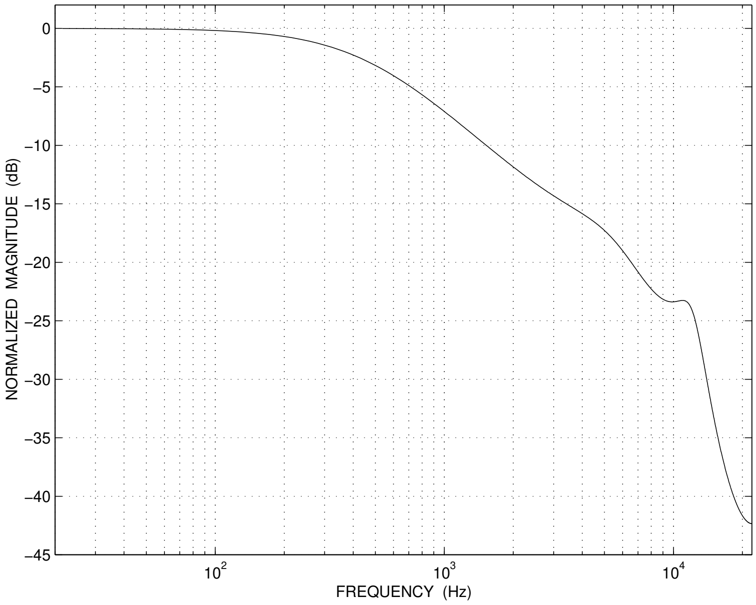

Figure 6.3: Spectral envelope of the LPC predictor used to detect signal transients.

the prediction residuals increase significantly in non-stationary regions of the audio signal, the energy of the prediction residuals is used to trigger the improved time resolution coding mode. A look-ahead delay of one frame (half the analysis window length) is necessary to permit the insertion of a "start" transition window.

6.2.3 Time-Frequency Transformation

As mentioned previously in Section 4.3.1.5, Section 4.4.2, and in Section 4.5.2, ASC uses a signal analysis/synthesis scheme based on the TDAC filter bank. Three main operations can be identified that correspond to the audio signal analysis, synthesis and reconstruction. First, the audio signal is analyzed into frames in the encoder according to an analysis window $h(n)$:

$$X_{TDAC}(k,m) = \sum_{n=0}^{N-1} h(n)x(n + m\frac{N}{2}) \cos\left[\frac{2\pi}{N}\left(k+\frac{1}{2}\right)\left(n+\frac{1}{2}+\frac{N}{4}\right)\right]. \tag{6.1}$$

The length of the analysis window is N and m is the time index that sets the 50% overlap between adjacent blocks of audio data as illustrated in Figure 6.2. As the input signal is real, only $N/2$ coefficients, as given by $X_{TDAC}(k,m)$ for the same time index m, are *unique* in each frame [Ferre.92]. This means that the system is critically sampled. Therefore, it is only necessary to retain and code spectral coefficients whose index is in the range $0 \leq k \leq \frac{N}{2} - 1$.

Assuming absence of quantization, the signal is synthesized in the decoder as:

$$\hat{x}(n,m) = \frac{1}{N}\sum_{k=0}^{N-1} X_{TDAC}(k,m)\cos\left[\frac{2\pi}{N}\left(k+\frac{1}{2}\right)\left(n+\frac{1}{2}+\frac{N}{4}\right)\right]. \tag{6.2}$$

Finally, the audio signal is reconstructed using a synthesis window $h(n)$ in the 50% overlap-add scheme:

$$\tilde{x}(n) = \sum_{m=-\infty}^{\infty} h(n - m\frac{N}{2})\hat{x}(n - m\frac{N}{2}, m). \tag{6.3}$$

ASC uses for both the analysis and synthesis windows the optimal window described in Section 4.5.6 and illustrated in Figure 4.17. In the absence of spectral modification, we have $\tilde{x}(n) = x(n-d)$, where the system delay is $d = N-1$, as we have concluded in Section 4.3.1.4. In the following the time block index m will be omitted for clarity.

ASC codes the audio signal by quantizing spectral coefficients at the output of a TDAC filter bank. As follows from Equation (4.16), it can be shown [Ferre.92] that the TDAC coefficients may also be computed as follows:

$$\begin{aligned} X_{TDAC}(k) = {} & \Re\{X_{ODFT}(k)\}\cos\left[\frac{2\pi}{N}\left(k+\frac{1}{2}\right)\left(\frac{1}{2}+\frac{N}{4}\right)\right] \\ & + \Im\{X_{ODFT}(k)\}\sin\left[\frac{2\pi}{N}\left(k+\frac{1}{2}\right)\left(\frac{1}{2}+\frac{N}{4}\right)\right], \end{aligned} \tag{6.4}$$

where $X_{ODFT}(k)$ represents the coefficients of the Odd-DFT. The Odd-DFT is computed using Equation (4.42) which we repeat here for convenience:

$$X_{ODFT}(k) = \sum_{n=0}^{N-1} h(n)x(n)e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n}.$$

The advantage of the Odd-DFT is that it is a complex transform whose sampled frequencies correspond to those of the TDAC filter bank. As we will show in Chapter 7, the Odd-DFT is also very interesting in post-processing.

The Odd-DFT is simply derived from the DFT by rotating the sampled frequencies of the latter by π/N, where N is the size of the transform [Bellan.89]. A consequence of this rotation, as illustrated in Figure 6.4, is that while the DFT of a real signal has $\frac{N}{2}+1$ unique spectral coefficients in the range $[0, \pi]$, the Odd-DFT has only $\frac{N}{2}$. As mentioned in Section 4.3.1.4, the computational cost of the Odd-DFT can be made equal to the computational cost of the DFT if the *decimation-in-time* method [Oppenh.75] is used for the fast computation of the direct transform, and if the *decimation-in-frequency* method [Oppenh.75] is used for fast computation of the inverse transform.

ASC uses the result of Equation (6.4) and, as illustrated in Figure 6.1, computes first the complex coefficients of the Odd-DFT. These coefficients contain important information (namely

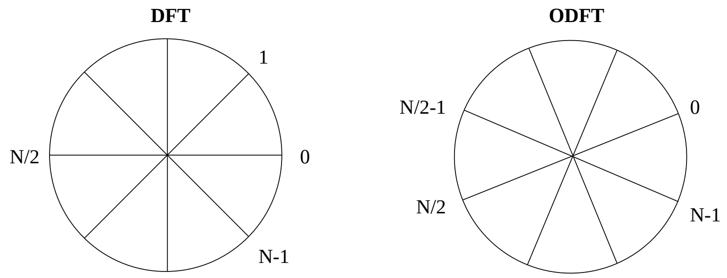

Figure 6.4: Difference between the sampled frequencies of the DFT and those of the Odd-DFT. In this Figure, N=8 and reflects one period (2π) of the digital frequency.

phase information) that is necessary for harmonic analysis [Ferre.96c] and for the estimation of the JND profile. A final processing stage, according to Equation (6.4), delivers the MDCT coefficients. Therefore, important psychoacoustic information is made available as a result of an extremely simple reformulation of the analysis filter bank. This represents a major contribution to the low computational complexity of ASC.

6.2.4 Harmonic Analysis

In this section we address harmonic analysis. The practical motivation is presented in Sub-section 6.2.4.1. A possible harmonic identification algorithm that takes advantage of the ASC analysis filter bank, is described in Sub-section 6.2.4.2.

The subject of harmonic coding and quantization is addressed in later sections. Specifically, we describe in Section 6.2.5 possible criteria to perform a global quantization and coding of all (including unresolved) harmonics. In Section 6.4 we compare the coding results obtained when accounting or not for the harmonic structure. As it will be shown, the quality improvement in the coding of strongly harmonic signals support our hypothesis that a careful modulation of the quantization noise in critical bands, taking into account unresolved harmonics, leads to a more efficient masking of the quantization noise.

6.2.4.1 Rationale

The main purpose of the harmonic analysis is to provide more information about the audio signal in order to achieve a better masking of the quantization noise, particularly in the case of unresolved harmonics. A better masking is reached through a more efficient noise shaping. This means that after the calculation of the critical band and inter-critical band Thresholds of Masking or Just Noticeable Distortions, contrary to what is usually done in most perceptual audio coders, the quantization noise should not be spread as white noise (*i.e.* with flat PSD) inside each critical band (or quantization band). In other words, if for example scalar uniform quantization is used, the quantization step-size should not be identical for all spectral coefficients inside a critical band. Otherwise, unresolved high level tones as well as low level tones (or other signal components) will be equally quantized. The probable outcome of this approach is that the former will be under-quantized and the latter will be over-quantized. Instead, the quantization noise should be colored as a function of the color of the signal. In this spirit, non-uniform quantization such as that used in Layer III of the MPEG-Audio standard [Brand.92b], is a better option. However, the total effectiveness of scalar non-uniform quantization has not been demonstrated for harmonic signals. The main reason is that this approach implements a local spectral coloration of the quantization noise as a result of a sum of masking effects and not as a result of a global spectral noise-shaping depending on the harmonic nature of the audio signal. We believe this contradicts the fact that the perception of the complex tone (Section 3.3.6) as well as its masking capability is a combined effect that should not be treated separately for each tonal component.

A more efficient modulation of the coding noise on a Bark or sub-Bark quantization band is possible, particularly in the higher frequency region of the audible spectrum, where many harmonics of the signal are likely to appear in one single quantization band.

Assuming scalar uniform quantization, we investigate the consequences of varying the quantization step-size inside the same quantization band. We believe that, in this way, and particularly in low bit rate coding, low level unresolved harmonics have a chance of being quantized other than to zero, in detriment of high level resolved and unresolved harmonics which are able to mask noise resulting from a coarser quantization step.

In the light of these arguments, ASC monitors the harmonic structure of the audio signal which is well defined on the linear frequency scale (and therefore not evident on the Bark scale), and uses this information to adapt the quantization operation to a suitable quantization mode[1].

6.2.4.2 The Algorithm

Several methods of pitch detection in the frequency domain are described by Kondoz [Kondoz94]. Most methods were originally developed to identify the pitch period of speech signals. In general and in our context, we found most methods to be of a limited interest and flexibility for the following reasons:

[1]It is interesting to note that this approach resembles that investigated by Tribolet and Cochiere (Section 5.3.2.4) for speech signals, which involved the parametrization of the speech spectrum into a spectral envelope and a spectral pitch pattern [Tribol.79].

- only one single pitch frequency is usually estimated,
- a computationally intensive procedure is frequently used that involves the minimization of a spectrum similarity metric,
- spectrum flattening stages are frequently used that remove important spectral density information.

Our new method is flexible enough to identify several pitches, although only one (the most relevant) is selected at the end for coding purposes [Ferre.96a]. It admits the estimation of the pitch limited to a fraction of the signal bandwidth since this is a likely scenario in music signals with many individual sounds, some of them having a relevant harmonic nature, some others not. This aspect is also valid for speech signals since only a fraction of the spectrum may be of a voiced nature.

The harmonic analysis is started by identifying all relevant peaks in the spectrum. Obviously, this assumes a time to frequency transformation with enough spectral resolution in order to allow detection of the lowest expected pitch (*e.g.* about 50 Hz), and sufficient spectral selectivity in order to allow discrimination between two very close tones. A simple algorithm for spectral peak detection based on the DFT such as that included in the Psychoacoustic Model 1 of the MPEG-Audio standard [ISO93] is adequate. It must be noticed however that a complex transform is necessary in order to evaluate the Power Spectral Density (PSD). A real transform, such as the 50% overlap TDAC filter bank which is frequently used in perceptual audio coding, does not deliver the correct information. In fact, it can be easily verified that the TDAC spectrum is given by an envelope corresponding to the Odd-DFT spectrum, and a modulating cosine function as follows[2]:

$$\mid X_{TDAC}(k) \mid = \mid ODFT(k) \mid \cos\left[\frac{2\pi}{N}\left(k+\frac{1}{2}\right)\left(\frac{1}{2}+\frac{N}{4}\right)+\phi_{ODFT}(k)\right] , \tag{6.5}$$

where $\phi_{ODFT}(k)$ is the phase of the k_{th} coefficient of the Odd-DFT.

Our interactive simulation and evaluation environment revealed that a particularly interesting but potentially misleading situation occurs very frequently at harmonic tones, as illustrated in Figure 6.5. This figure shows that as a result of the cosine modulation, a single tone might be erroneously taken for two tones.

For simplicity, we assume that the peak detection algorithm delivers the number of peaks found, $peaklast$, and delivers two vectors, $pos[i]$ and $lev[i]$, which contain respectively the position (frequency index) and level (in dB) of each spectral peak.

[2] This result is similar to that derived by Tribolet and Crochiere [Tribol.79], which relates the spectrum of the DCT with that of the DFT.

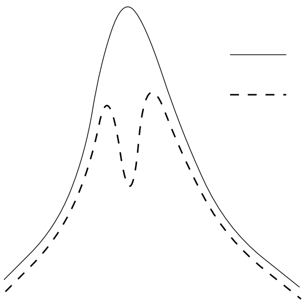

Figure 6.5: The TDAC spectrum is upper-bounded by the Odd-DFT spectrum. In specific cases, this relation might be misleading (see text).

If the vector $pos[i]$ contains the position of the tonal components of a harmonic complex, the associated pitch frequency, $pitchd$, can be computed, for example, by minimizing the following error function:

$$error = \sum_{i=1}^{peaklast} (pos[i] - i \times pitchd)^2 \ . \tag{6.6}$$

The solution is given by

$$pitchd = \frac{\sum_{i=1}^{peaklast} i \times pos[i]}{\sum_{i=1}^{peaklast} i^2} \ . \tag{6.7}$$

The harmonic analysis is implemented using this result and assumes further that the highest peak in the spectrum (which is probably the harmonic component less affected by noise) belongs to the harmonic complex.

Using the highest spectral peak, the algorithm defines first a set of very crude initial pitch estimates. For each initial estimate of pitch, the algorithm tries iteratively to find a next matching harmonic and in this case it adjusts the estimated pitch frequency. For each estimate the algorithm stops when it detects harmonic discontinuity, or when the deviation to the next harmonic exceeds a predefined error. Finally, the harmonic structure that is likely to provide a higher coding gain, is selected for coding.

The algorithm has been detailed by Ferreira [Ferre.96c]. With minor reformulations, the algorithm may also be used in other applications where the identification of several pitches is important.

6.2.5 The Quantization Noise Profile

The quantization noise profile (QNP) corresponds to the final noise shaping in frequency that meets the coding bit rate constraints.

In ASC, the QNP results from the estimated JND profile, it may receive the contribution of an existing harmonic structure in the audio signal, and it is modified as a function of bit rate constraints.

ASC estimates the JND profile in a conventional way (*e.g.* as recommended in the MPEG Audio Standard) using the simplified psychoacoustic model presented in Section 3.3.4.3, and using tonality information as proposed in Section 4.4.2.

If a relevant harmonic structure is not detected in the audio signal, ASC quantizes the TDAC spectral representation of the signal using the JND estimates, without giving any particular shape the quantization noise within the limits if each quantization band. However, if a relevant harmonic structure is detected, then the quantization and coding mode is switched to a global harmonic noise shaping in combination with the JND estimates. In this case, the harmonic structure information is transmitted to the decoder as side information: pitch frequency, number of harmonics and average level. We detail below the implication of each one of these two possibilities.

6.2.5.1 Quantization

For all possible coding circumstances, ASC quantizes the TDAC spectral coefficients using scalar uniform quantization. In the light of the arguments presented in Chapter 2, this choice has been made in order to benefit from the following main advantages:

- simplicity: in implementation and computation,
- robustness: to error propagation (as opposed for example to Vector Quantization),
- precision: since selective (over/under) quantization of a specific coefficient is easily addressed,
- low system complexity: in order to simplify the design and implementation of the rate loop control.

The "zero" region of the scalar uniform quantizer has been widened to ± 0.66 (normally it would be limited to ± 0.5) in order to avoid the unpleasant oscillations of low-level tonal noise.

In future versions of ASC, we plan to make comparative performance evaluations using other types of quantizers.

6.2.5.2 Coding of Quantization Step-Sizes

The quantization step-sizes specify the spectral coloration of the quantization noise on a quantization-band by quantization-band basis (if the QNP does not include the contribution of a harmonic pattern), or on a spectral coefficient by spectral coefficient basis (if the QNP includes the contribution of a harmonic pattern).

A most important aspect that determines the coding performance of ASC, specially at very low bit rates, is the efficient coding of all side information. In fact, since ASC does not use interframe coding (as pointed out in Section 2.2), the intraframe coding of all quantization step-sizes must be as efficient as possible so as to minimize its relative cost in bit rate.

Towards this purpose, the quantization step-sizes resulting from the QNP are decimated to avoid small variations, are logarithmically quantized and entropy encoded. Before being transmitted to the decoder, they are re-ordered and their position is run-length encoded. This processing is illustrated in Figure 6.6. For this particular example comprising eigth quantization

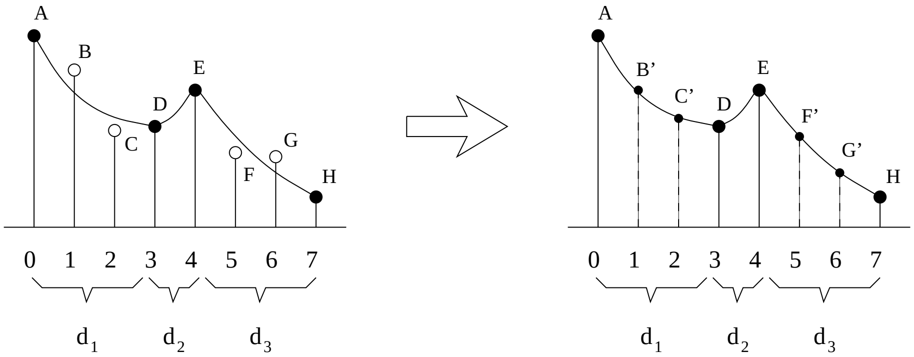

Figure 6.6: ASC decimates the quantization step-sizes as illustrated. The coding of spectral coefficients assumes the interpolated QNP as illustrated on the right-hand side.

step-sizes, only four of them (A, D, E, H) would be effectively transmitted to the decoder. The order of transmission would be: A, E, D, H, *i.e.* all maxima are followed by all minima. The correct order can be reestablished by transmitting the distances d_1, d_2 and d_3, and an additional bit specifying if the first step-size is a maximum or a minimum. After reconstruction, new scale factors B', C', F', G' are synthesized as illustrated in Figure 6.6. Obviously, the encoder must also use these interpolated step-sizes to quantize the spectral coefficients.

The interpolation of the quantization step-sizes is implemented using the following interpolation function [Ferre.95c]:

$$\left[1 - \cos\left(\pi \frac{i+1}{tr+1}\right)\right]^{EXP} , \quad 0 \leq i \leq tr - 1 , \tag{6.8}$$

where tr is the number of interpolated values. This function is represented in Figure 6.7 for several values of the parameter EXP (in steps of 0.1). ASC uses EXP=0.5 since this leads, in general, to an interpolated QNP slightly more conservative than the undecimated (original) QNP.

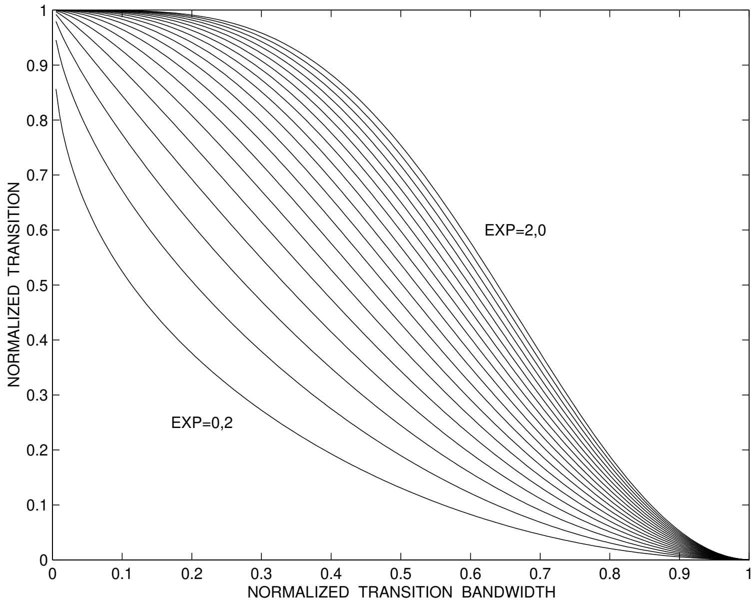

Figure 6.7: Possible solutions for the interpolation of quantization step-sizes (see text).

A final remark regards the efficient coding of all step-sizes pertaining to a set of short windows. The coding of short windows occurs when ASC switches to the improved time resolution coding mode (Section 6.2.2). Since a harmonic structure is not expected to be found in non-stationary regions of the audio signal, harmonic coding is not allowed in the improved time resolution coding mode. On the other hand, provided that the time resolution of each short window is sufficiently high (in the order of $1-2$ ms), a very simple fact takes place that can be exploited to foster coding efficiencies.

In fact, at most two clear trends exist in both the energy and the JND estimates of the signal, before and after the start of the non-stationarity. Figure 6.8 illustrates these possibilities in the case of 4:1 window switching. In this figure, the dark areas as well as the height of each short window, suggest symbolically the two tendencies. Depending on which short window the beginning of the signal non-stationarity falls, ASC sends to the decoder only one or two sets of quantization step-sizes (one for each group of short windows), as well as additional information regarding the reconstruction of the QNPs pertaining to each individual short window.

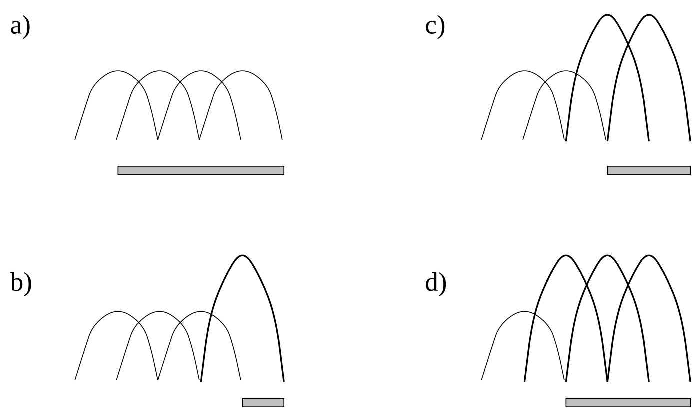

Figure 6.8: Modeling of the QNP of 4 short windows according to the start of the signal non-stationarity (dark areas): a) single QNP model for all short blocs, b), c), d) two QNPs for the short windows before and after the start of the non-stationarity.

6.2.5.3 Harmonic Quantization and Coding

If a harmonic structure is identified in the signal, it becomes possible to discriminate specific spectral coefficients in the spectrum (corresponding in particular to unresolved harmonics) which tolerate a higher quantization step than that predicted for neighboring spectral coefficients. In order to verify this hypothesis, we have conducted several experiments involving highly harmonic signals such as "pitch pipe" and "German male speech" [Ferre.96c]. As ASC uses scalar uniform quantization and an optimized analysis window that has been chosen to allow a spectral resolution of 43 Hz, we have quantized the harmonic tones using the quantization step-size estimated by the psychoacoustic model for scalar uniform quantization, and also over-quantized them by a fraction related to the log-average level of the harmonic tones. We concluded that if we over-quantize the spectral line exactly in the center of each tone[3], this effect will not become perceptible until the degree of the over-quantization is very high.

We used these results to perform a more efficient noise shaping with resolved and unresolved harmonics, particularly in low bit rate coding. While fine-tunning the procedure, we evaluated three possibilities:

- over-quantization of the harmonics inside a quantization band, which is accompanied by a concomitant reduction of the basic quantization step-size so as to maintain the same overall quantization noise estimated by the psychoacoustic model for that quantization band,

[3] Due to the high selectivity allowed by the analysis window, the most significant part of a pure tone is represented by three adjacent spectral lines.

- simple over-quantization of the harmonics by an amount depending on the overall over/under coding of the signal,
- fixed over-quantization defined by the log-averaged level of the harmonic tones.

As the last two criteria led to the best results, ASC uses a hybrid solution. First, the JND profile becomes the combination (superposition on a logarithmic scale) of the conventional JND estimate with a harmonic pattern whose level is related to the log-averaged level of the detected harmonic complex. Second, as the QNP results from the JND, as explained in Sub-section 6.2.5.4, the degree of harmonic over-quantization simply follows the general principle of the rate/distortion control.

The harmonic structure is transmitted to the decoder and the associated special quantization mode is triggered only when there is a quality advantage that compensates for the cost of side information. This cost amounts to 24 bits per frame according to the following distribution:

- pitch frequency: represented by a 13 bit float,
- number of harmonics: represented by a 6 bit integer,
- average harmonic level: logarithmically quantized to 5 bits.

6.2.5.4 Rate/Distortion Control

For transparent coding, the QNP profile should not exceed (*i.e.* should not be above) the JND profile. However, given that the *Perceptual Entropy* [Johns.88a] of an audio signal varies with time and that there are usually constraints of low or constant bit rate coding, quite often the audio signal must be undercoded, *i.e.* the QNP must be increased beyond the JND. This increase should be performed in a graceful manner so as to lead to a smooth and unobjectionable increase in loudness of the unmasked quantization noise. At this point, the two conflicting theories presented in Section 4.3.1.7 and in Section 5.3.2.4 come into play, because they suggest different approaches to the problem of *partial* masking of the quantization noise.

The first theory reflects an objective measure of perceptual distortion (NMR). Extrapolating from the JND level, it postulates that the increased QNP should be parallel to the JND profile. The second theory reflects an objective measure of the sensation of loudness and postulates that the increased QNP should minimize the loudness of the unmasked quantization noise.

In order to benefit from the best of both theories without dealing neither with complex models of loudness, nor with computationally intensive iterative approximation procedures, ASC uses a very simple approach that is supported by three main ideas.

- The functions of rate control and distortion control are not decoupled as in the MPEG-1 Audio Layer 3 standard. Thus, by increasing the QNP in a suitable fashion, the bit rate should decrease automatically, and inversely.

- *Overcoding* is achieved by shifting downwards the JND profile, *i.e.* the QNP is parallel to the JND.

- *Undercoding* is achieved by raising the JND profile in such a way that the quantization noise increases faster in the spectral valleys that it does in the spectral peaks. This method resembles to an approximate "water filling" procedure as it will be illustrated with examples in Section 6.4.

This approach to the control of the rate/distortion in ASC is implemented whether or not the JND includes a harmonic quantization pattern.

6.2.6 Huffman Coding

Independently of the value of specific coding parameters such as the sampling frequency and the average coding bit rate, ASC uses Huffman CodeBooks to code the quantized spectral coefficients, the quantization step-sizes, and run-length values.

A total of 10 Huffman CodeBooks are used: eight are dedicated to the coding of quantized spectral coefficients (COF), one CodeBook is used to code the differences between the transmitted quantization step-sizes (SCL), and another CodeBook is used to code the spectral distance (RUN) between the transmitted quantization step-sizes (as illustrated in Figure 6.6). The details concerning each Huffman CodeBook are shown[4] in Table 6.1.

Three main parameters specify the size (number of codes) of each Huffman CodeBook: the maximum absolute value (MAV) of each (integer) input value, the number of input values (n-tuples) that are converted into a single Huffman code, and whether or not the sign of each input value is included in the Huffman code. The chosen combination of these factors as shown in Table 6.1, is essentially the result of existing correlations that bring significant compression gains.

The special (dummy) CodeBook COF0 is used to signal that the spectral region to which this CodeBook is assigned, contains coefficients only quantized to zero. As a consequence, these coefficients are not transmitted to the decoder.

[4]The Efficiency column in Table 6.1 shows the difference in bits between an uncompressed coding of each input value, and the Huffman coding of that value. For example, for COF2, this difference is $\log_2(2 \times 2 + 1) \rightarrow$ the average number of bits of a COF2 coded value which has been found to be 1.30 .

Table 6.1: Characteristics of the Huffman CodeBooks used in ASC. The figures under Efficiency denote the reduction in bits per input value.

CodeBook	Values/Code	MAV	Number Codes	Signed Values	ESCape	Efficiency
COF7	2	16	289	NO	YES	$4.09 \rightarrow 3.44$
COF6	2	12	169	NO	NO	$3.70 \rightarrow 2.83$
COF5	2	9	100	NO	NO	$3.32 \rightarrow 1.98$
COF4	2	6	49	NO	NO	$2.80 \rightarrow 2.24$
COF3	4	4	625	NO	NO	$2.32 \rightarrow 1.33$
COF2	4	2	625	YES	NO	$2.32 \rightarrow 1.30$
COF1	4	1	81	YES	NO	$1.58 \rightarrow 0.64$
COF0	–	0	–	–	–	–
SCL	1	32	65	YES	YES	$6.02 \rightarrow 3.93$
RUN	1	32	33	NO	YES	$5.04 \rightarrow 3.07$

CodeBooks COF7, SCL and RUN include Escape codes allowing the coding of input values whose MAV is higher than that specified for each one of those CodeBooks. The Escape mechanism has been described by Bosi *et al.* [Bosi96].

Several CodeBooks exist to code quantized spectral coefficients in order to exploit the structure of different statistical distributions. These are mainly determined by the spectral localization of the coded audio signal and by the coding bit rate. ASC implements an optimum assignment of Huffman CodeBooks to spectral regions using a technique that has been originally developed for the AT&T PAC coder [Johns.92], and that has been recently included in the MPEG-2 AAC standard [Bosi96]. The technique may be summarized as follows:

1. *Coder sections* are first defined approximating the boundaries of quantization bands.
2. Each section is assigned a CodeBook whose MAV is the integer equal to or immediately higher than the largest quantized coefficient in that section.
3. Sections are then merged, two at a time, the final MAV being the largest of those of the two sections, so as to minimize the overall bit count of all Huffman codes.

As the merging process saves side information, the iteration of the latter step makes that the overall bit count decreases until only a small number of Huffman CodeBooks is used to span the entire coded spectral bandwidth. For stationary signals the final number of different sections is typically in the order of 3 – 4. The final number of sections, their width and the indexes of the assigned CodeBooks are transmitted to the decoder.

ASC uses the same set of Huffman CodeBooks irrespective of coding parameters such as sampling frequency and coding bit rate. In order to allow for this flexibility, different Huffman CodeBooks were populated and trained under different coding bit rates. The precise conditions were as follows:

- 48 kHz sampling frequency, constant bit rate coding mode, frame size of 512 samples, and 4:1 window switching,
- training of CodeBooks SCL, RUN, COF1, COF2 and COF3 using 32 $Kbit/s$ coding,
- training of CodeBooks COF4 and COF5 using 48 $Kbit/s$ coding,
- training of CodeBooks COF6 and COF7 using 72 $Kbit/s$ coding.

Stable CodeBooks were obtained after iterating three times this training procedure with typical audio material. The specialization of some Huffman CodeBooks for different coding bit rates explain the non-monotonic evolution of the coding efficiency shown in Table 6.1. This efficiency represents the compression gain by Huffman encoding each input value.

6.2.7 The Bit Stream Structure

In order to facilitate bit stream access and editability, ASC generates a bit stream that gathers all the information necessary to decode each frame individually. As illustrated in Figure 6.1, this function is implemented by a multiplexer that reorders all data to be transmitted to the decoder, and that provides means for frame synchronization. A general map to the ASC bit stream is shown in Figure 6.9. It consists basically of three units: a 20 bit fixed header, a variable length

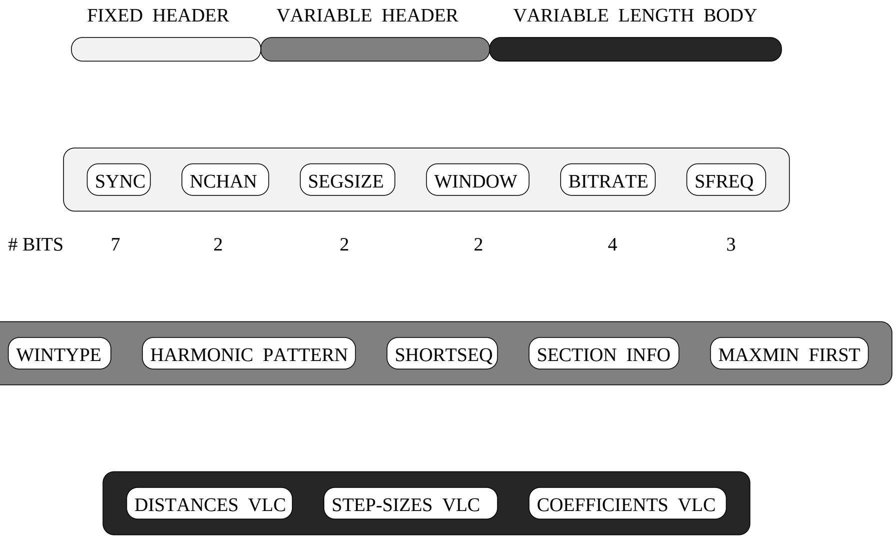

Figure 6.9: General structure of the bit stream generated by ASC. The unit identified as Variable Length Header displays all possible fields in that header although some of them are mutually exclusive (see text).

header, and a variable length body.

The fixed header includes a synchronization pattern (SYNC), and information about the number of audio channels (NCHAN), the frame size (SEGSIZE), the window switching mode[5] (WINDOW), the coding bit rate[6] (BITRATE), and the sampling frequency (SFREQ). The size of this header is the same for all bit rates. For 16 $Kbit/s$ coding of 16 kHz sampled signals for example, the fixed header represents about 8% of the available bit rate for frames with 256 samples.

The variable header contains additional information that depends on the fixed header configuration. Figure 6.9 displays all possible fields although some of them are mutually exclusive. For example, if the field WINDOW in the fixed header specifies that window switching is allowed, then the variable header might transmit one or two additional fields:

1. the type of window being used in the current frame (WINTYPE), the possibilities are those represented in Figure 6.2: “long”, “start”, “stop”, and “short”,
2. the reduced QNP models (SHORTSEQ) for short windows (as illustrated in Figure 6.8) in case WINTYPE corresponds to “short”.

The remaining possible fields of the variable header are:

HARMONIC PATTERN includes the parameters of a quantization harmonic pattern as described in Sub-section 6.2.5.3,

SECTION INFO includes the description of the spectral organization into coder sections and identifies the associated CodeBooks as described in Sub-section 6.2.6,

MAXMIN FIRST specifies if the first transmitted quantization step-size is a local maximum or a local minimum, as described in Sub-section 6.2.5.2.

The third unit (Variable Length Body) of the ASC bit stream gathers all variable length (Huffman) codes. As illustrated in Figure 6.9, the sequence of data is as follows: first the coded distances between quantization step-sizes, then the coded quantization step-sizes, and finally the quantized spectral coefficients.

As results from the above description, error protection is not currently implemented in ASC. However, since ASC does not exploit inter-frame redundancies, a higher robustness to errors is

[5]The possible window switching modes are those identified in Figure 6.2 and are symbolically denoted by TDAC0, TDAC2, TDAC4, and TDAC8.

[6]Constant bit rate coding mode is assumed and one out of 16 possible bit rate values between 16 $Kbit/s$ and 128 $Kbit/s$ can be represented in the bit stream.

achieved. In fact, upon loss of frame synchronization, very simple error concealment techniques can be used to fill the gap of a lost frame.

6.3 The Decoder

6.3.1 The Decoder Structure

A simplified diagram of the ASC decoder is illustrated in Figure 6.10. As can be concluded

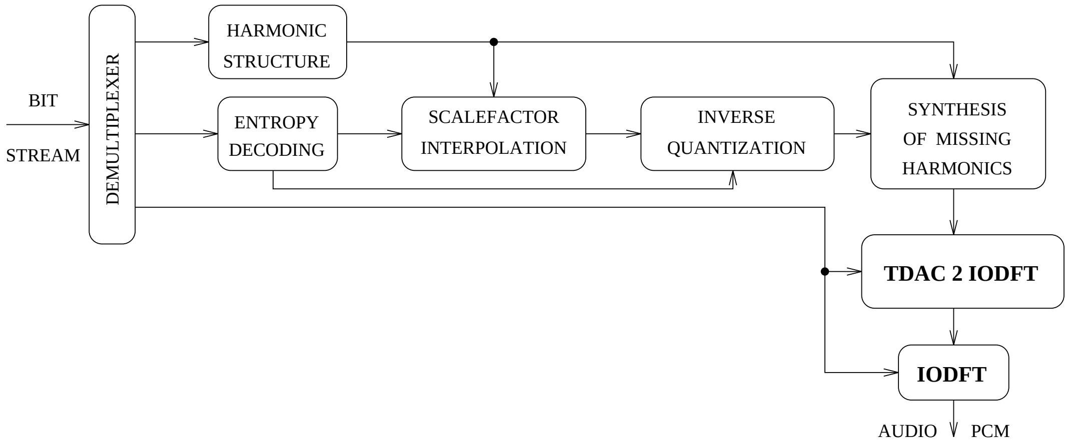

Figure 6.10: A simplified diagram of the ASC decoder.

by comparing the building blocks presented in this figure to those shown in Figure 6.1, the complexity of the decoder is significantly lower than that of the encoder.

Decoding consists of a set of actions that are taken while reading a coded bit stream, the first of which is synchronization. This set of actions corresponds to the processing steps taken at the encoder in an inverse order. Since these have already been detailed previously, in the following we just summarize the decoding aspects of major relevance.

The demultiplexer insures the correct frame synchronization and verifies the decoder settings at each frame. It then decodes information of a variable header (mainly related with the harmonic structure information) which is dependent on the information of the fixed header, and finally reads all Huffman codes related to quantization step-sizes, quantized coefficients, and run-length values.

The interpolation of scale factors is performed as in the encoder and any quantization contribution due to an existing harmonic pattern is combined with the interpolated scale factors to reconstruct the QNP.

6.3.2 Synthesis of Missing Harmonics

After inverse quantization, an important operation takes place if a harmonic structure has been transmitted. In fact, in the case of very coarse quantization of spectral coefficients, there is the possibility of missing harmonics appearing within a transmitted harmonic structure, *i.e.* there is the possibility of some relevant tonal components being quantized to zero. In this case, a better subjective quality is reached by synthesizing the missing harmonics using only the corresponding sign bits and quantization step-sizes. In order to allow for this processing, the encoder detects if relevant harmonic components are quantized to zero. If this is the case, the encoder sends the sign bit of each "missing" harmonic as side information [Ferre.96a].

The reconstructed spectral coefficients are then inverse transformed and the resulting time segments are overlapped and added as illustrated in Figure 6.2 to recover the audio signal with the same sample resolution and sampling frequency of the original signal.

6.4 A Few Results

6.4.1 ASC Evaluation at High Bit Rates

As a transform-based coder for music signals, ASC delivers higher coding gains for longer audio frame lengths as would be expected from the theory presented in Chapter 4. This is particularly true for music signals since the quasi-stationary regions of these signals tend to be much longer than the average syllabic duration of speech signals.

Using an early version of ASC that did not include the harmonic analysis and coding technique, a study has been made regarding the bit rate necessary for transparent coding of a representative "mix" of critical audio signals, as a function of the frame length (and hence of the coding delay) [Ferre.95a]. The most important results of this study are shown in Figure 6.11. Although a similar study for the current version of ASC is not yet available[7], we believe a few conclusions can be drawn from those results that are still valid for the most recent version of ASC.

1. Assuming a sampling frequency of 44.1 kHz and a frame size of 512 samples, ASC codes transparently any monophonic audio using a bit rate not exceeding 96 $Kbit/s$. As expected, this bit rate decreases if the frame size increases, and inversely.

2. As the frame size gets wider than about 20 ms, the smooth monotonic evolution of the curve displayed in Figure 6.11 reflects the effectiveness of the window switching mechanism.

[7] It would probably reveal a similar trend in a curve below that shown in Figure 6.11.

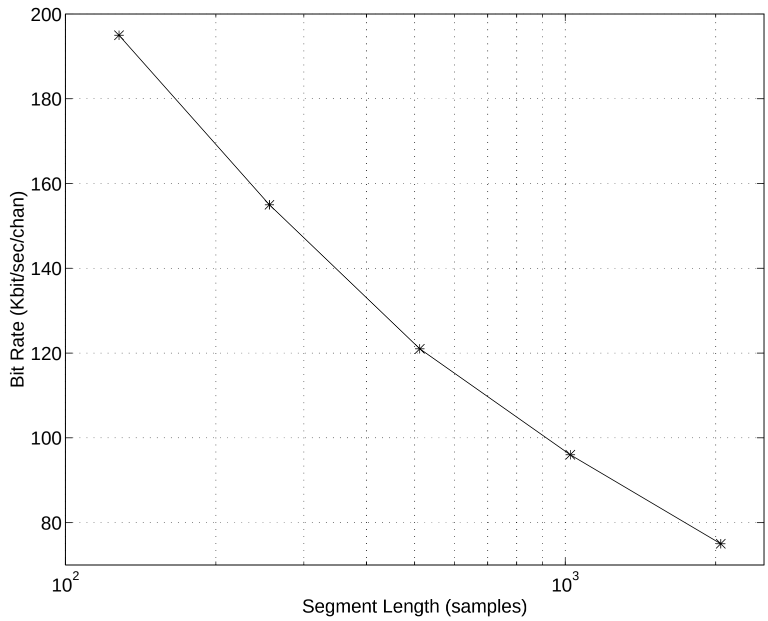

Figure 6.11: ASC coding bit rates for the transparent coding of a "mix" of critical audio signals sampled at 44.1 kHz, as a function of the segment length (which is two times the frame length). These results were derived using a version of ASC significantly less optimized than the current version.

In fact, the accurate control of pre-echo effects becomes much more critical for large frame sizes.

3. If the length of the analysis/synthesis window is in the order of 2048 samples or above, post-echo effects become noticeable that reduce the practical interest of very large frame sizes. Obviously, this implies also a longer coding delay which is not of a practical interest either.

6.4.2 ASC Evaluation at Low Bit Rates

In the following we concentrate on the coding of audio signals at low bit rates. Specifically, we present in the remaining of this sub-section our own evaluation of the coding performance of the full version of ASC, taking as a reference the MPEG-1 Audio standard, Layer 2 and Layer 3. In the next Section we report on the results of the ASC participation in the competitive tests of the MPEG-4 Audio standardization activities.

In order to test the harmonic analysis and coding technique, we used two strongly harmonic sound items: "German male speech" [Jakubo.88] and "pitch pipe". The coding parameters were as follows: 44100 Hz sampling frequency, frame size of 512 samples, 4:1 window switching, and constant bit rate coding at 32 $Kbit/s$. At this bit rate, the maximum signal bandwidth transmitted to the decoder is 7.5 *kHz*.

Figure 6.12 illustrates the coding of an audio segment belonging to "pitch pipe". The length of this segment corresponds to the duration of the analysis window: 23 *ms*. The spectrum of its Odd-DFT reveals that the signal has a well defined harmonic structure with 25 harmonics below 7.5 *kHz*. This figure also represents psychoacoustic data in the ASC quantization partition

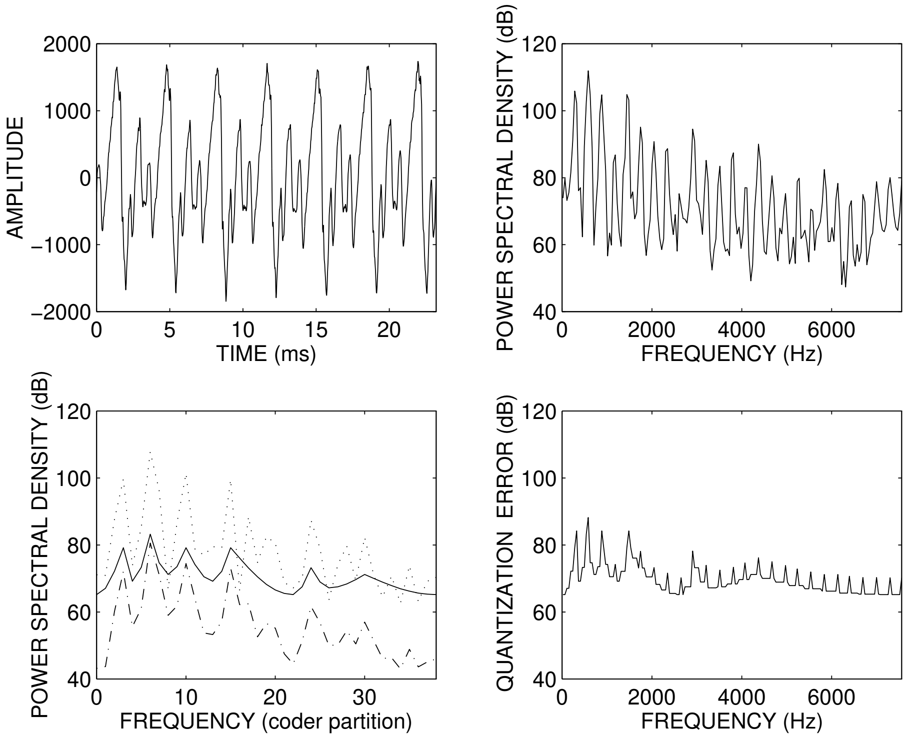

Figure 6.12: Coding of a harmonic signal by ASC. Upper left: time segment. Upper right: power density of the Odd-DFT spectrum. Lower left: spectral density in a Bark related frequency scale of the signal power (dotted line), of the estimated Just Noticeable Distortion (dashed line), and of the final Quantization Noise Profile (solid line) that complies with bit rate constraints. Lower right: spectral density of the Quantization Noise Profile in the linear frequency scale after taking into account the relevant harmonic structure of the signal.

domain that is related to a sub-Bark scale. This data consists of: the signal power density (upper line), the JND estimates (lower line) and the final QNP that meets the bit rate constraints (line in the middle). Considering that this representation is simply a mapping of the signal spectrum on a non-linear frequency scale, it can be seen that in the low frequency region the harmonic components are clearly resolved while in the high frequency region a merging effect takes place that makes the involved harmonic components unresolved in this frequency scale.

As the coder settings imposed a spectral resolution of 43 *Hz* and a good spectral selectivity, ASC detected the existence of the harmonic structure and the associated average pitch frequency of about 292.5 *Hz*. The harmonic detection as well as the quantization are performed in the same linear frequency scale, *i.e.* with the same fine spectral resolution. The final QNP is therefore obtained as the superposition, on a logarithmic scale, of the basic quantization profile (after its mapping from the *sub-Bark* scale to the *Hertz* scale) with the synthesized harmonic quantization pattern, as described in Sub-section 6.2.5.3. The final quantization profile in the linear frequency scale, after the superposition of both quantization contributions, is also illustrated in Figure 6.12. It can be seen that the 25 spectral peaks in the QNP match very approximately those of the original signal spectrum.

The effect of the harmonic analysis and coding technique after coding 20 adjacent audio frames similar to that illustrated in Figure 6.12 is shown in Figure 6.13. It can be seen that

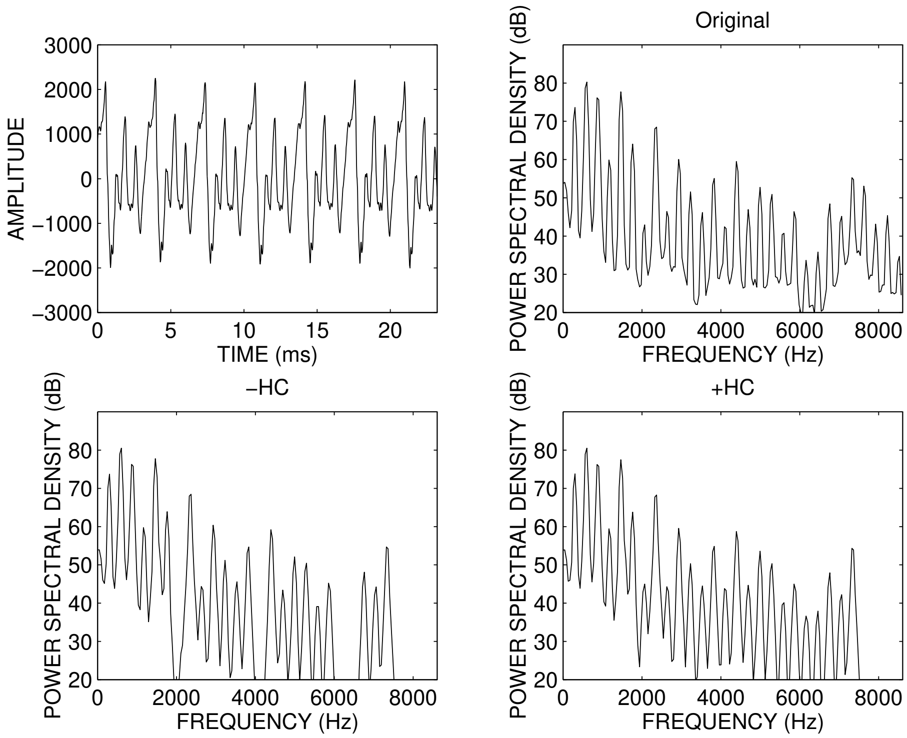

Figure 6.13: Coding (using ASC) of a 232 *ms* time segment of "pitch pipe" with and without the harmonic analysis and coding technique. Upper left: 1/10 of the total time segment. Upper right: spectrum of the original signal. Lower left: spectrum when coding without the technique (-HC). Lower right: spectrum when coding with the technique (+HC).

without the proposed technique, 4 tonal components are not coded at all, while with the proposed technique, these components are coded or synthesized at the output. This is reached at the cost of a coarser quantization of the remaining tonal components, particularly those with higher spectral power. Therefore, as we shall see later, it is not surprising that the Signal to Noise Ratio of the decoded signal with all the harmonics is inferior to that of the decoded signal without some of the harmonics.

In order to draw conclusions in a relative perspective, we have compared critical audio material coded by ASC, to the same audio material coded by the MPEG-1 Audio Layer 2 and Layer 3 coders. Specifically, we coded critical (mono) audio signals sampled at 44.1 kHz, at constant bit rate of 32 $Kbit/s$ using ASC, MPEG Layer 3, and MPEG Layer 2. As critical audio material we selected "German male speech" (4.2 sec long), "pitch pipe" (6.3 sec long), and "castanets" (2.3 sec long). The SNR obtained for each coder and audio signal is shown is Table 6.2.

Table 6.2: Signal to Noise Ratio of signals coded at 32 $Kbit/s$ by ASC (without [-HC] and with [+HC] the harmonic analysis and coding technique), and by the standard MPEG Layer 3 and Layer 2.

item	**SNR** (dB)			
	ASC -HC	**ASC +HC**	**MPEG Layer 3**	**MPEG Layer 2**
male speech	17.18	16.90	16.11	15.66
pitch pipe	15.35	14.02	13.93	11.07
castanets	6.87	6.87	6.04	5.46

Two results are presented for ASC concerning the utilization or not of the harmonic analysis and coding technique. As the results in Table 6.2 denote, the ASC harmonic analysis and coding technique is always disabled for "castanets" because a well defined harmonic structure is never found in the signal. Figure 6.14 shows the original and coded versions of the first attack of "castanets". As this representation probably suggests (together with the results of Table 6.2), the signal coded by ASC sounds not as band limited as all the remaining encoded signals. There are practically no pre-echo effects for the ASC and Layer 3 coded signals but these are very clear for the Layer 2 coded signal.

For "pitch pipe" and "German male speech" the coding with ASC reveals that the SNR obtained when coding with the harmonic analysis and coding technique is lower that when coding without it. This is not surprising since as pointed out previously, the coding or synthesis of low level harmonics is achieved at the cost of a coarser quantization of the remaining tonal components, particularly those with higher spectral power. As a consequence, a lower SNR results. But, provided that the additional quantization noise is effectively masked, a higher subjective coding quality is obtained. In fact, this has been concluded for both harmonic signals. For "pitch pipe", the lower SNR signal sounds more close to the original, specially if listening to these signals dichotically. The improved coding quality of "German male speech", when using the harmonic analysis and coding technique, is even more evident since the coded signal sounds

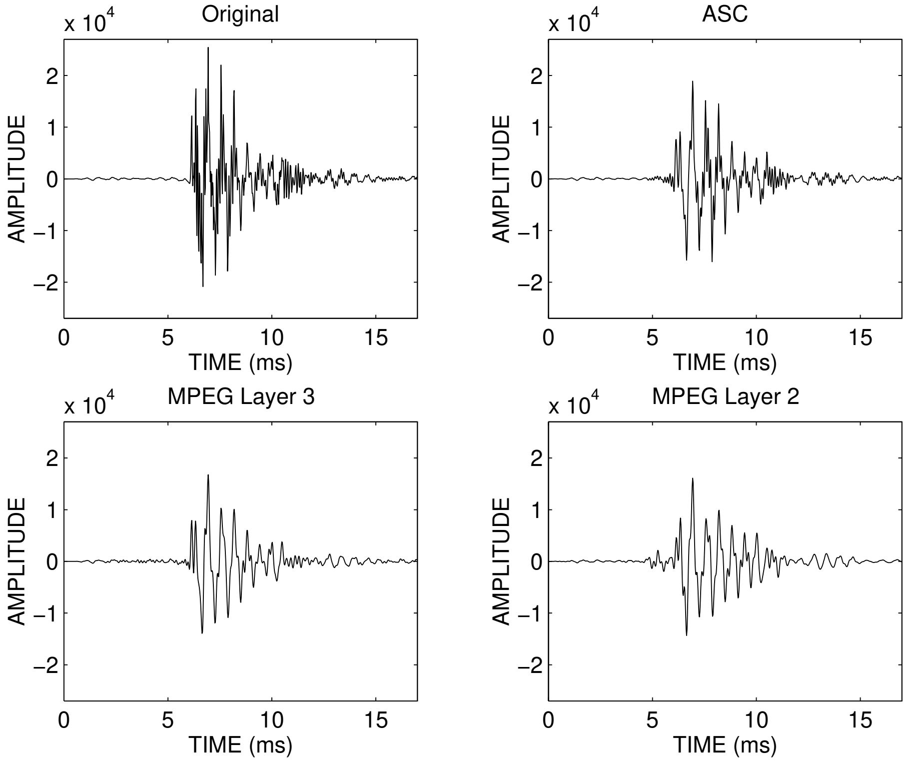

Figure 6.14: First attack of "castanets" coded at $32Kbit/s$ by ASC (upper right), MPEG Layer 3 (lower left), and MPEG Layer 2 (lower right).

much more natural and with much less artifacts than when the signal is coded without the technique.

When comparing these signals to those coded with Layer 3 and Layer 2, it is in general true that a lower SNR signal in Table 6.2 sounds worst. The exception is for "pitch pipe" which reveals no obvious difference in subjective quality when it is coded either by ASC or by Layer 3.

Figure 6.15 shows the spectrum of the exact same "pitch pipe" signal represented in Figure 6.13, when coded by ASC (repetition from Figure 6.13), with Layer 3, and with Layer 2. It can be concluded that the highest SNR obtained for the ASC coded signal is probably also a result of the fact that it has a larger bandwidth and it keeps all the harmonics till 7.5 kHz.

In the light of the considerations presented in Section 2.2, the relative performance evaluation is not complete without relating other important coder specifications besides coding bit rate and subjective/objective quality. Therefore, we give in Table 6.3 some more information regarding ASC, MPEG Layer 3, and MPEG Layer 2 coders.

The figures in Table 6.3 indicate in particular that ASC operates under much tougher conditions than MPEG-1 Layer 3. In fact, ASC uses a shorter frame length and does not exploit

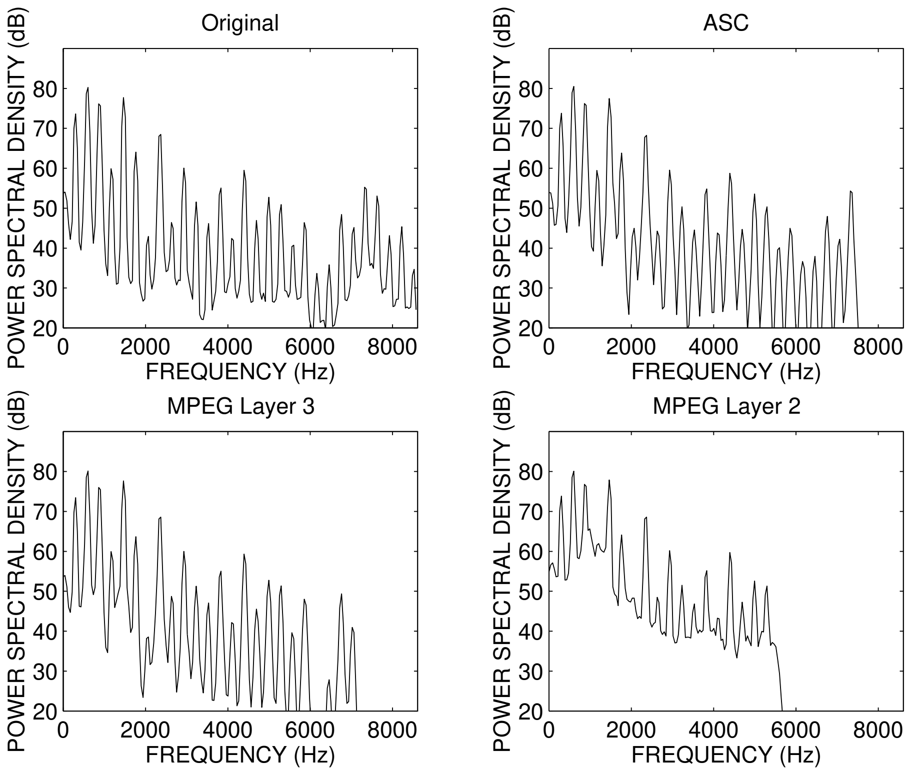

Figure 6.15: Spectrum of "pitch pipe" (upper left) after being coded at $32Kbit/s$ by ASC (upper right), by MPEG Layer 3 (lower left), and by MPEG Layer 2 (lower right).

inter-frame redundancies. On the other hand, without knowing the exact figures, we estimate that the algorithmic complexity of ASC lies between that of MPEG Layer 2 and that of Layer 3, possibly closer to the former than to the latter. Given that we do know the influence of each one of these aspects (coding bit rate, subjective quality, system delay, algorithmic complexity) on the performance Equation (2.3), we can only be excited about the global performance of ASC taking into consideration that the MPEG-1 Audio Layer 3 coder has been for a long time an acknowledged representative of the state-of-the-art in high quality audio coding.

Table 6.3: Coder specifications.

Coder	**FRAME LENGTH** (samples)	**CODING DELAY** (ms)	**INTERFRAME CODING** (frames)
ASC	512	46.4	No
MPEG Layer 3	576	63.9	Yes (2)
MPEG Layer 2	384	37.7	Yes (3)

6.5 ASC in MPEG-4

6.5.1 Introduction

The goals of the phase 4 of the MPEG-Audio standardization activities, currently in progress, have been summarized in Section 1.1.4. The competitive phase of the MPEG-4 Audio took place between November 1995 and January 1996, following an MPEG-4 Call for Proposals that was issued in July 1995. The main objective of this competitive phase has been the identification of relevant technology for inclusion in the audio part of the MPEG-4 standard.

ASC has been submitted by INESC to the MPEG-4 Audio Sub-Group in November 1995, as an audio coding proposal highly flexible, addressing a wide range of coding bit rates and two functionalities. The official input document to the ISO/IEC has the reference ISO/IEC/M425.

The functionalities and the coding bit rates at which ASC has been submitted to the MPEG-4 Sub-Group were as follows:

- compression at 16 $Kbit/s$, 24 $Kbit/s$, 40 $Kbit/s$, and 64 $Kbit/s$,
- error-resilience at 16 $Kbit/s$ and 64 $Kbit/s$.

These particular tests have been performed at four different test centers: FhG-Germany (FHG), Motorola-USA (MOT), Mitsubishi-USA (MIT), and NTT-Japan (NTT).

Several anchor signals have been generated that played the role of reference coders, especially for the 16 $Kbit/s$ coding tests. Most of them consisted in a simple form of contamination of the original audio signal with Gaussian noise whose level has been controlled to generate various "MN" signals[8] [ISO96a]. Different sets of "MN" anchors were generated for each bit rate, in order to cover the full range of quality listened to in each test.

Standardized reference coders were also used for the compression tests at 16 $Kbit/s$, 24 $Kbit/s$, and 40 $Kbit/s$:

- compression at 16 $Kbit/s$: ITU-T G.722 Standard using 64 $Kbit/s$ coding,
- compression at 24 $Kbit/s$: MPEG-2 Layer 3 (Low Sampling Frequency Mode),
- compression at 40 $Kbit/s$: MPEG-1 Layer 3.

For the 16 $Kbit/s$ compression tests, the audio material consisted essentially of speech signals sampled at 16 kHz, and organized in three groups according to the type of speech (plus sounds)

[8] "MN" is an abbreviation standing for Modulated Noise Reference Unit (MNRU), which is specified by the ITU-T Recommendation P.81 regarding subjective testing.

used in each test site: Japan and Europe/USA ("S"), only in Japan ("J"), or only in Europe/USA ("E").

For the 24 $Kbit/s$ and higher bit rates compression tests, the audio material consisted of music signals (mostly of single instruments) sampled at 48 kHz.

The 5-point MOS subjective grading scale was used for all the tests.

The list of all the proponents to the MPEG-4 Audio competitive tests is given in Table 6.4. This table includes also abbreviations that are necessary to interpret the results that will be presented later in this section. The details concerning the organization and the realization of all

Table 6.4: Proponents to the MPEG-4 Audio competitive tests.

Proponent	Name Abbreviation
Alcatel/Philips/RAI	APR
AT&T	ATT
Bosch/Cselt/Matra/Mavt	MAV
CCETT	CCE
INESC	INE
JVC	JVC
Matsushita	MAT
Mitsubishi	MIT
Motorola	MOT
NEC	NEC
NTT	NTT
NTT DoCoMo	DCM
Philips	PHI
Samsung	SAM
Sony	SON
U Erlangen/FhG	UER
U Hannover/Deutsche Telekom	UHD
Sony IPC	IPC

tests at the different test sites, as well as the analysis of the results, are described in an MPEG document with reference ISO/IEC/N1144 [ISO96a].

6.5.2 Pre-Screening Tests

After the submission period of proposals to the MPEG-4 Audio competitive tests, a pre-screening phase took place. The purpose of this phase was to detect, for all proposals, any non-conformance with the MPEG-4 requirements specified in document ISO/IEC/N999. In this case the proposal would be rejected and, as a consequence, the number of tests would be reduced.

The pre-screening tests declared in particular that the ASC error-resilience functionality was non-conformable with the test conditions requirements, and therefore should be excluded from testing. Our interpretation of those conditions was that only a mechanism for error detection and frame synchronization should be provided in order to resume synchronization after a frame loss due to errors. We were then informed that, although not explicitly stated in the requirements document, error correction was also implied.

The net effect of this appreciation was that ASC became a compression only proposal at 16 $Kbit/s$, 24 $Kbit/s$, 40 $Kbit/s$, and 64 $Kbit/s$. As a consequence of this decision, a second order effect also took place that removed ASC from testing at $64Kbit/s$ compression. In fact, at this bit rate, which was the main focus of the NBC MPEG-Audio activities, '*the test group decided to test only submissions that address functionalities in addition to compression*' [9]. As a consequence, ASC has been tested only for the functionality of compression at 16 $Kbit/s$, 24 $Kbit/s$, and 40 $Kbit/s$.

6.5.3 Ranking Criteria

The ASC bit streams submitted to the MPEG-4 Audio competitive tests, had been deliberately generated with a short audio coding frame in order to improve other coding aspects than subjective quality, namely: the time resolution when accessing to bit streams, the coding delay and the elimination of inter-frame error propagation. We took this decision because in the *MPEG-4 Plan Package Description* document, these were indicated as important new or improved functionalities.

Following our request (and protest) in document ISO/IEC/M574, those coding aspects have been (informly) addressed during the analysis of the test results (mainly during the 33_{rd} MPEG meeting in Munich), and have also been mentioned in the MPEG-4 Audio Test Results Report. However, those aspects were not consequently considered for ranking the proposals. Therefore, the "best" coders were judged so using a figure of merit relating subjective quality and bit rate which was even miscalled (in the light of arguments presented in Section 2.2) coding performance.

Although we recognize that considering too many performance parameters might difficult the process of ranking the proposals, or even make it an unmanageable process, we felt strongly that ignoring such important aspects as coding delay, was inconsistent with the spirit of previous MPEG documents and totally unfair to our proposal. Considering that other submitted proposals to the MPEG-4 Audio competitive tests had an extremely relaxed coding delay, which benefits the subjective quality of coded signals, as we argue in Chapter 2, and as we demonstrate in Section 6.4, it became clear to us that ASC entered the MPEG-4 Audio competitive tests with an obvious handicap.

[9] We regard this decision as a political one.

6.5.4 MPEG-4 Results and Ranking of Proposals

In this sub-section, we address only those MPEG-4 Audio tests where ASC has been allowed to participated *i.e.* 16 $Kbit/s$ compression, 24 $Kbit/s$ compression, and 40 $Kbit/s$ compression.

The test results regarding 16 $Kbit/s$ compression are shown in Figure 6.16. The anchor

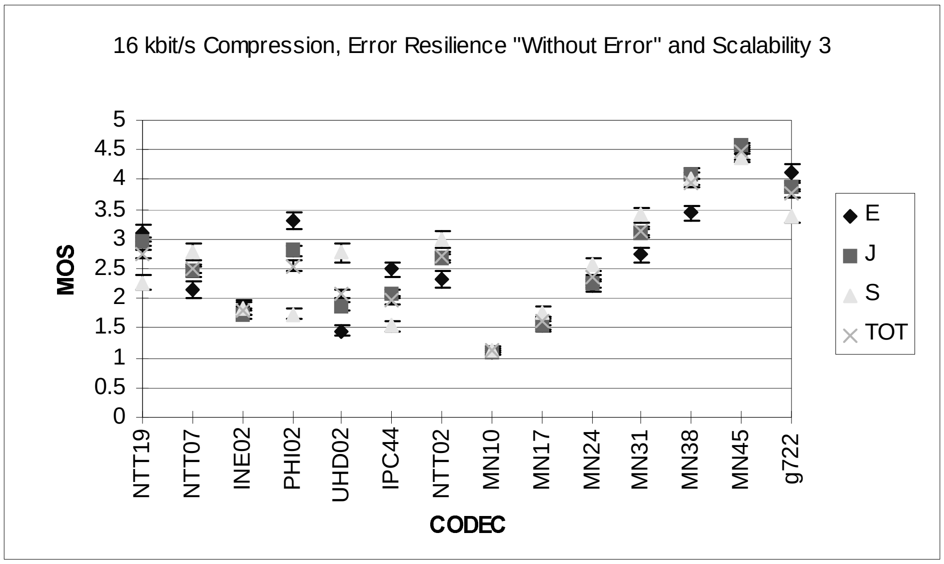

Figure 6.16: MPEG-4 Audio test results for the functionality of compression at 16 $Kbit/s$. The test signals whose first two initials are "MN" as well "g722", are reference or anchor signals (see text).

signals are identified by the initials "MN" and "g722". The proposals as well as their proponents are identified using the abbreviations of Table 6.4. The type of test audio material and the test sites are denoted by the letters "E", "J", and "S". For each one of these cases, the higher and lower scores signaled in Figure 6.16 denote the 95% confidence interval [ISO96a]. The average results are indicated by TOT.

In an absolute perspective, the average rating for ASC, at 16 $Kbit/s$ coding, is poor. Comparing to other proposals, ASC reveals to have the smallest variation over audio test material and over test sites. On average, at least two proposals perform worse than ASC for specific test items. The partial results (not shown in Figure 6.16) obtained in the MOT and NTT test sites, indicate that the ASC scores are among the best three scores over all codecs [ISO96a].

These results are nevertheless very encouraging because it is obviously very difficult for a high quality audio coder such as ASC, to compete with *speech specific coders* in a test where most of the test items are speech signals. We estimate that if ASC had been allowed to use a higher frame length than that used: 256 samples, its MOS scores would have been at least one point higher that those obtained in this test. These estimates are based on simulations we have presented to the MPEG 4 Audio Sub-group showing that, as it would be expected, if the frame size is doubled, the subjective quality of the coded material increases substantially.

The test results for 24 $Kbit/s$ compression are shown in Figure 6.17. In this figure, the

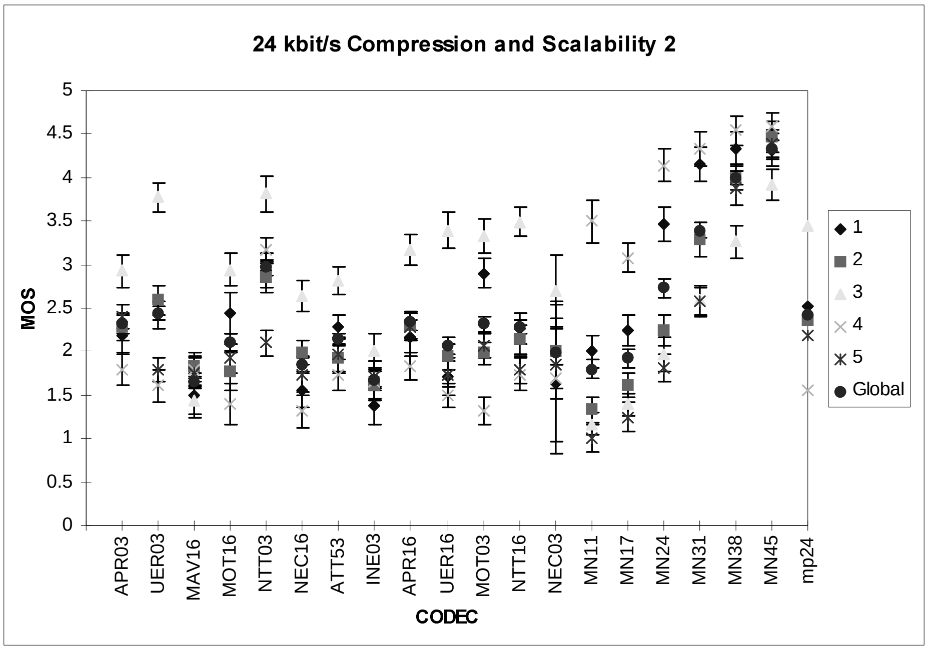

Figure 6.17: MPEG-4 Audio test results for the functionality of compression at 24 $Kbit/s$. The test signals whose first two initials are "MN" as well "mp24", are reference or anchor signals (see text). The symbols numbered 1 – 5 denote specific test items.

symbols numbered 1 to 5 denote specific test items as specified in Table 6.5.

Table 6.5: Test items used in the 24 $Kbit/s$ and 40 $Kbit/s$ compression tests.

Item Number	Signal
1	*harpsichord*
2	*female vocal*
3	*haydn*
4	*castanets*
5	*male speech*

As in the previous test, ASC reveals a small score variation over all test signals. As most ASC competitors, ASC also reveals that the best audio coding quality is achieved for a strongly harmonic signal (*haydn*) but, contrary to a general tendency shared by all the ASC competitors, the worst coding quality is not achieved for strongly non-stationary signals such as *castanets*.

We believe that if ASC had used an audio frame longer than 512 samples, its scores would have risen by at least one point in the MOS scale.

Finally, the test results regarding the functionality of 40 $Kbit/s$ compression are shown in Figure 6.18. At this bit rate, ASC reveals clearly a higher coding quality than that reached by a

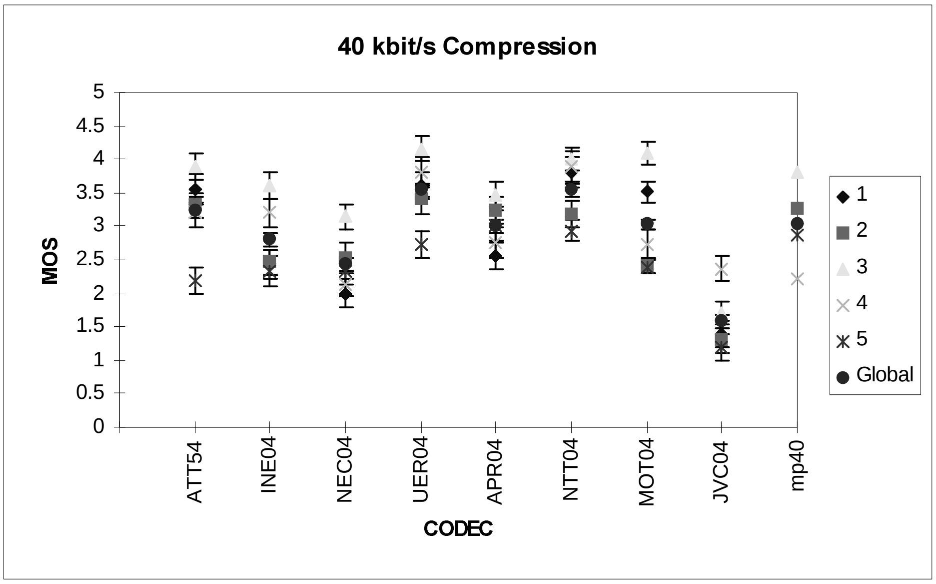

Figure 6.18: MPEG-4 Audio test results for the functionality of compression at 40 *Kbit/s*. The test signal identified as "mp40" is an anchor signal. The symbols numbered 1 – 5 denote specific test items (see text).

few ASC competitors for the same test item. It is also interesting and perhaps revealing to note that contrary to a general tendency in the results, the highest coding quality of ASC is obtained for a strongly harmonic signal (*haydn*) and for a strongly non-stationary signal (*castanets*).

Another conclusion may also be derived from Figure 6.18: the ASC scores as well as those of the reference coder MPEG-1 Audio Layer 3, are reasonably similar. This indicates that our independent evaluation of the ASC performance relative to the MPEG-1 Audio standard, as detailed in Section 6.4, is fair and reasonably unbiased.

Again, the extrapolation is also pertinent here and therefore inevitable: had ASC been allowed to use a frame size similar to that used by some of the ASC competitors, its relative ranking would certainly be much more favorable.

Overall, and despite the above difficulties and other factors ignored during the MPEG-4 Audio competitive tests (*e.g.* flexibility and overhead resulting from addressing several functionalities with a single coder/decoder proposal), the final results revealed two more interesting conclusions concerning the ASC coding proposal:

- ASC was one of the only two submitted proposals that addressed a large range of coding bit rates (16, 24, 40, and 64 *Kbit/s*) using the same coder/decoder executable,
- As the bit rate of the compression tests increases, ASC reaches a higher relative ranking position.

It is therefore inevitable to foresee that even more interesting results would have been obtained at 64 $Kbit/s$ compression if ASC had been allowed to be tested at this bit rate.

As a conclusion, we feel that the involvement of ASC in the MPEG-4 Audio tests could have been much more successful if we had relaxed freely the coding delay. In spite of this, we feel that the outcome of the MPEG-4 Audio adventure is twofold. First, ASC had an effective impact on the draft of the first MPEG-4 Verification Model and this certainly encourages further algorithmic research. Second, despite the limited number of tests ASC went through, the available test results indicate that ASC is an advanced audio coder whose unusual flexibility and performance validate our pragmatic approach concerning design and optimization.

6.6 Summary

We have described in this Chapter the structure and operation of the ASC encoder and decoder.

The performance of ASC has been illustrated using the results of two different evaluation tests: our own independent tests taking the MPEG-1 Audio coding standard as a reference, and the formal subjective tests carried out by the MPEG-4 Audio Sub-group, during the competitive phase of the MPEG-4 Audio standardization activities, on all coding proposals (including ASC) submitted to the MPEG-4 Audio.

Chapter 7

ASC - Additional Functionalities

omne ignotum pro magnifico
- [if it is unknown it is magnificent]

Tacitus, Roman historian, 2nd century.

7.1 Introduction

In this Chapter we exploit the ASC analysis/synthesis filter bank in order to implement three functionalities beyond coding: equalization, synthesis and modulation of artificial noise, and time-scale expansion. Although this research has been inspired by the MPEG-4 standardization objectives and activities, it reflects also our interest in an incursion in the area of special effects in audio. Specifically, we have been motivated by the following goals:

- to minimize the use of independent pre/post-processing modules in order to implement the above functionalities,
- to understand current pre/post-processing practices and techniques and to develop new (and preferably) simplified techniques within the analysis/synthesis framework of ASC,
- to gain specific insight regarding the spectral manipulation of audio signals in order to feed-back that specific knowledge to the optimization of perceptual audio coding techniques.

The implementation of the above functionalities is conveniently addressed by inserting pre-processing functions and/or post-processing functions in the ASC filter bank, as it is shown in Figure 7.1. This figure may be seen as a fragmentation of the TDAC filter bank shown in Figure 4.5, and as a generalization of the ASC encoder and decoder diagrams shown in Figure 6.1 and in Figure 6.10. For simplicity, we assume in this Chapter that the coding building block

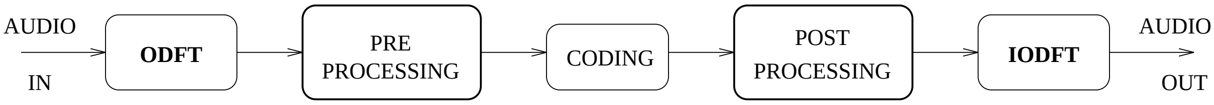

Figure 7.1: Reformulation of the ASC analysis/synthesis framework in order to accommodate pre/post-processing.

in Figure 7.1 is omitted[1] which means we consider a more general analysis/synthesis framework than the TDAC filter bank. Analytically, this generalization is obtained by modifying Equation (6.1) and Equation (6.2) according to the decomposition suggested by Equation (6.4). As a result, a direct Odd-DFT transformation is implied in the 50% overlap frame analysis:

$$X(k,m) = \sum_{n=0}^{N-1} h(n)x(n+m\frac{N}{2})e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n} \,, \tag{7.1}$$

an inverse Odd-DFT transformation in implied in the 50% overlap frame synthesis:

$$\hat{x}(n,m) = \frac{1}{N}\sum_{k=0}^{N-1} X(k,m)e^{j\frac{2\pi}{N}(k+\frac{1}{2})n} \,, \tag{7.2}$$

and the reconstruction is given by Equation (6.3), without modification, which we repeat here for convenience:

$$\tilde{x}(n) = \sum_{m=-\infty}^{\infty} h(n-m\frac{N}{2})\hat{x}(n-m\frac{N}{2},m)\,.$$

We recall that m is the time block index and $h(n)$ is the analysis/synthesis time window.

Each functionality is addressed in an independent section of this Chapter. We present in Section 7.2 a study of the difference between circular convolution and linear convolution, and we propose a simplified procedure to approximate, in the frequency domain, an arbitrary equalization curve, while insuring total inaudibility of the allowed but highly attenuated aliased signal components. In Section 7.3 we present a study of the types of impairment encountered in old recording and film sound tracks, and we propose a simple process of controlled contamination of a high quality audio signal so as to make it sound like an impaired old recording. In Section 7.4 we describe a new time-scale expansion algorithm that combines frequency-scale modification and time interpolation. The algorithm is non-iterative and is constrained to a *blind* magnitude and phase modification on a frame basis. The performance of the algorithm is illustrated with both synthetic and natural audio signals. Finally, Section 7.4 summarizes this Chapter and discusses future work.

[1]Further research is therefore needed to evaluate integration effects.

7.2 Equalization

7.2.1 Introduction

The advantages of digital audio equalizers over analog equalizers have been addressed and demonstrated by several authors [Henriq.90][Kraght92]. There are essentially two main aspects of particular relevance: quality and efficiency. In terms of quality, typical problems of analog equalizers like group delay distortion, interband interference, difficulty to achieve a perfectly flat frequency response, poor global system controllability; may be completely eliminated or attenuated within some desired mathematical precision by using digital processing techniques. With respect to efficiency, it is possible to implement very efficient high quality digital filters using multirate structures [Crochi.83], or using well-known frequency domain techniques [Oppenh.75]. These reasons, complemented by the fact that there are commercially available low-cost, powerful DSP processors, lead to very attractive implementations capable of a successful competition in the sensible market of high quality audio equipment.

Symmetric Finite Impulse Response (FIR) filters are usually considered for digital equalization because they are inherently linear phase and thus they eliminate group delay distortions. Several digital equalizers use the coefficients of the impulse response of these filters to perform a direct linear convolution in the time domain with the audio data. The number of bands implemented by this type of equalizers in normally low (*e.g.* 10). This practical limitation is due to the fact that specifying many bands requires very high spectral selectivity which, in turn, implies very long digital filters. This leads obviously to a proportionally higher computational load and long system delays that may not be suitable for certain applications [Henriq.90].

A much more efficient implementation of linear convolution may be achieved by using frequency domain techniques such as the known methods of *overlap-add* (OA) and *overlap-save* (OS) [Oppenh.75]. However, several precautions must be taken in order to insure that the circular convolution that is in fact implemented by these methods, corresponds in practice to the linear convolution. This implies that after a modification of the frequency response of the equalizer, a new filter must be designed using some filter design tool, the most common of which is the Parks MacClellan program [McCle.73b]. This is explained by the fact that the OA/OS methods specify a maximum filter length that can not be exceeded because otherwise unwanted aliased signal components are introduced in the processing.

The design of a digital filter constrained to a predefined filter length involves normally a considerable computational effort. This is illustrated is Figure 7.2 (dashed path), where it is shown that between the specification of the desired frequency response of the equalization filter,

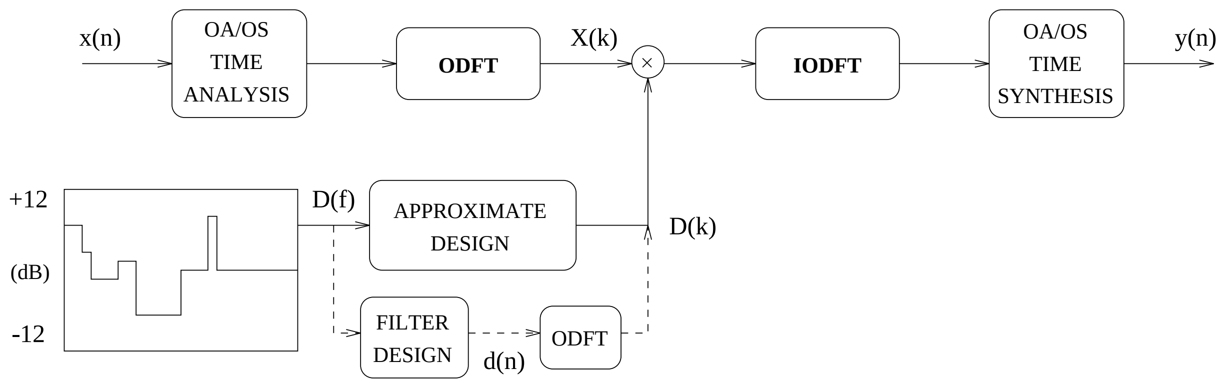

Figure 7.2: Reformulation of the conventional OA/OS methods in order to simplify the design of the equalization filter.

$D(f)$, and a suitable discrete spectral approximation, $D(k)$, a discrete filter $d(n)$ complying with the length constraint, is first obtained.

Since this is the most critical aspect that is associated with the highly efficient frequency domain OA/OS methods, we investigate in this section a fast and efficient approximation to the desired equalization filter that ignores the length constraint. The new approach is also illustrated in Figure 7.2 (solid path). Naturally, as a corollary of this approach, aliased components should be expected. However, instead of seeking their mathematical elimination as the OA/OS methods do, it is sufficient to eliminate their perceived presence in the processed signal. This may be achieved using a simple strategy that we will follow here and that comprises three steps:

1. to study the difference between linear convolution and circular convolution,

2. to identify and isolate the unwanted aliased components,

3. to find and evaluate ways to the objective and subjective minimization of the unwanted aliased components.

This section is organized as follows: in Sub-section 7.2.2 we shown how our analysis/synthesis framework relates to the classic OA and OS methods, and we revisit the theory developed in Section 4.5.5 expressing the circular convolution operation as a function of the linear convolution and a set of *minimized* aliasing terms. In Sub-section 7.2.3, a 20-band graphic equalizer is designed considering masking concepts, the whole equalization response is characterized and illustrated, and one example of frequency approximation to arbitrary equalization settings is shown. In section 7.2.4 the results of an informal subjective test evaluation are presented.

7.2.2 Circular Convolution and Linear Convolution

The diagram in Figure 7.2, which characterizes our analysis/synthesis framework, also reflects the scheme of a *generalized* OA/OS method presuming a 50% overlap between consecutive signal transformations[2].

In fact, although the OA and OS methods are traditionally presented as two different methods [Oppenh.75][Kulp88], they can be seen as the dual of each other, as we illustrate in Figure 7.3. This means both methods can be reduced to a simple scheme that corresponds to our

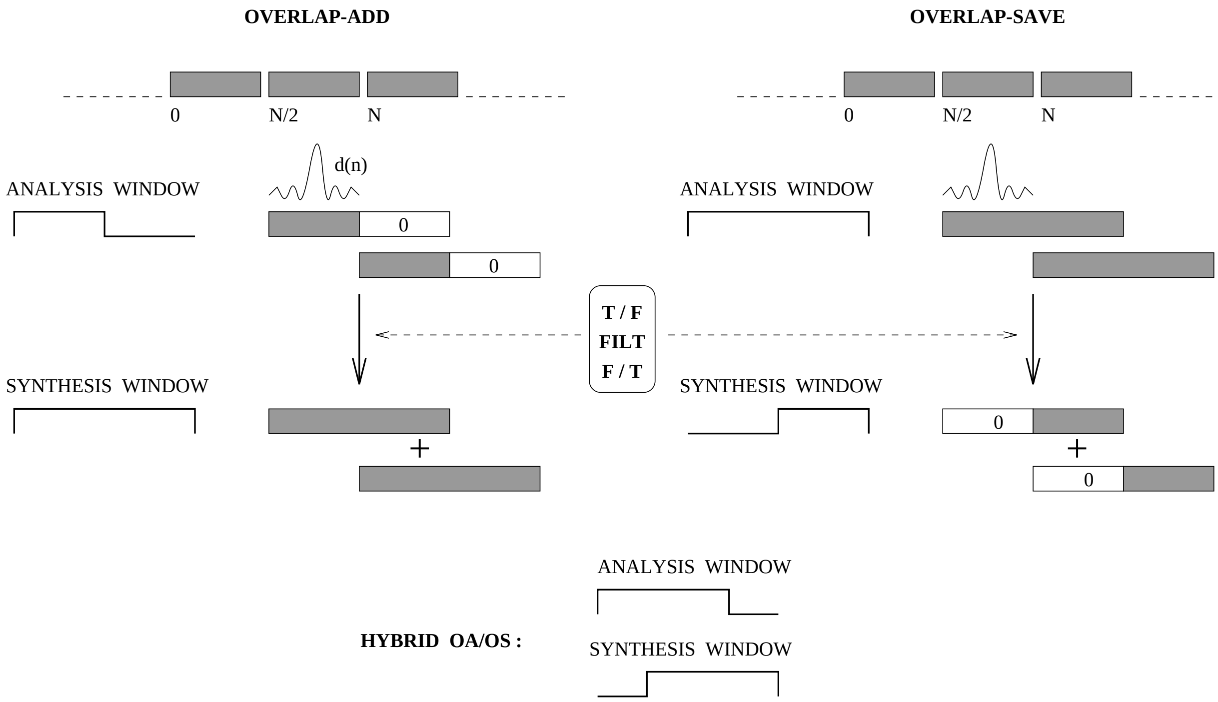

Figure 7.3: Simplified scheme of the overlap-add (OA) and overlap-save (OS) methods when $N_{data} = N/2$. These methods are the dual of each other in the sense that the analysis and synthesis windows, as well as their role, are reversed in the two cases. As illustrated, an hybrid OA/OS method is obtained by simply using suitable analysis/synthesis windows.

analysis/synthesis framework. Briefly, as illustrated in Figure 7.2 and in Figure 7.3, a segment of audio data (comprising two frames of $N/2$ samples each), is multiplied by an analysis window and then transformed to the frequency domain using the DFT or, as in our context, the Odd-DFT. This spectral representation is then multiplied by the spectral representation of the impulse response of the desired equalization filter $d(n)$. This product is then inverse transformed to the time domain and multiplied by a synthesis window. Successive segments are finally overlapped

[2] The computational efficiency of the OA or of the OS methods are higher when the length of the audio data effectively processed and the length of the digital filter come close to half the length of the transform [Kulp88].

and added to reconstruct the output signal.

In reality, this scheme implements the circular convolution in the time domain. In order to reduce the circular convolution to a pure linear convolution, the OA/OS methods specify two important conditions:

1. The length of the transformation, N, is related to the data length, N_{data}, and to the filter length, N_{filter}, by:
$$N \geq N_{data} + N_{filter} - 1 \,. \tag{7.3}$$
We will assume throughout this Chapter that $N_{data} = N/2$, to be consistent with the 50% overlap.

2. The analysis and synthesis time windows should be such that the filtering corresponds to a linear time convolution using $d(n)$ and not a modification thereof.

If $d(n)$ is a linear-phase FIR filter, as implied in Figure 7.3, whose length complies with $N_{filter} \leq N/2 + 1$, then the above conditions are met using a pair of windows as illustrated in Figure 7.3.

Considering the 50% overlap, in the OA method, the analysis window is a half length rectangular window and the synthesis window is a full length rectangular window. The role of each window is reversed in the OS method and each window is itself reversed in time. The half length rectangular window is usually seen as keeping half segment of audio data and then zero padding the remaining half of the segment. This window is responsible for avoiding (OA) or discarding (OS) unwanted aliased components of the circular convolution and hence only those components which corresponding to a linear convolution operation are retained.

A hybrid OA/OS method is obtained by just modifying the shape of the analysis/synthesis time windows, as illustrated in the bottom of Figure 7.3.

Since our analysis/synthesis framework assumes symmetric and identical analysis/synthesis time windows, a correspondence to the OA/OS methods can be established only if $d(n)$ is a zero-phase filter. The implication of this modification is depicted in Figure 7.4. The processing is the same as indicated above but the major new aspect is that a hybrid OA/OS method can be found involving symmetric and identical analysis/synthesis time windows. Unfortunately these are such that their frequency response is rather poor, and their impulse response stimulates the appearance of blocking noise. These aspects are highly undesirable in audio coding. Smoother and full-length analysis/synthesis windows should be used instead but this stimulates the appearance of aliasing effects due to the circular convolution. Under this assumption, it does not make sense to constrain $d(n)$ to a length smaller than N.

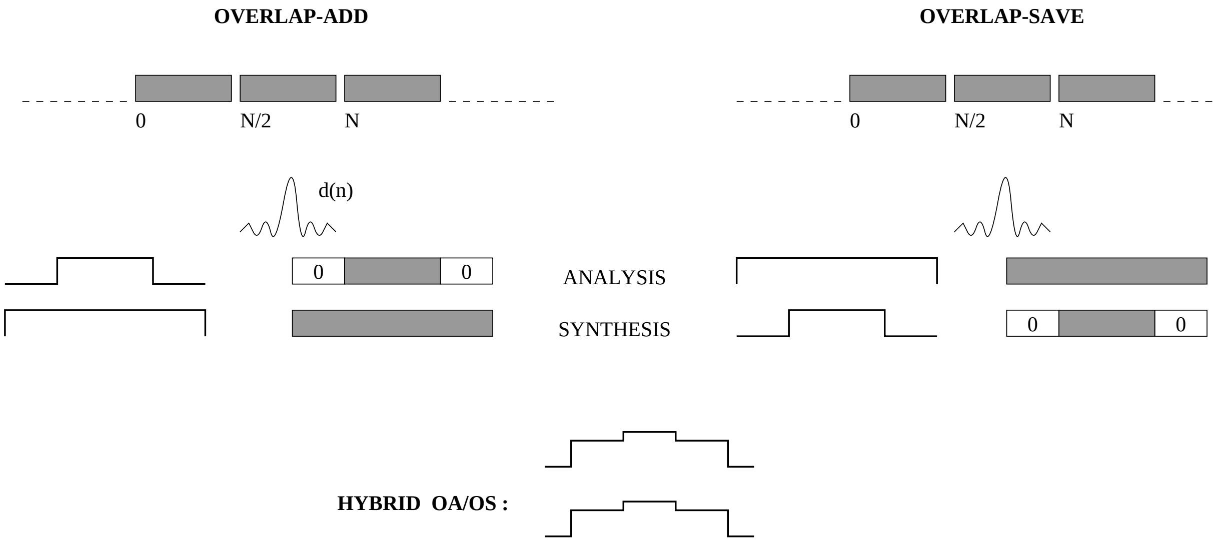

Figure 7.4: Illustration of the analysis/synthesis time windows used in a 50% OA/OS method when the equalization filter is zero-phase. Bottom: a hybrid OA/OS method is possible where both windows are symmetric and identical.

By not constraining neither the length of $d(n)$, nor that of the analysis/synthesis windows, the above OA/OS methods reduce to the conditions of the ASC analysis/synthesis framework. The whole issue is therefore to design optimal, full length analysis/synthesis windows such that undesirable circular convolution terms are minimized for an arbitrary, full-length equalization filter $d(n)$. Such a study has been presented in Section 4.5.5. Two main conclusions have been reached after the terms associated to linear filtering had been isolated from the aliasing terms:

1. the resulting linear filter is a time-varying, smoothed version of $d(n)$, which we have been referring to as $\hat{d}(n)$ (Equation 4.63),

2. time-domain aliasing terms could be minimized leading to optimized analysis/synthesis time windows.

Specifically, if the analysis time window is denoted by $h(n)$ and the synthesis time window is denoted by $f(n)$, then the resulting linear filtering is given by Equation (4.76), and the aliased terms are given by Equation (4.77) and Equation (4.78). To verify how these results simplify to the known OA/OS methods, we take the example of the OA method illustrated in Figure 7.4, where $h(n)$ is given by:

$$h(n) = 0.0\ , \qquad 0 \leq n \leq \frac{N}{4} - 1\ ,$$
$$h(n) = 1.0\ , \qquad \frac{N}{4} \leq n \leq \frac{3N}{4} - 1\ ,$$

$$h(n) = 0.0\ , \qquad \frac{3N}{4} \leq n \leq N-1\ , \tag{7.4}$$

$f(n)$ is given by:

$$f(n) = 1.0\ , \quad 0 \leq n \leq N-1\ , \tag{7.5}$$

and $d(n)$ is a zero-phase filter complying with Equation (7.3). If these windows are involved in the OA or OS convolution, it results that $\hat{d}(n-p)$ in Equation (4.76) will be exactly $d(n-p)$. On the other hand, it can also be easily proven that aliased terms resulting from Equation (4.77) would appear for $\frac{N}{4}+1 \leq n-p \leq \frac{N}{2}$ and those from Equation (4.78) would appear for $-\frac{N}{2}+1 \leq n-p \leq -1-\frac{N}{4}$ for an arbitrary $d(n)$. However, if our zero-phase filter $d(n)$, in accordance with the inequality condition of Equation (7.3), has coefficients only in the range $-\frac{N}{4} \leq p \leq \frac{N}{4}$, it results that all the aliased terms due to the circular convolution are zero. However, if the filter length is just one tap longer than that considered, then disturbing aliased terms will appear because, in particular, the window discontinuities in Equation (7.4) are very severe. This phenomenon is one possible form of the blocking noise well-known in the area of transform coding [Jayant84].

Taking into account perceptual aspects, the advantage of using smooth and optimized $h(n)$ and $f(n)$ windows is twofold:

- A relaxed design of the equalization filter is possible without satisfying Equation (7.3) because the existing aliasing terms may be attenuated beyond the Threshold of Masking, *i.e.* it is possible to make those aliased terms inaudible in the presence of the audio signal.
- The modification implied in Equation (4.76) on any desired band-pass filter $d(n)$ has the potential to improve substantially its spectral characteristics.

These aspects are very helpful in simplifying the filter approximation to any desired equalization curve. In fact, we have concluded that in order to obtain good filter characteristics, it was sufficient to consider flat response within the pass-band limits of each equalizer band, and to modulate the transition between adjacent equalizer bands using a simple smoothing function. Several possible solutions have been compared and evaluated taking into special consideration the symmetry between the band-pass filters of the equalizer above and below 0 dB. We have concluded that good results were obtained using a function given by Equation (6.8).

7.2.3 A 20-Band Graphic Equalizer

We have seen in Section 3.3.2 that masking effects are maximized when both masker and masked components lie within the limits of the same critical band [Hellman72][Scharf70].

In order to further reduce the possibility of the attenuated aliased components from being heard, and given that most of their energy is concentrated in the region of the equalized spectrum [Ferre.96b], we have designed the different equalizer bands of the equalizer so as to keep them close to the limits of the critical bands. The center frequencies of all 20 band of the equalizer are shown in Table 7.1.

Table 7.1: Center frequency of each band-pass filter of the equalizer.

Band	**Center Frequency** (Hz)
1	47
2	140
3	234
4	352
5	492
6	633
7	820
8	1008
9	1242
10	1523
11	1875
12	2320
13	2930
14	3680
15	4805
16	6352
17	8507
18	11414
19	15210
20	20883

Table 7.2: An example of equalization settings.

Band	**Setting** (dB)
1	8
2	0
3	-8
4	0
5	0
6	-12
7	0
8	12
9	0
10	-2
11	-8
12	-4
13	5
14	0
15	0
16	0
17	0
18	0
19	0
20	0

The length of the analysis/synthesis time window is 1024 samples which means we can reach a spectral resolution of 47 Hz for the sampling frequency of 48 kHz. Transition bandwidths occupy 2 Odd-DFT spectral lines in the low frequency bands, and 8 Odd-DFT spectral lines in the high frequency bands. In order to obtain a reasonable symmetry in the response of the equalizer, transition bands are modulated differently above and below 0 dB, and at low or at high center frequencies, using a different parametrization of Equation (6.8) for each band.

Ignoring the aliasing terms, the frequency response of the equalizer when all bands are individually set to $+12dB$ and to $-12dB$, is depicted in Figure 7.5. The ripple around 0 dB when any band is maximally accentuated or attenuated is less than 0.15 dB. Furthermore, when adjacent bands are set to the same accentuation or attenuation value, the frequency response

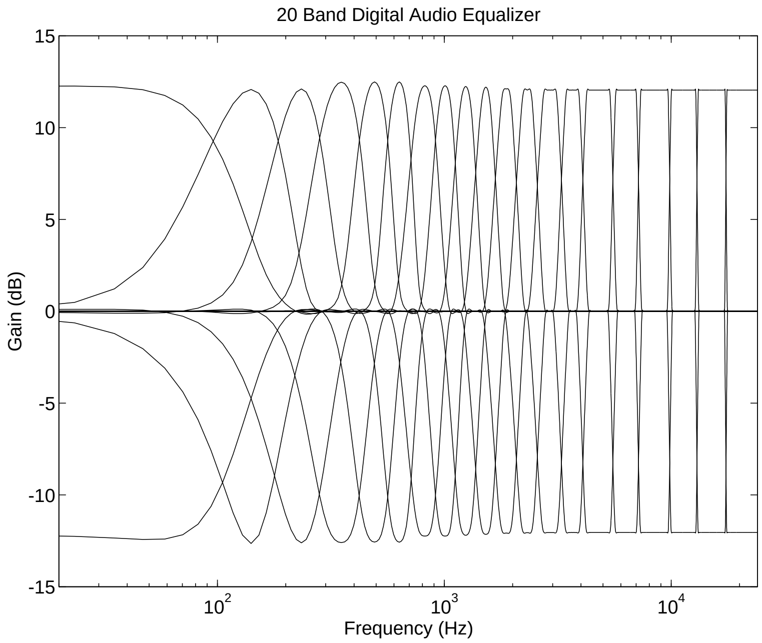

Figure 7.5: Frequency response of the 20-band digital audio equalizer.

between their limits is perfectly flat and free of significant energy of aliased residuals produced by other equalized bands. In the limit, if all bands are set to the same arbitrary equalization parameter, then the resulting frequency response is absolutely flat and there are no aliased residuals. With exception for the system delay, this corresponds to a multiplication of each time sample by the same constant. This is a desired property which is not met at all or only slightly approximated by other digital equalizers [Henriq.90].

As an example, we have defined a set of hypothetical equalization settings as specified in Table 7.2. A direct map of these settings to the frequency plane taking into consideration the center frequencies and the bandwidth of each band, is represented in Figure 7.6. The approximation algorithm smooths out the transitions between adjacent bands as illustrated in Figure 7.7. A transformation of the smoothed spectral equalization curve to the time domain yields filter $d(n)$. The final equalization filter $\hat{d}(n)$ is obtained by smoothing the shape of $d(n)$ in the time domain using the functions of the analysis/synthesis time windows (Equation 4.76). The N coefficients of the impulse response of the final equalization filter are depicted in Figure 7.9. The corresponding frequency response is illustrated in Figure 7.8. It may be verified that this frequency response exhibits high spectral selectivity and smoothness.

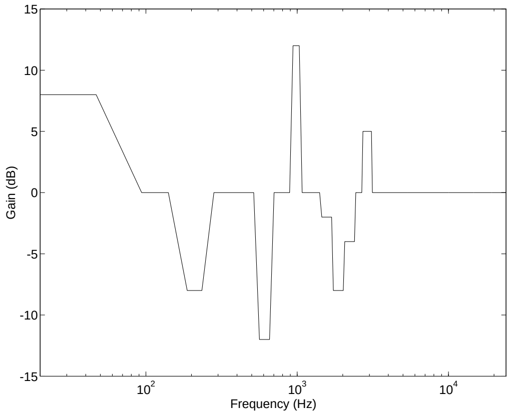

Figure 7.6: Example of a desired equalization curve.

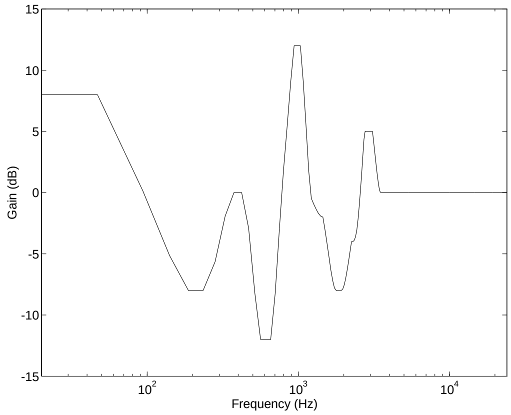

Figure 7.7: Example of a smoothed equalization curve.

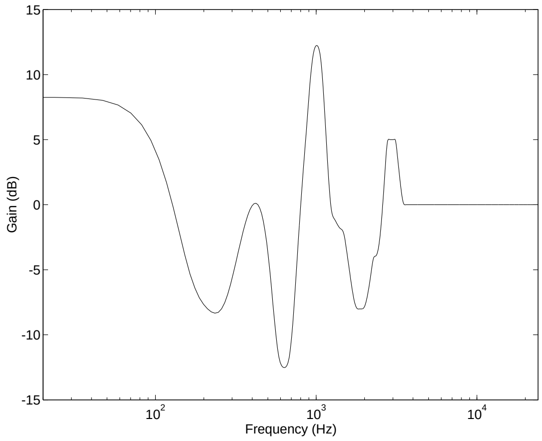

Figure 7.8: Frequency response of the approximated equalization filter.

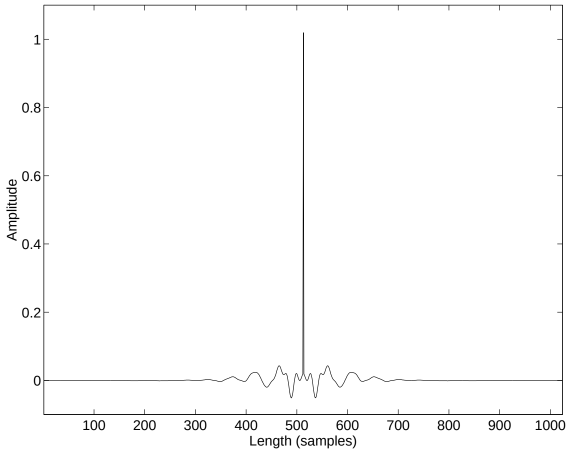

Figure 7.9: Impulse response of the approximated equalization filter.

7.2.4 Subjective Tests

Since the impulse coefficients of the (ideal) equalization filter are available, we have the possibility to determine exactly the aliased terms. In fact, they correspond to the difference between the result of the equalization process as illustrated in Figure 7.2, and the result of a linear convolution operation in the time domain between the audio signal and the FIR filter $\hat{d}(n)$. We used this approach to perform a few listening tests.

Three experienced listeners were involved and three particularly critical test items (single instruments) were used: "bass guitar", "castanets" and "harpsichord". The latter two were taken from the SQUAM Compact Disc [Jakubo.88]. For each arbitrary equalization settings, two equalized signals were generated as explained above and as illustrated in Figure 7.10. The

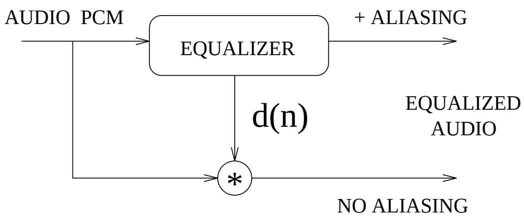

Figure 7.10: Generation of two equalized signals for subjective evaluation. One version is obtained by linear convolution in the time domain using the ideal equalization filter. The other version is obtained by filtering in the frequency domain.

listeners were asked to identify which signal was corrupted by aliased signal components. We used both non-simultaneous monaural presentations, as well as simultaneous binaural presentations, to improve the possibilities of detection. We concluded that even for very demanding equalization settings like those illustrated in Figure 7.8, the listeners could not clearly identify the two versions.

We compared the spectrum of the ideally convolved signal to the difference signal between the two versions and noticed that, for singularly equalized bands, where the energy of the difference signal is more concentrated, the minimum SNR value found was higher than 28 dB. This value happens to exceed largely the masking requirements for transparent encoding of audio signals, as we have concluded in Section 3.3.4. It is therefore possible to conclude that our optimized transform window, the careful design of equalization bands, and a smooth and simple approximation to the spectral equalization curve, lead to the generation of aliased signal components which, for the examples tested, were completely masked by the signal and hence, not audible.

7.3 Noise Synthesis and Modulation

7.3.1 Introduction

An aspect of particular concern in perceptual audio coding is the presence of signal impairments. The most obvious examples are audio signals originating from old recordings or from old movie sound-tracks. Perceptually, an impairment is clearly perceived as a contamination of the main audio signal by spurious components. Such distinction is not made by a perceptual audio coder that takes all audio components as resulting from a single source of audio information.

Conceptually, the implication of a signal impairment on the coding process is twofold:

- depending on the type of impairment, the opportunities for masking increase which means a more efficient coding may be achieved,

- the impairment misleads the psychoacoustic model causing wrong bit allocation and coding decisions which means a worse coding quality may be expected.

These two possibilities are not mutually exclusive since we assume the coding algorithm does not have a way to anticipate which signal components reflect the impairment.

In order to evaluate the implications of coding impaired audio material, both impaired and unimpaired versions of the same audio material should be available. Since the "clean" versions of the impaired audio material are in general not available, it is necessary to synthesize impaired audio material from unimpaired audio material in order to compare the coding results of both versions. Such a study has been carried out in the context of an European RACE project (R2122 – Cougar) [Keyhl95]. Very detailed subjective tests are reported on MPEG-1 Audio Layer 2 and Layer 3 coded material [Keyhl95]. We will address here a sub-part of that study concerning the analysis, the modeling and the synthesis of typical signal impairments.

In Sub-section 7.3.2 we analyze impaired audio material in order to model a few types of typical signal impairments. In Sub-section 7.3.3 we use these models to filter and to contaminate unimpaired audio material.

As examples of impaired audio material, we used the sound-track from a 1944 movie ("Bluebeard") provided by the BBC, and the sound-track from a telecine documentary about the "Beatles" provided by the RTP (the Portuguese Public Television).

7.3.2 Analysis of Impaired Audio Material

Several forms of impairments have been found that were grouped into three categories:

- *background noise:* a form of stationary random noise that is modulated as a function of the audio signal activity,
- *scratches:* a form of band-limited non-stationary spikes that are perceived individually and whose magnitudes approach the limits of the dynamic range of the audio signal,
- *impulsive noise:* a burst of band-limited non-stationary spikes of moderate amplitude (relative to the dynamic range of the audio signal) that are not perceived individually, creating sometimes the auditory sensation of crepitation (crackling).

In the following, we will illustrate each category of impairment and we will address possible causes.

Background Noise The color of the background noise is the result of additive and multiplicative contributions of all the signal conditioning and processing phases along the audio production chain, from the point of origination (*e.g.* the acoustical conditions, the recording equipment) to the point of reproduction.

The spectral profile of the background noise is reasonably stable in moments of silence, *i.e.* where there is no relevant audio signal activity. Figure 7.11a shows the time and spectral representation of a short segment of background noise found in "Bluebeard". This particular

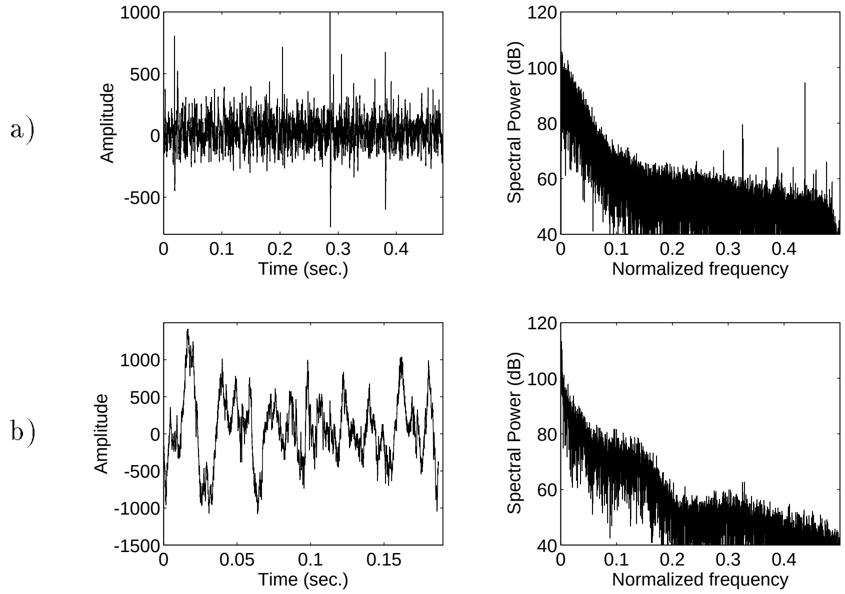

Figure 7.11: Time and spectral representation of two examples of background noise. a) the upper example was taken from the sound-track of a 1944 movie and b) the lower example was taken from the sound-track of a telecine documentary about the Beatles.

spectral profile can be reasonably well approximated by three straight lines as shown in Figure 7.12. The two strong spectral tones existing in the spectrum of the "Bluebeard" background noise, at about 15.6 kHz and 31 kHz, are believed to be the result of interference from the telecine line frequency and its second harmonic.

Figure 7.11b shows a short example of background noise taken from the "Beatles" documentary. The spectral profile has a more complex form which is probably related with the capture of sound in the conditions of a live concert.

A modulation of the background noise as a function of the audio signal activity is clearly perceived in "Bluebeard".

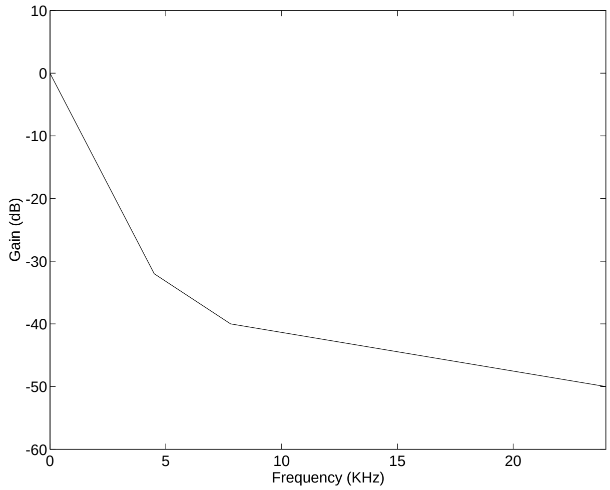

Figure 7.12: Simplified spectral model of the background noise found in the "Bluebeard" sound-track.

Scratches Scratches do not occur very frequently nor with a well defined regularity. Scratches appear to be to result of the introduction of a strong impulse in the production or post-production phase. Thus, their temporal shape corresponds to the impulse response of the processing chain down-stream after this insertion point, up to the point of reproduction. Neighbor scratches tend to have similar shapes which indicates they are probably introduced in the same point of the processing chain.

Different forms of scratches can be found and their intensity may approach the limits of the dynamic range of the sound programme. Usually, they have a duration in the order of 20 ms to 100 ms, which means they are perceived individually. Although normally they occur in isolation, sometimes two or more scratches occur very close to each other, as it is illustrated in Figure 7.13.

Impulsive noise While scratches are perceived individually, even when they occur in a burst sequence, impulsive noise corresponds to a regular succession of scratches, typically with lower intensities and shorter durations, but occurring much more frequently. As a consequence, they invoke the *fusion mechanism* of the human auditory system producing an auditory impression between crepitation and continuous noise. This is clearly related to the *precedence effect* (Section 3.4.3.1) that explains the perception of a sound consisting of a first signal occurrence followed by attenuated replicas or echoes. It is normally considered that a separation in the order of 50 ms is the threshold between the perception of different successive scratches and the perception of a "fused pattern".

Impulsive noise appears to be the result of other causes than those producing the scratches

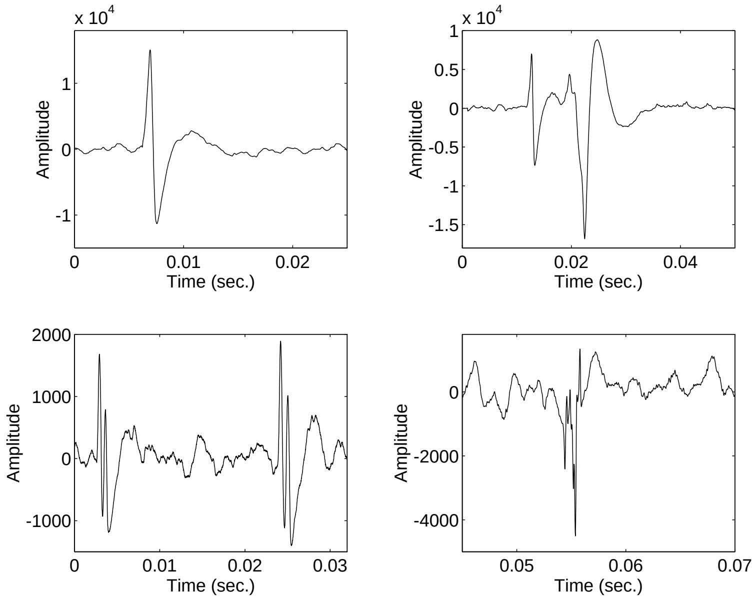

Figure 7.13: Example of scratches: isolated (upper left), close but separated (upper right and lower left), and merged (lower right).

described above. A deficient analog electrical connection can be such a cause. Figure 7.14

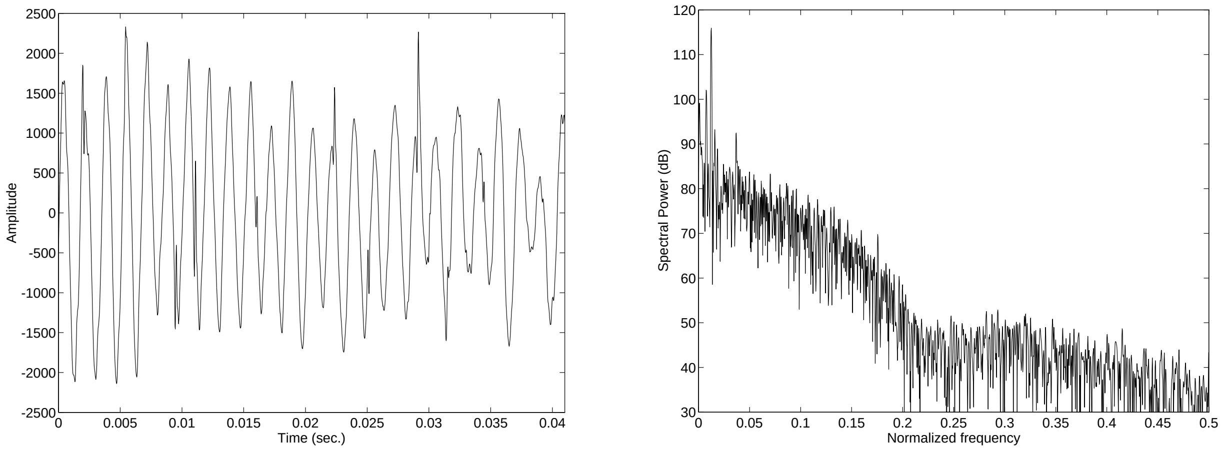

Figure 7.14: Time and spectral representation of a tonal signal contaminated by impulsive noise.

represents a short segment of a tonal signal contaminated by impulsive noise. In order to characterize statistically the small scratches, they should be isolated from the audio signal using for example predictive techniques.

7.3.3 Synthesis of Impaired Audio Material

Since different types of impairments can be identified as seen above, it is very convenient to be able to synthesize each type of impairment independently, in order to study isolated and combined effects. Figure 7.15 shows the modifications introduced in the ASC analysis/synthesis framework in order to inject background noise, scratches and impulsive noise into an arbitrary audio signal.

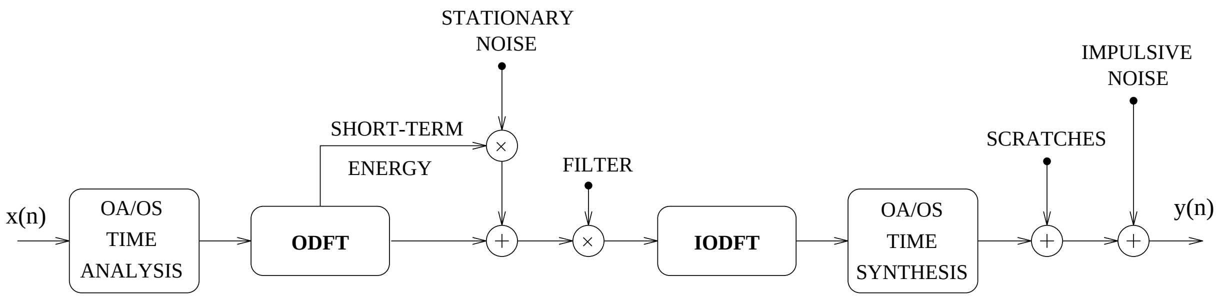

Figure 7.15: Adaptation of the ASC analysis/synthesis filter bank in order to synthesize impaired audio material from unimpaired audio material.

Background noise is added to the audio signal in the frequency domain. The magnitudes of the noise spectral components are randomized and are modulated by the short-term (11 ms) energy of the audio signal. For simplicity and in order to avoid unpleasant metallic-like synthesized noise, the phases of the noise components are made equal to zero. Filtering is also implemented in the frequency domain. The spectral model of the background noise is used to shape the synthesized noise and to filter the spectrum of the unimpaired audio signal.

Due to their non-stationary nature, scratches and impulsive noise are synthesized in the time domain. For simplicity, we add directly the available time patterns (*i.e.* the impulse responses) of scratches and impulsive noise to the partially impaired audio signal.

Good results are obtained using uniform or Gaussian distributions to modulate scratches in time and in amplitude. The sign of each scratch is randomly chosen in order to avoid the accumulation of a DC component. This is not, however, a serious concern because the frequency analysis of an average scratch pattern reveals that it has a “zero” at DC. This is to be expected since most audio circuits use capacitors or transformers to insure DC isolation.

In order to synthesize impulsive noise, a richer impulse response diversity is used. For simplicity, we choose randomly from a collection of 4 different (but similar particularly in terms of bandwidth) time patterns.

The synthesis of background noise plus scratches, and the synthesis of impulsive noise, are illustrated in Figure 7.16. The background noise is a synthesized approximation based on the

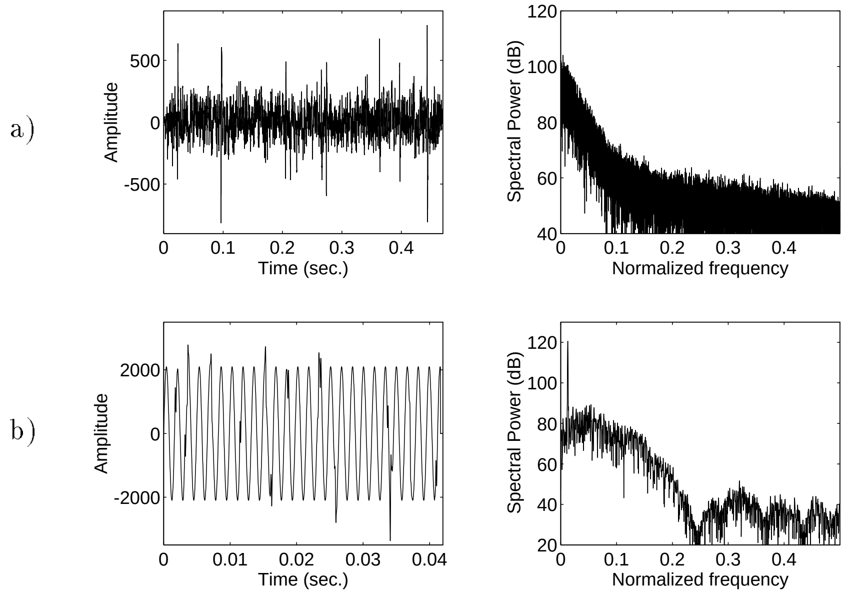

Figure 7.16: Examples of synthesized impairments: a) colored noise plus scratches (upper figure), b) impulsive noise added to a pure tone (lower figure).

model shown in Figure 7.12, to the spectral profile of Figure 7.11a. Scratches are then added according to a Gaussian time distribution and an uniform amplitude distribution. It can be seen that the time and spectral representation of the synthesized impairment are very similar to those shown in Figure 7.11a. This is consistent with the fact that both signals sound alike.

Figure 7.16b illustrates the synthesis of impulsive noise. A pure tone with a frequency equal to the main tone of the signal represented in Figure 7.14 was synthesized. Four representative impulse patterns were extracted from this signal, were modulated in time (with a Gaussian distribution) and amplitude (with a uniform distribution), and were finally added to the synthesized tone. The spectral profile of the synthesized signal (Figure 7.16b) mimics reasonably well the spectral profile of the impaired signal (Figure 7.14), which agrees with the fact that both types of impulsive noise sound alike.

An extension of the algorithm has been implemented to synthesize stereo impairments. Since stereo impaired audio material was not available for analysis and characterization, simple heuristic rules were used to create stereo impairments:

- In the case of background noise, the colored noise is first synthesized separately for each channel. Then, a system parameter forces any desired degree of correlation between the two channels.

- In the case of scratches and impulsive noise, amplitudes are modulated independently but the same Gaussian distribution model is used to define the timing of a scratch or spike in both channels. However, an additional Gaussian model is used to introduce a small delay between the disturbances in the two channels.

Formal listening tests involving unimpaired source material, source material impaired with background noise only, and source material impaired with both scratches and background noise, have indicated that:

- the coding quality is in general adversely affected by the presence of signal impairments,

- depending on the coding scheme, the presence of scratches (in addition to background noise) has the potential to cause significantly worse results than when only background noise is present.

It can be concluded that rather than improving the opportunities for masking the coding noise, an impaired audio signal stresses the operation of perceptual audio coders, leading in general a lower coding quality. As possible solutions to this problem, it has been recommended the removal of signal artifacts in a pre-processing stage prior to coding, or the provision for a safety margin in the coding process [Keyhl95].

7.4 Time Scale Expansion

7.4.1 Introduction

Time-scale modification of audio signals is desired, for example, in applications involving non-linear audio editing such as fast/slow forward or backward audio replay. Ideally, only the presentation rate of the audio material should be modified without affecting neither its intelligibility nor its quality. This implies that the pitch and the timbre coloration must be preserved, and that signal artifacts should not be audible. Time domain techniques, as well as frequency domain techniques, have been used to modify the time-scale of audio signals.

Time domain techniques preserve easily the time waveform of the signal, and hence the original pitch. In fact, the main principle of time domain techniques is the repetition or the discarding of suitable fragments of the original signal in order to perform a desired time-scale expansion or compression. When these operations are combined with decimation or interpolation, they raise or lower the pitch while keeping the original time-scale of the signal. In the case of speech specific signals, an additional requirement in pitch modification is that the spectral envelope (formant structure) should be maintained [Macon96]. The major difficulty in maintaining signal quality with time domain techniques consists in the elimination of artifacts due to discontinuities arising when concatenating time fragments of the original audio signal.

Frequency domain techniques involve the modification of a spectral representation of the audio signal, possibly with reference to a specific model of source production (as in the case of speech). Decimation and/or interpolation operations are then combined with the frequency-time transformation to synthesize the rate modified audio signal. Frequency domain techniques are in general computational intensive since usually they involve iterative procedures.

We will present a new frequency domain approach to time-scaling that is based on frequency-scale modification *in addition* to the magnitude and phase modifications of the spectral representation. Time-scaling is then obtained simply by controlling directly the sampling frequency of the Digital-to-Analog converter or, if this frequency can not be changed, by using an additional and independent interpolation/decimation processing module. Our approach was developed to satisfy specific and severe requirements:

1. the algorithm is constrained to share the ASC analysis/synthesis framework described in Sub-section 7.1, which is based on the 50% overlap Odd-DFT,

2. the algorithm should be non-iterative and *blind*, *i.e.* it should perform all relevant spectral manipulations on a single frame basis, without the need to look into the past[3] nor into the future.

As will be detailed later (Sub-section 7.4.3.4), the first functional limitation resulting from the above constraints is that only time-scale expansion is allowed. Secondly, only integer modification factors are possible. As a consequence, our algorithm focuses on applications which slow down the presentation rate of audio material in order to improve its intelligibility (*e.g.* of spoken sentences when one learns a foreign language), or in order to allow time discrimination of very rapid acoustic events.

This section is organized as follows. Sub-section 7.4.2 presents a short review of the most common time-scale modification techniques applied to PCM audio data, and introduces the main differences to the new technique. In Sub-section 7.4.3 we describe the main principles underlying our technique and we propose a new method to estimate the phase and the center frequency of an audio component beyond the frequency resolution of the analysis/synthesis filter bank. The main structure of the frequency-scaling and spectral modification algorithm is described in

[3] An obvious consequence of this requirement is that it precludes phase unwrapping.

Sub-section 7.4.4 and its performance is illustrated, with a few examples, in Sub-section 7.4.5.

7.4.2 A Review of Known Techniques

7.4.2.1 Time Domain Techniques

Time domain techniques are frequently used to time-scale an arbitrary audio signal, including speech, without assuming an explicit source production model. The main advantage of time domain techniques is the low algorithm complexity. Yet, for modest time-scale modifications, the quality of the synthesized signals has been independently reported to be very high [Verhel.93][Yim96][Lin95]. Given an input signal, the basic approach of time-domain techniques consists in the generation of a new signal by concatenating time frames taken from the original sound either in an overlapped fashion (corresponding to time-scale expansion), or in a discontinuous fashion (corresponding to time-scale compression). The concatenation is performed so as to avoid audible artifacts associated with amplitude or phase discontinuities. To prevent these artifacts, some small overlap between adjacent frames is usually enforced in which fading and alignment techniques are used. Therefore, under this class of algorithms, methods for time-scale modification differ mainly on the fading criterion, or on the correlation or similarity metrics underlying the alignment criterion.

As long as the time-scale modification is moderate [Moulin.95], the time waveform of the original signal, as well as its local spectral shape, are preserved in the synthesized signal. Therefore, the quality of the synthesized signal is close to the quality of the original signal.

7.4.2.2 Frequency Domain Techniques

Frequency domain techniques assume that an input signal can be decomposed as a sum of time varying sinusoidal components. Time-scaling is achieved by combining a suitable manipulation of the phase and magnitude of these components, with decimation and/or interpolation operations.

A well known frequency domain technique is based on the phase vocoder [Dolson86][Flanag.66]. This method first computes the short time Fourier transform (STFT) of the original signal. The "jump" in time samples between adjacent transformations is known as the analysis hop. A phase unwrapping procedure follows, allowing to obtain the phase increments between adjacent transformations for each spectral coefficient (by adding an integer multiple of 2π according to the wrapped phase). From this, the "instantaneous" frequencies of the main sinusoidal components of the signal are estimated. Time-scaling is obtained by changing the temporal evolution of the signal using a synthesis hop different from the analysis hop (their ratio specifies the time-scaling factor). Before the overlap-add of the (windowed) inverse STFTs, the phases of the main sinusoidal components of the signal are modified to account for the difference between the analysis hop and the synthesis hop or, in other words, to guarantee the "horizontal phase coherence". A recently submitted paper proposes a solution to improve the performance of the phase vocoder based time-scaling, by addressing the problem of "vertical phase coherence" [Laroch.97].

Due to the fact that frequency domain techniques deal simultaneously with the complex time and frequency structures of an arbitrary audio signal, quality and complexity reasons force a number of functional limitations. For example, the time-scaling algorithm might not address time compression and time expansion simultaneously, and it might be constrained to integer time-scaling factors. Quatieri and his colleagues [Quatie.95], for example, designed a frequency domain technique that extends the temporal envelope of an acoustic signal by a desired integer factor. They propose a 50% overlap-add scheme based on a 21-band perfect reconstruction filter bank. Time-scale expansion is achieved by interpolating the magnitude and the unwrapped phase of the signal associated to each filter. The phase is further scaled according to the time-expansion factor (in order to maintain the original frequency trajectory), and is corrected by an offset factor in order to maintain the phase relation among selected subband channels contributing the most to distinct signal events [Quatie.95]. A limitation results from this technique in that no more than two acoustic events are assumed to occur simultaneously. Furthermore, it suffers from a low spectral resolution.

In general, frequency domain techniques have a number of problems due to intrinsic limitations in time/frequency resolution, and due to the accumulation of phase and magnitude errors introduced in the spectral modification. Artifacts result when audio signals include strong non-stationarities and simultaneously occurring sounds, of a harmonic nature, with different pitches [Pielem.96]. In these cases, transient smearing, harmonic smearing and phase dispersion are noticeable [Quatie.95].

7.4.2.3 Time Scaling Using Speech Models

Most published algorithms for time-scale modification were specifically designed for speech signals [Verhel.93][Roucos85][Quatie.95][Portnof81], Moulines and Laroche [Moulin.95] present a detailed review. For example, Quatieri and McAulay [Quatie.92] describe a speech specific time-scale modification algorithm that relies on the sinusoidal model [Rabiner78] for speech analysis/synthesis. The articulation rate of voiced speech is modified by manipulating the speech duration on the basis of the pitch period. The magnitude and phase of the components of the sinusoidal model (sine-wave synthesizer) are modified so as to maintain the temporal structure of the original waveform (a feature referred to as *shape invariance*) in the new desired time-scale. The performance of the algorithm is dependent on an accurate pitch period estimation and when a pitch period does not clearly exist, for example during unvoiced regions of the speech signal, the articulation rate does not change at all or does not change as desired.

Explicit Fourier techniques have also been used to modify the time-scale of speech. One example is the algorithm named *least-squares error estimation from the modified short-time Fourier transform magnitude* (LSEE-MSTFTM) [Griffin84]. The central idea is to use a frequency domain analysis of the original and synthesized signal in order to iteratively generate a better synthesized rate-modified signal, in the sense of least squared error between the magnitudes of the original and synthesized spectra, at corresponding time instants. As this algorithm does not account for phase alignment between adjacent frames, its convergence to a high quality synthesized signal is very slow and, therefore, is computationally intensive. This disadvantage was eliminated in a modified version [Roucos85] of the algorithm that included phase in the similarity metric underlying the alignment. This phase alignment was based on time-domain correlation between successive windowed frames. As a consequence, the resulting *Synchronized Overlap-and-Add* algorithm (SOLA) produces a time-scale modification factor slightly different from the factor initially specified.

Another example of a speech specific algorithm performing time-scale expansion or compression and operating in the frequency domain, is described by Portnoff [Portnof81]. It also assumes the classical speech model [Rabiner78]. The parameters of this model, as well as the pitch contour, are interpolated/decimated (*i.e.* linearly time-scaled) leading to the desired time-scale modification of speech while maintaining, ideally, signal quality and speaker-dependent features. The speech model uses a linear time-varying filter to approximate the transmission characteristics of the vocal tract and the spectral characteristics of the glottal pulse. The central idea of the algorithm is that with an appropriate signal analysis and synthesis scheme based on the STFT, the coefficients of the STFT are related to the parameters of the speech model (either considered as a harmonic representation for voiced speech, or as a time-varying spectrum for unvoiced speech), such that their manipulation leads to the synthesis of rate-changed speech. This manipulation involves decimation or interpolation of the spectral representation and a non-linear phase modification along the time axis (not along the frequency axis).

Although it is claimed [Portnof81] that the algorithm produces high-quality time-scaled speech for a wide range of time scaling factors, it is also reported [Portnof81] that for time-scale expansion greater than 4:1, a reverberating effect becomes objectionable, and that for time-scale compression greater than 3:1, the synthesized speech becomes rough. In general, the algorithm is quite complex and computationally intensive, primarily because it does not process critically sampled data (*i.e.* at the Nyquist rate). On the other hand, phase unwrapping [Tribol.77] is needed which implies a signal activity analysis in order to identify the initial time after which the phase of a spectral coefficient can be unwrapped. Furthermore, the spectral interpolation/decimation operation associated to the time-scaling of the STFT coefficients, makes use of anti-imaging/aliasing filters that must be redesigned for each new selection of the decimation/interpolation factors defining the rational time-scaling change.

7.4.2.4 Relation to the New Technique

As with other frequency domain techniques, our technique also assumes the audio signal can be represented as a sum of sinusoids. However, instead of a simple stationarity assumption, our technique also takes into account explicit temporal modulation effects.

All of the relevant processing is performed in the original time-scale of the signal. Time-scaling is implemented through a spectral expansion[4] of the spectral coefficients modified in amplitude *and* phase. After the frequency-time transformation and the overlap-add reconstruction, a final interpolation operation in the time domain leads to a time expanded signal that preserves the original pitch, timbre and formant structure. Another innovation in our technique is a new method for accurate phase *and* frequency estimation beyond the frequency resolution of the filter bank.

Our time-scale expansion technique does not assume any particular source production model. Instead, we rely on an accurate spectral manipulation that takes into consideration the specific nature of the filter bank and the shape of the analysis/synthesis time windows[5]. As a result, our spectral expansion approach insures that speaker dependent features, such as pitch and formant structure, are preserved in the time-expanded audio signals, provided that the filter bank has enough frequency resolution to resolve the reciprocal of the pitch period. Our algorithm processes critically sampled data, has a much lower complexity than other frequency domain schemes including that proposed by Portnoff [Portnof81], and elicits an entirely new approach: Instead of focusing on time interpolation/decimation of the spectral representation, it focuses on frequency expansion/compression of the spectral representation.

[4] Spectral compression, and hence time-scale compression, is not possible because, due to design constraints, phase unwrapping can not be implemented.

[5] For simplicity, we assume here that the analysis/synthesis time windows do not change dynamically.

7.4.3 The New Technique

7.4.3.1 The Central Idea

The objective of the algorithm is to extend the temporal envelope of a quasi-stationary signal by a desired integer factor, while preserving its spectral structure. This is achieved through a suitable spectral manipulation along the frequency axis (not along the time axis as in other algorithms [Portnof81][Quatie.95]). This objective is implemented in three steps, as illustrated in Figure 7.17. The first step encompasses the most relevant processing, the remaining two are

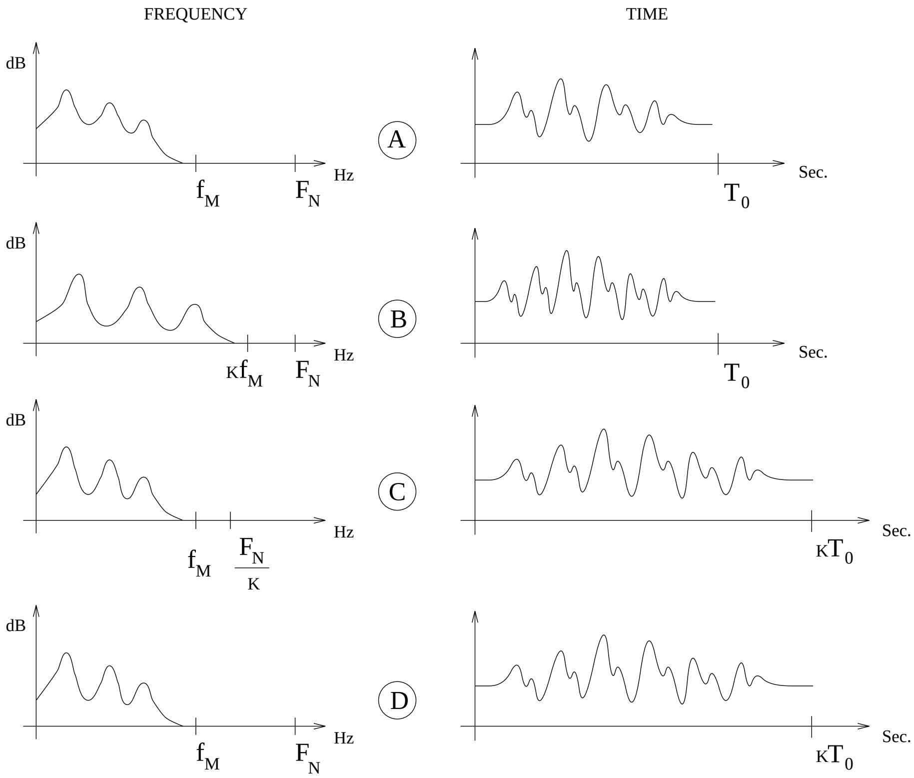

Figure 7.17: Illustrative representation of the new frequency domain approach to time-scale expansion. From top to bottom: (A) original signal, (B) spectral modification, (C) same as B with sampling frequency $\frac{2}{K}F_N$ instead of $2F_N$, (D) same as C after interpolation by K.

trivial operations.

The signal represented in A is assumed to have its significant spectral content limited to f_M. The Nyquist frequency is represented as F_N. The first step consists of a spectral modification (in amplitude and phase), followed by expansion along the frequency axis by the desired time expansion factor, K, as illustrated in B. This spectral expansion is a delicate operation because the frequency and phase relationships among the most relevant signal components must be preserved in order to maintain the fine temporal structure of the signal. It should be noted that this spectral expansion does not correspond, for several reasons, to the frequency domain resampling technique described by Moulines [Moulin.95], one of the reasons being that we preserve the bandwidth of each spectral component during the frequency expansion process. On the other hand, it must be emphasized that a clear difficulty arises, in particular, in shifting narrow-band tones in frequency due to the fact that the frequency of the tone can be any real number below f_M while the discrete transform only gives access to a finite number of sampled frequencies. As it will be shown in Sub-section 7.4.3.5, this problem has been circumvented by exploiting special properties associated to the Odd-DFT filter bank when combined with the chosen analysis/synthesis time window (Sub-section 7.4.3.3).

If the sampling frequency of the signal is maintained after this processing step as $2F_N$, *i.e.* if the time-scale of the signal is unchanged, then the "pitch" of the signal is raised[6] by K. If, as represented in C, the sampling frequency of the signal is simply changed to $\frac{2}{K}F_N$, then the pitch is maintained and the time-scale is expanded by K. This represents the second processing step (trivial resampling). In order to recover the original sampling frequency, the signal must be interpolated by K, as represented in D.

The sequence of steps illustrated in Figure 7.17 denotes an obvious implication: The signal bandwidth is reduced by the time-scale expansion factor. This problem is eliminated if the interpolation operation (step D) is performed before the frequency expansion operation (step B).

A similar approach may be used to perform time-scale compression in case phase unwrapping is available. In this case, B is a spectral compression operation which lowers the "pitch" while maintaining the time-scale. In order to compress the time-scale while keeping the original pitch, the processing step C changes the new sampling frequency of the signal to $2KF_N$, and then D decimates the signal by K. It must be noticed however, as pointed out by Portnoff [Portnof81], that while time-scale expansion by an arbitrarily high factor is physically feasible, time-scale compression even by a modest factor might destroy the information contents of the signal and its corresponding intelligibility. This problem is easily illustrated with a voiced sound since time compression is meaningless if it leads to a signal with a single pitch period or a fraction of it.

7.4.3.2 The Research Context and Related Constraints

In order to ease bit-stream access, editability and error robustness, ASC does not make use of interframe coding (Chapter 6). All processing is made on a frame-by-frame basis, thus avoiding dependencies between successive frames, a feature that reduces the coding/processing delay. This

[6] To be exact, pitch raising in this context also includes a formant shift.

constraint makes phase unwrapping difficult or even impossible since a detailed signal analysis along the time axis is not available. A direct consequence, as will be analytically demonstrated in Sub-section 7.4.3.4, is that only time-scale expansion (and not time-scale compression) by an integer factor is allowed. Nevertheless, we will show that a spectral modification based only on the phase and magnitude information available from a 50% overlap[7] analysis/synthesis scheme, is sufficient to perform the desired spectral expansion while maintaining the temporal envelope of the signal. When this processing is combined with a suitable control of the clock driving the Digital-to-Analogue converter or, if the sampling frequency is maintained, with external time interpolation, then this spectral modification equates to time-scale expansion.

7.4.3.3 Properties of the Odd-DFT Based Analysis/Synthesis System

The Filter Bank In order to implement time-scale expansion, we exploit the processing potential afforded by the 50% overlap analysis/synthesis framework based on Equation (7.1), Equation (7.2), and Equation (6.3). This means that the ability to analyze, to manipulate and to synthesize an arbitrary audio signal depends on the combined properties of the Odd-DFT and analysis/synthesis time window. As we have seen in Section 4.5, the analysis/synthesis time window should provide:

- perfect reconstruction in the absence of spectral modification,
- good spectral selectivity,
- good attenuation of time aliasing effects due to spectral sub-sampling and filtering.

Although optimized windows can be found as we have shown in Section 4.5.6, an additional requirement here is that the analysis/synthesis window should be analytically tractable in order to ease the spectral analysis and modification processes.

A simple and yet satisfactory solution, given the above requirements can be derived from the Hanning window as explained in the next sub-section.

The Analysis/Synthesis Window The $N+1$ point Hanning window is symmetrical with both first and last coefficients equal to zero:

$$hann(n) = \frac{1}{2}\left[1 - \cos\frac{2\pi}{N}n\right] , \qquad 0 \leq n \leq N . \tag{7.6}$$

[7]This is relative to the length of the analysis/synthesis time window.

A symmetrical N point Hanning window with all coefficients different from zero is given by:

$$hann_{mod}(n) = \frac{1}{2}\left[1 - \cos\frac{2\pi}{N}(n+\frac{1}{2})\right] = h^2(n)\ , \qquad 0 \le n \le N-1\ . \tag{7.7}$$

This result can be seen as the component-to-component product between two identical and symmetrical windows expressed by a simple trigonometric function:

$$h(n) = \sin\frac{\pi}{N}(n+\frac{1}{2})\ , \qquad 0 \le n \le N-1\ . \tag{7.8}$$

This solution has already been encountered before (Equation (4.23), (4.43), and (4.62)) because it is a simple solution satisfying the perfect reconstruction requirement. This window is frequently used in audio coding because it also exhibits a good spectral selectivity.

The time window given by Equation (7.8) possesses an additional advantage that is very important here: it is analytically tractable. In fact, the frequency response of $h(n)$ is easily obtained:

$$|H(w)| = \frac{\cos\frac{N}{2}w}{2}\sqrt{\left[\frac{1}{\sin\frac{1}{2}\left(\frac{\pi}{N}-w\right)} + \frac{1}{\sin\frac{1}{2}\left(\frac{\pi}{N}+w\right)}\right]^2}\ , \tag{7.9}$$

revealing that the bandwidth of the main lobe (pass-band) is $\frac{6\pi}{N}$ and that the normalized magnitude in dB of the envelope of its frequency response, for $|w| > \frac{3\pi}{N}$, is given by:

$$|\widehat{H(w)}|_{dB} = 20\log\left|\frac{1}{\sin\frac{1}{2}\left(\frac{\pi}{N}-w\right)} + \frac{1}{\sin\frac{1}{2}\left(\frac{\pi}{N}+w\right)}\right| + 20\log\left(\frac{1}{2}\sin\frac{\pi}{2N}\right)\ . \tag{7.10}$$

This result will be used in Sub-section 7.4.3.5 to derive a model for accurate frequency estimation.

Time-Frequency Transformation of Special Signals When the analysis/synthesis window is combined with the Odd-DFT filter bank, a number of interesting results are obtained. For example, any DC component of a real signal, $x(n)$, is concentrated on a single spectral line in the set of *unique* Odd-DFT coefficients (*i.e.* for $0 \le k < \frac{N}{2}$). In fact, when $x(n) = A$ and when $h(n)$ is as in Equation (7.8), the coefficients of the Odd-DFT are given by

$$X(k) = N\frac{A}{2}\left\{e^{-j\frac{\pi}{2}(1-\frac{1}{N})}\delta(k) + e^{j\frac{\pi}{2}(1-\frac{1}{N})}\delta(k+1-N)\right\}\ , \tag{7.11}$$

which shows that the single non-zero coefficient in the range of unique coefficients, occurs when $k = 0$. $\delta()$ is the Dirac delta function.

Another interesting effect takes place when $x(n)$ is a pure tone whose frequency is such that, due to the contribution of $h(n)$ (see below), it corresponds to a sampled frequency of the

Odd-DFT, $\ell\frac{2\pi}{N}$, with ℓ in the range $[1,\ \frac{N}{2}-1]$. In fact, if $x(n)$ is given by

$$x(n) = A\sin\left(\frac{2\pi}{N}\ell n + \phi\right) \tag{7.12}$$

then,

$$X(k) = \frac{NA}{4}\left\{e^{-j\left(\frac{\pi}{2N}-\phi\right)}\delta(k-\ell+1) + e^{-j\left(\pi-\frac{\pi}{2N}-\phi\right)}\delta(k-l)\right\}\ , \qquad 0 \le k \le \tfrac{N}{2}-1\ . \tag{7.13}$$

The meaning of this result is that in the set of unique Odd-DFT coefficients, an exactly sampled pure tone is mapped to two consecutive spectral lines. As explained elsewhere [Ferre.95b], this is the result of a (suppressed carrier) double-side-band modulation effect caused by the sine window $h(n)$. Further analysis of the phase of these spectral lines reveals that the fixed phase, ϕ, in Equation (7.12), is available from the phase of the spectral line $k = \ell - 1$ after adding $\frac{\pi}{2N}$. Moreover, the difference between the phase at $k = \ell$ and the phase at $k = \ell - 1$ is exactly $\pi(\frac{1}{N} - 1)$, regardless of ϕ.

If the frequency of a pure tone is not exactly sampled by the analysis filter bank (*i.e.* the sine window $h(n)$ combined with the Odd-DFT), then the energy of $x(n)$ will spread to more than two spectral lines of the Odd-DFT, but the above conclusions still hold very approximately if an additional phase offset is considered. In fact, if $x(n)$ is given by

$$x(n) = A\sin\left(\frac{2\pi}{N}(\ell + \Delta\ell)n + \phi\right)\ , \tag{7.14}$$

where ℓ is an integer number in the range $1 \le \ell \le \frac{N}{2} - 1$, and $\Delta\ell$ is a real number in the range $0.0 \le \Delta\ell < 1.0$, then the spectral line whose index is $k = \ell$, is a local maximum, and the fixed phase ϕ in (7.14) can be recovered from the phase at $k = \ell - 1$ by adding $\frac{\pi}{2N}$ and a convenient offset. Figure 7.18 illustrates a few representative cases where $N = 256$, $\ell = 12$, $\phi = 0.0$ and $\Delta\ell$ varies from 0.05 to 0.95.

Tests with synthesized tones led to the conclusion that the offset can be approximated by $\Delta\ell\pi$. After comparing the offset estimated by this simple model to the exact offset of synthesized tones, we concluded that the estimation error varies with N and, to a much less extent, with the index of the spectral line. For example, if N is 1024, ℓ is 12 and $\Delta\ell$ varies across the frequency resolution of the transform ($\frac{2\pi}{N}$) in small steps (200 points), the maximum relative error is less than 0.18%. As this offset estimator depends on $\Delta\ell$, this means that an accurate estimation of the fractional frequency $\Delta\ell$ is also necessary in order to compute a reliable estimation of the fixed phase ϕ. In Sub-section 7.4.3.5 we propose a new method to estimate $\Delta\ell$.

A final relevant observation deals with time modulated tones. For simplicity, let us consider

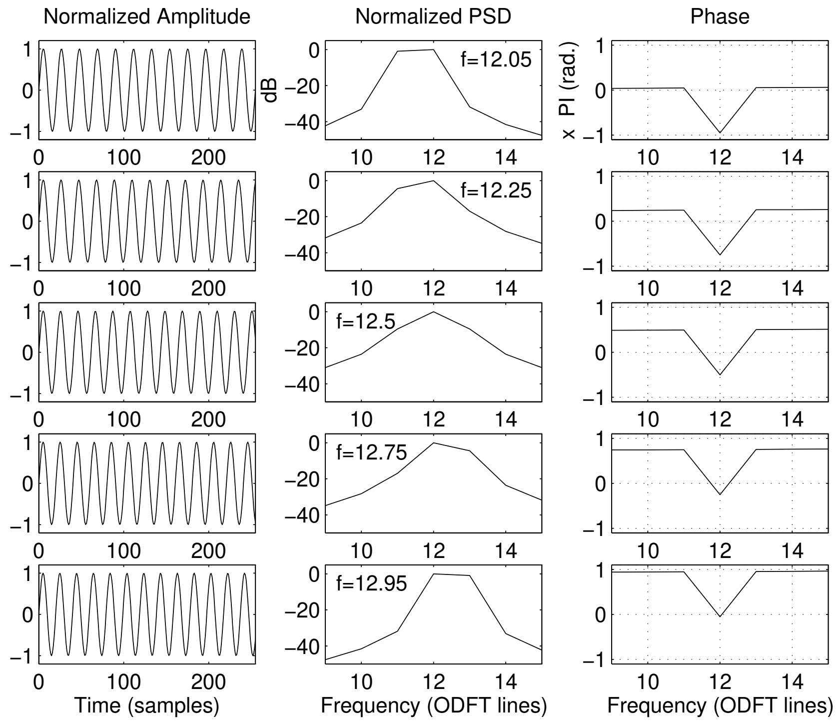

Figure 7.18: Phase offset when $\Delta\ell$ varies from 0.05 (top) to 0.95 (bottom). N is 256 and ℓ is 12.

a tone modulated by a real exponential:

$$x(n) = Ae^{-\frac{\alpha}{N}(n)} \sin\left(\frac{2\pi}{N}(\ell + \Delta\ell)n + \phi\right) . \tag{7.15}$$

Several tests have indicated that the most relevant implication of the time modulation is a phase modification around the center frequency of the tone as illustrated in Figure 7.19. In this figure, a non-modulated tone with $N = 256$, $\Delta\ell = 0.5$ and $\ell = 12$ is compared to its modulated version when $\alpha = 2.0$. The spectral magnitude is only slightly affected but the symmetry of the strongest spectral coefficients is maintained which is consistent with the fact that in both situations $\Delta\ell = 0.5$. In particular, if ℓ is sufficiently different from either 0 or $\frac{N}{2} - 1$, then neither the spectral magnitude nor the phase are modified significantly as a function of the frequency of the tone, for the same temporal envelope and $\Delta\ell$. This is also illustrated in Figure 7.19 for tones with $\ell = 12$ and $\ell = 24$.

Equation (7.16) gives the Odd-DFT spectral coefficients in the range $[1,\ \frac{N}{2} - 1]$ (thus ignoring the contributions existing in the complementary spectral range), when $x(n)$ is as in Equation (7.15). This expression corroborates the above observations and also predicts that the phase modification is dependent on the sign of α and is more accentuated for higher values of $|\alpha|$ and

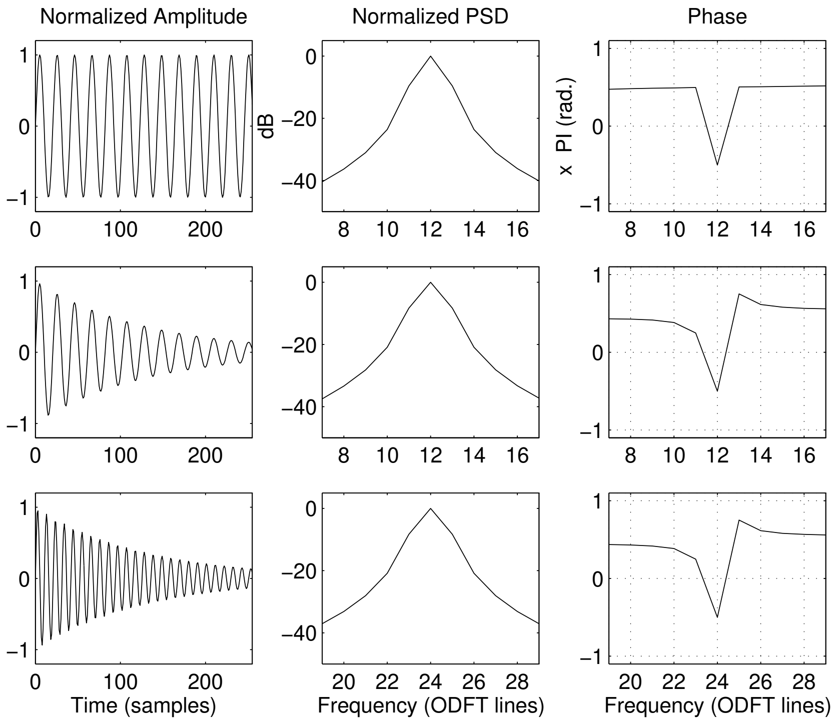

Figure 7.19: Spectral modifications when a tone is modulated in time by a real exponential. In all cases N is 256 and $\Delta\ell = 0.5$. From top to bottom: $\ell = 12$ and $\alpha = 0.0$, $\ell = 12$ and $\alpha = 2.0$, $\ell = 24$ and $\alpha = 2.0$.

$|\Delta\ell - \frac{1}{2}|$.

$$X(k) = \frac{A}{4}\sqrt{1 - 2e^{-\alpha}\cos(2\pi\Delta\ell) + e^{-2\alpha}}\; e^{j\left(\phi - \arctan\frac{e^{-\alpha} sin(2\pi\Delta\ell)}{1-e^{-\alpha}\cos(2\pi\Delta\ell)}\right)} \times \left\{ \frac{e^{-j\left(\frac{\pi}{2N} + \arctan\frac{e^{-\frac{\alpha}{N}} sin\frac{2\pi}{N}(k-\ell-\Delta\ell+1)}{1-e^{-\frac{\alpha}{N}}\cos\frac{2\pi}{N}(k-\ell-\Delta\ell+1)}\right)}}{\sqrt{1 - 2e^{-\frac{\alpha}{N}}\cos\frac{2\pi}{N}(k-\ell-\Delta\ell+1) + 2e^{-\frac{2\alpha}{N}}}} + \frac{e^{j\left(\frac{\pi}{2N} - \pi - \arctan\frac{e^{-\frac{\alpha}{N}} sin\frac{2\pi}{N}(k-\ell-\Delta\ell)}{1-e^{-\frac{\alpha}{N}}\cos\frac{2\pi}{N}(k-\ell-\Delta\ell)}\right)}}{\sqrt{1 - 2e^{-\frac{\alpha}{N}}\cos\frac{2\pi}{N}(k-\ell-\Delta\ell) + 2e^{-\frac{2\alpha}{N}}}} \right\} \tag{7.16}$$

As long as the temporal modulation is relatively smooth compared to the duration of the analysis/synthesis window, the conclusions derived above for exponentially modulated tones, also apply to signals with more complex time modulation patterns. As a result, the temporal envelope of a frequency shifted tone is approximately maintained by keeping the relative phase relationships among neighboring spectral coefficients. In Sub-section 7.4.4 this approach will also be extended to signals whose bandwidth is comparable to, or wider than the bandwidth of the analysis window main lobe.

7.4.3.4 Constraints on the Frequency Expansion

Given the system constraints introduced in Sub-section 7.4.3.2, a question arises concerning the possibility of an arbitrary frequency expansion factor. If the frequency of a stationary input signal is f_a and the frequency of its shifted version is f_b then, it becomes obvious that their ratio also relates $\ell_a + \Delta\ell_a$ with $\ell_b + \Delta\ell_b$, *i.e.*

$$\frac{f_b}{f_a} = \frac{\ell_b + \Delta\ell_b}{\ell_a + \Delta\ell_a} = \beta. \tag{7.17}$$

On the other hand, recognizing that the fixed phase ϕ_a can be expressed as a function of a global time offset n_0:

$$\phi_a = \frac{2\pi}{N}(\ell_a + \Delta\ell_a)n_0 \tag{7.18}$$

then, it is a simple matter to conclude that $\phi_b = \beta\phi_a$. This means that the fixed phase must be modified by the same frequency expansion factor. This condition is essential to preserve the fine time structure of a complex multitone signal within each time block m.

Furthermore, given that the phase is wrapped, the fixed phase of both input and modified signals at a given time block m are reduced to the fundamental interval $[-\pi, \pi]$. If the signals are quasi-stationary across two consecutive time blocks then, phase continuity between blocks must be enforced. These conditions may be condensed as follows:

$$\beta\left\{\frac{2\pi}{N}(\ell_a + \Delta\ell_a)(m\frac{N}{2} + n_0) - 2\pi r_a\right\} = \frac{2\pi}{N}(\ell_b + \Delta\ell_b)(m\frac{N}{2} + n_0) - 2\pi r_b\ , \tag{7.19}$$

where r_a and r_b are integer numbers. However, since we do not have access to the term[8] $2\pi r_a$ in Equation (7.19), phase continuity can only be guaranteed if βr_a is an integer number. It follows that as r_a depends on the frequency, β must be an integer number. The most obvious consequence of this conclusion is that the system constraints eliminate the possibility of down-shift in frequency (*i.e.* $\beta < 1$ meaning spectral compression) and only up-shift (spectral expansion) by integer factors are allowed. In other words, only time-scale expansion by integer factors are permitted.

7.4.3.5 Estimating the Fractional Frequency $\Delta\ell$

As a multiple integer of a frequency given by a real number is also a real number, then $\Delta\ell_a$, the fractional frequency of a signal relative to the frequency resolution of the analysis filter bank,

[8] In fact, this information is available only if phase unwrapping is implemented.

as considered in Equation (7.17), must be accurately estimated so as to implement a coherent spectral expansion of *all* components of that signal. Failure to reach this goal will result in frequency and phase discontinuities. In order to derive this fractional frequency estimator[9], we will first find a practical model of the analysis/synthesis window frequency response.

As the analysis window (7.8) defines the transfer function of each spectral coefficient of the Odd-DFT, the spectral representation of a narrow-band input tone corresponds to samples at frequencies $(k+\frac{1}{2})\frac{2\pi}{N}$, with $0 \leq k \leq \frac{N}{2}-1$, of the frequency response of the analysis window (7.9) centered around the frequency of the tone. Its main lobe (pass-band) spans three spectral coefficients across its bandwidth ($\frac{6\pi}{N}$) for a non-exactly sampled tone ($\Delta\ell \neq 0$) and two spectral coefficients for an exactly sampled tone ($\Delta\ell = 0$). This conclusion suggests the possibility of estimating $\Delta\ell$ by relating the magnitudes of the three spectral coefficients within the pass-band. An existing smooth time modulation of the narrow-band input tone does not alter significantly this relationship as seen previously (end of Sub-section 7.4.3.3).

A simplified model of the frequency response of the analysis window can be derived from Equation (7.9) and from Equation (7.10). The discontinuity existing in Equation (7.9) within the passband can be circumvented by using a simple model for the entire main lobe:

$$|\widehat{H(w)}|_{dB} = G \log\left[\cos\frac{N}{6}(w-w_0)\right] \ , \tag{7.20}$$

for $|w-w_0| < \frac{3\pi}{N}$. G is a gain parameter and was set to 30 to approximate the main lobe of the analysis window. As the main feature of interest in the stopband is its envelope, then Equation (7.10) can be used for the stopband model. Figure 7.20 compares this composite model to the exact frequency response of the analysis/synthesis window.

If the frequency of the input tone is $w_0 = \frac{2\pi}{N}(\ell+\Delta\ell)$, then the spectral coefficient at $k=\ell$ will be a local maximum. The frequencies at $k=\ell-1$ and $k=\ell+1$ which fall within the passband, are respectively $w_{\ell-1} = \frac{2\pi}{N}(\ell-\frac{1}{2})$ and $w_{\ell+1} = \frac{2\pi}{N}(\ell+\frac{3}{2})$. Using Equation (7.20), $\Delta\ell$ can be expressed as a function of the attenuation at these frequencies:

$$\Delta\ell = \frac{3}{\pi}\arctan\frac{\sqrt{3}}{2}\frac{1}{\frac{1}{2}+10^{\frac{X(\ell-1)_{dB}-X(\ell+1)_{dB}}{G}}} \ . \tag{7.21}$$

This equation was used to estimate the fractional frequency $\Delta\ell$ when $N=1024$ and $\ell=12$. The results are presented in Figure 7.21 where the estimated $\Delta\ell$ is compared to the exact fractional frequency of 200 synthesized tones. G has been adjusted to 28 instead of 30 in order to reduce the asymmetry of the estimation error. Another possibility to improve the $\Delta\ell$ estimate using

[9] It should be noted that the specific combination of the chosen analysis/synthesis time window with the Odd-DFT filter bank, makes other methods of fractional frequency estimation such as the "quadratic fit" or the "phase vocoder" [Brown96], inappropriate or less accurate.

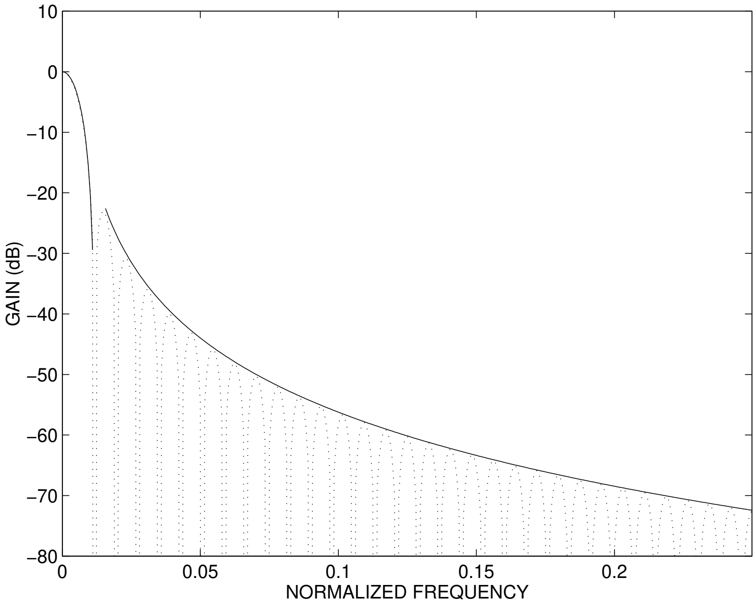

Figure 7.20: Passband and stopband approximations (solid line) to the frequency response of the analysis/synthesis window (dotted line).

Equation (7.21) consists in adjusting the value of G according to the actual data, as will be shown in Sub-section 7.4.4.

7.4.4 Spectral Expansion of Narrow and Wideband Signal Components

In this sub-section, we describe a procedure to implement frequency expansion of all the components of an audio signal. Both narrow-band and wideband components are subjected to magnitude and phase changes in order to preserve the temporal envelope in the original time-scale, and to preserve the pitch and the fine time structure in the new time-scale, as illustrated in Figure 7.17. Because the bandwidth of each shifted component is maintained (see below), broadband signals such as tonal pulses are also taken into account, although the overlap-add procedure, using a fixed window length has an inherent limitation regarding time resolution.

The major algorithmic steps are summarized below.

1. Following the time-frequency transformation, all spectral peaks are found using a peak detector similar to the one used in the Psychoacoustic Model 1 of the MPEG-Audio Standard [ISO93]. These peaks are further decimated in order to avoid low level variations and in order to retain the most *relevant* signal components. A decimation procedure similar to that presented in Section 6.2.5.2 is used. The location of each spectral peak (local maximum) as well as the location of the center of each spectral valley (local minimum)

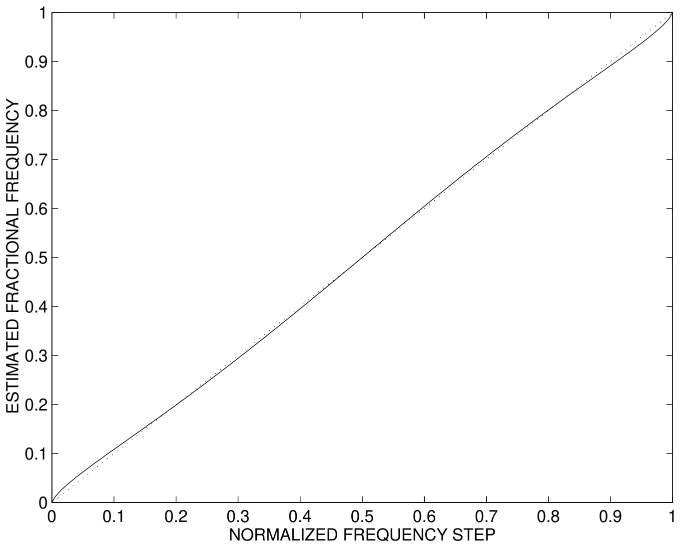

Figure 7.21: Estimated fractional frequency (solid line) as a function of the exact synthesized fractional frequency (dotted line) for 200 tones. $N = 1024$ and $\ell = 12$.

are stored.

2. The center frequency of each local maximum is computed using a two step approach. As ℓ is known, $\Delta\ell$ is first estimated using Equation (7.21) with $G = 28dB$. The value of G is then adjusted to

$$G = \begin{cases} \dfrac{X(\ell)_{dB} - X(\ell-1)_{dB}}{\log\left[\cos\frac{\pi}{6}(2\Delta\ell+1)\right]}, & \text{if } \Delta\ell > 0.5\,, \\[2ex] \dfrac{X(\ell)_{dB} - X(\ell+1)_{dB}}{\log\left[\cos\frac{\pi}{6}(3-2\Delta\ell)\right]}, & \text{otherwise}\,. \end{cases} \tag{7.22}$$

A better estimate of $\Delta\ell$ is obtained by recomputing (7.21) with the new value of G.

3. As the time expansion factor, β, is specified and ℓ_a and $\Delta\ell_a$ are known, the new frequency position ℓ_b as well as $\Delta\ell_b$, are obtained using Equation (7.17).

4. The bandwidth of each spectral component is found using

$$BW = \begin{cases} \max\left(6.0, \dfrac{\pi(2\Delta\ell+1)}{\arccos\left(10^{\frac{X(\ell-1)_{dB}-X(\ell)_{dB}}{30}}\right)}\right), & \text{if } \Delta\ell > 0.5\,, \\[3ex] \max\left(6.0, \dfrac{\pi(3-2\Delta\ell)}{\arccos\left(10^{\frac{X(\ell+1)_{dB}-X(\ell)_{dB}}{30}}\right)}\right), & \text{otherwise}\,. \end{cases} \tag{7.23}$$

These expressions were derived from Equation (7.20).

5. The Power Spectral Density at the center frequency $(\ell + \Delta\ell)$ of each signal component is estimated using its bandwidth, as obtained previously, at the closest discrete frequency $(k = \ell)$, as follows:
$$PSD_{MAX} = X(\ell_a)_{dB} - 30\log\left[\cos\frac{\pi}{BW}(2\Delta\ell_a - 1)\right] . \tag{7.24}$$

6. The fixed phase, ϕ, for each signal component is estimated by removing the constant phase contributions identified in Sub-section 7.4.3.3:
$$\phi_a = \phi(\ell_a - 1) + \pi(\frac{2}{N} - \Delta\ell_a) . \tag{7.25}$$
This estimation assumes that any time modulation is sufficiently smooth so that, as a consequence, the phase at $k = \ell_a - 1$, *i.e.* $\phi(\ell_a - 1)$, is not modified substantially.

7. New magnitudes are synthesized at $k = \ell_b - 1$, $k = \ell_b$ and at $k = \ell_b + 1$ as given by Equations (7.26), (7.27) and (7.28) respectively.
$$\hat{X}(\ell_b - 1)_{dB} = PSD_{MAX} + 30\log\left[\cos\frac{\pi}{BW}(2\Delta\ell_b + 1)\right] \tag{7.26}$$
$$\hat{X}(\ell_b)_{dB} = PSD_{MAX} + 30\log\left[\cos\frac{\pi}{BW}(2\Delta\ell_b - 1)\right] \tag{7.27}$$
$$\hat{X}(\ell_b + 1)_{dB} = PSD_{MAX} + 30\log\left[\cos\frac{\pi}{BW}(2\Delta\ell_b - 3)\right] \tag{7.28}$$
The new magnitudes from $k_b = \ell_b + 2$ to the next local minimum on the right are synthesized by maintaining the original differential magnitude in dB:
$$\hat{X}(k_b) = X(k_a) + \hat{X}(\ell_b + 1) - X(\ell_a + 1) .$$
Similarly, the new magnitudes from $k_b = \ell_b - 2$ to the next local minimum on the left are synthesized using the original differential magnitude in dB:
$$\hat{X}(k_b) = X(k_a) + \hat{X}(\ell_b - 1) - X(\ell_a - 1) .$$
The magnitude of each local minimum is further modified using Equation (7.10) and the spectral distance to one of the two surrounding local maxima, whichever produces a predominant effect. This avoids the generation of signal artifacts due to components with significant energy in spectral valleys.

8. The phase of each frequency expanded component is synthesized by first reconstructing the phase at $k = \ell_b - 1$ as follows:
$$\hat{\phi}(\ell_b - 1) = \beta\phi_a - \pi(\frac{2}{N} - \Delta\ell_b) . \tag{7.29}$$

The phases from $k_b = \ell_b$ to the next local minimum on the right, and from $k = \ell_b - 2$ to the next local minimum on the left, are synthesized by maintaining the original differential phases:

$$\hat{\phi}(k_b) = \phi(k_a) + \hat{\phi}(\ell_b - 1) - \phi(\ell_a - 1).$$

9. The above spectral expansion and modification process generates holes which are likely to cause signal artifacts. The problem is attenuated if these holes are filled with coefficients whose magnitude and phase are interpolated. We used a function presented in Section 6.2.5.2 and elsewhere [Ferre.95c] to interpolate the magnitude and phase differences across all tr coefficients of the spectral hole. The expression of this function is

$$\sqrt{1 - \cos\left(\pi \frac{i+1}{tr+1}\right)}\ , \quad 0 \leq i \leq tr - 1\ ,$$

for the left-hand side of a local minimum, and

$$\sqrt{1 - \cos\left(\pi \frac{tr-i}{tr+1}\right)}\ , \quad 0 \leq i \leq tr - 1\ ,$$

for the right-hand side of a local minimum. The interpolation of phases was preferred to a simple randomization in order to avoid critical phase transitions, in the order of $-\pi$, as seen in Sub-section 7.4.3.3.

7.4.5 A Few Results

Our time-scaling algorithm is based on a frequency expansion of spectral coefficients modified in amplitude and phase, as illustrated in Figure 7.17, step Ⓑ. This is where the most important processing takes place. The final stages of resampling and time interpolation are straightforward, are independent of the main algorithm, and should *not* affect the quality of the signal. Therefore, we can make an objective assessment of the quality of the whole spectral modification by comparing the output of the algorithm at step Ⓑ, with the ideal output signal. The latter can be easily generated by synthesizing the desired spectral modification subjected to the time envelope of the original signal.

Using this approach, we tested our algorithm with stationary harmonic signals, time modulated tones, broadband signals (tonal pulses), and a sine sweep. Real world audio signals were also used in a subjective quality evaluation. Instead of peaceful synthetic test signals, we use tough ones in order to understand the response of the algorithm to difficult conditions. We will first describe the response of the algorithm to three different test signals, when the frequency expansion factor is 2, the sampling frequency is 48 kHz, and $N = 1024$.

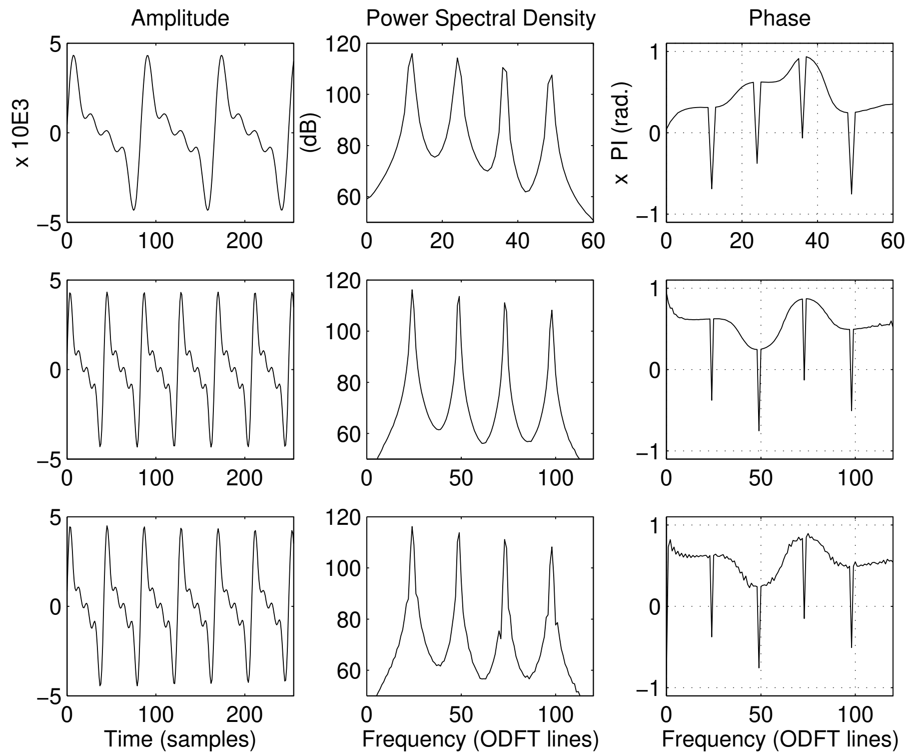

Figure 7.22: Frequency expansion of a four tone harmonic signal by a factor of two. Top: original signal. Middle: synthesized (frequency expanded) signal. Bottom: processed signal. For clarity only 1/4 of the time segment is displayed. The sampling frequency is 48 kHz, $N = 1024$ and $(\ell + \Delta\ell) = 12.3$.

The first example is the signal illustrated in Figure 7.22 (top).

As its time and spectral representation reveal, it was synthesized with four tones, harmonically related, whose fundamental frequency corresponds to $(\ell + \Delta\ell) = 12.3$. Also shown is the signal ideally expanded in frequency by a factor of two. Finally, the signal on the bottom of the figure shows the output of the algorithm to the input signal when the selected time-expansion factor is two. In all cases, the spectrogram corresponds to a 1024-point time-frequency Odd-DFT transformation based on the sine window (7.8). For clarity only one fourth of the time segment is displayed. The fixed initial phases of all four tones of the input signal are zero.

By comparing the input signal to its ideally expanded version, a few conclusions can be drawn. First, as the spectral separation between the tones of the expanded version is larger than in the case of the input signal, the spectral valleys reach lower PSD values. Second, as the $\Delta\ell$ of the second and fourth harmonics of the input signal are equal to the $\Delta\ell$ of the first and second harmonics of the ideally expanded version, respectively, then their phases are also equal as can be concluded from Equation (7.29). For the same reason, the magnitude shape of these spectral peaks is also similar, as suggested by Equations (7.26), (7.27) and (7.28).

A comparison of the ideally expanded version to the output processed signal reveals that the fine temporal structure of the original signal is maintained. This is a consequence of the magnitude and phase modifications/compensations implemented by the algorithm in the process of frequency expansion, which are in agreement with those of the ideal output.

This example assumes signal stationarity and absence of temporal modulation. For most audio signals, these assumptions are reasonably realistic. For example, at 44.1 kHz sampling rate, time segments with 1024 samples (which is an interesting value for N, allowing fine frequency resolution), represent about 12 ms for the 50% overlap signal analysis/synthesis. This duration is typically associated to quasi-stationarity for speech signals. With the exception of sharp and short audio signals such as drum sounds, clicks and impulsive noise, most music signals have longer stationary durations [Ferre.98].

Figure 7.23 shows a time modulated tone as input to the algorithm, its ideal (synthesized) frequency expanded version by a factor of two, and the output of the algorithm when using the same factor for the frequency expansion. The frequency of the tone corresponds to $(\ell+\Delta\ell) = 12.3$

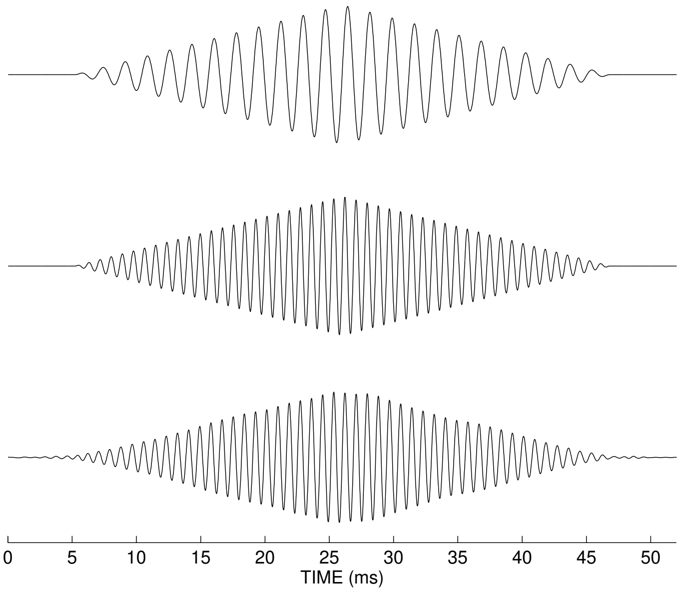

Figure 7.23: Frequency expansion of a time modulated tone by a factor of two. Top: original signal. Middle: synthesized (frequency expanded) signal. Bottom: processed signal. The sampling frequency is 48 kHz, $N = 1024$ and $(\ell + \Delta\ell) = 12.3$.

Odd-DFT spectral lines. It can be seen that the output signal exhibits the same characteristics of the ideal signal, namely the same time envelope of the original signal. However, a small fluctuation of this shape is noticeable. This is related to the fact that the duration of the signal implies about four overlap-add analysis/synthesis operations, where the phase modification did account for the local phase relationship due to the temporal modulation, but did not account for an additional constant phase offset at $k = \ell - 1$ (besides that considered in Equation (7.25)), due to the time modulation. As explained in Sub-section 7.4.3.3, this additional phase contribution is ignored based on the assumption that the shape of the time modulation is smooth relative to the length on the analysis window (which is not at all the case of Figure 7.23) and also in order to keep the complexity of the algorithm low.

Surprisingly, the results obtained with a tone modulated by a time envelope significantly shorter than the length of the analysis window, are quite good. This is illustrated in Figure 7.24 where a tone with the same frequency of the previous example is converted into a tonal pulse. The processing conditions were maintained. It can be seen that the output signal closely

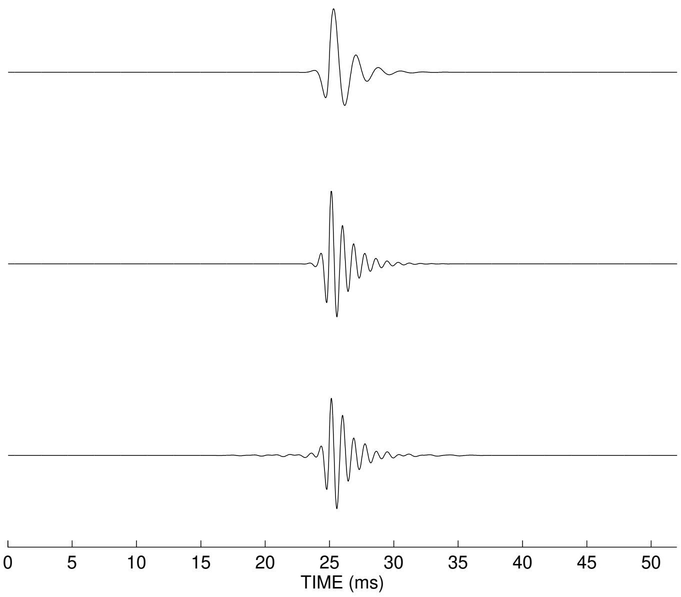

Figure 7.24: Frequency expansion of a tonal pulse by a factor of two. Top: original signal. Middle: synthesized (frequency expanded) signal. Bottom: processed signal. The sampling frequency is 48 *kHz*, $N = 1024$ and $(\ell + \Delta\ell) = 12.3$.

resembles the ideal frequency-expanded reference signal, the most obvious difference being a local magnitude misalignment, which has no impact on the sound quality. In addition, the fact that the length of the analysis window, which defines the time resolution of the filter bank, is about 20 ms, and that no serious pre-echo effects are present, as happens frequently in audio coding [Jayant93], also comes as an encouraging result.

Table 7.3 presents the Signal-to-Noise Ratios (SNR) showing the error in the time domain between the ideal output signal and the actual output signal for four different test signals. The first three correspond to the signals described above. The fourth signal is a sine sweep whose frequency increases from 100 *Hz* to 1 *kHz* in 40 ms. In all cases the duration of the signal corresponds to four frames with 50% overlap.

Table 7.3: Qualitative evaluation of the time-scaling algorithm for 4 increasingly difficult signals. The SNR figures indicate the errors in the time domain introduced by the spectral modification process relative to the ideal output signal (see text). $N = 1024$ and the duration of the signals is 4 audio frames ($4\frac{N}{2}$).

item	**harmonic complex**	**modulated tone**	**tonal pulse**	**sine sweep**
SNR (dB)	28.4	15.4	14.6	1.6

The SNR figures in Table 7.3 are not appropriate to inform about signal quality since the ideal and actual output signals sound indistinguishable, except for the sine sweep, where differences are noticeable but not annoying. Therefore, the SNR figures should be interpreted qualitatively rather than quantitatively.

Not surprisingly, the best score is obtained by the harmonic complex because it is a stationary signal and this condition minimizes the errors introduced by the algorithm. In all other cases, the deviations from stationarity get more and more evident, given that the duration of the strong changes is comparable to or less than the size of the transform (N). The tonal pulse has a lower SNR than the modulated tone because the slope of its time envelope relative to the size of the transform is sharper. The low SNR in the case of the sine sweep reflects the fact that due to the nature of its analysis/synthesis filter bank, the algorithm has an inherent difficulty in dealing with frequency modulated signals. Nevertheless, this objective assessment of the performance of the algorithm for this class of signals is not commensurate with the subjective assessment based on listening tests[10], the former being much less encouraging than the latter. Further research is needed to deal with frequency modulated signals and to improve the performance of the algorithm for audio signals which are either strongly non-stationary, or quasi-stationary but with complex time envelopes.

The performance of the time-expansion algorithm for arbitrary audio signals is quite acceptable, particularly in the case of wideband speech signals. In fact, objectionable signal artifacts are not generated even for scaling factors as high as four. In some cases, the most evident distortion is an unobjectionable form of reverberation due to insufficient frequency resolution of the filter bank. Phase smearing exists during unvoiced signal segments but this has not a perceptual impact since phase is not too much important in these regions, and since the algorithm does not force a phase coherence between successive synthesis segments, as it happens frequently with other frequency domain algorithms[11] [Moulin.95].

In the case of very wideband signals such as impulsive noise, and sound attacks, the algorithm fragments the wideband character of the signals during the process of spectral expansion. As

[10]Only one trained listener using headphones was involved.

[11]Moulines [Moulin.95, page 187] indicates that the resulting unvoiced speech has a "tonal" quality.

a consequence, transient smearing is clearly audible in the case of very sharp sound attacks. When strongly non-stationary sounds occur, a solution to this problem would be to switch to a finer time resolution based analysis/synthesis, as usually done in the context of perceptual audio coding.

Distortions also arise due to signal components of very low frequency. In fact, the algorithm is unable to perform frequency estimation and expansion of components whose center frequency is below the frequency resolution of the filter bank, *i.e.* less than f_s/N, where f_s is the sampling rate and N is the length of the analysis/synthesis window. The main reason is that in this case, the algorithm presented in Sub-section 7.4.3.5 is not able to estimate the fractional frequency and phase.

Figure 7.25 shows a short-time segment of a female speech extracted from track 49 of the SQUAM compact disc [Jakubo.88], as well as its time expanded replica by a factor of two. This

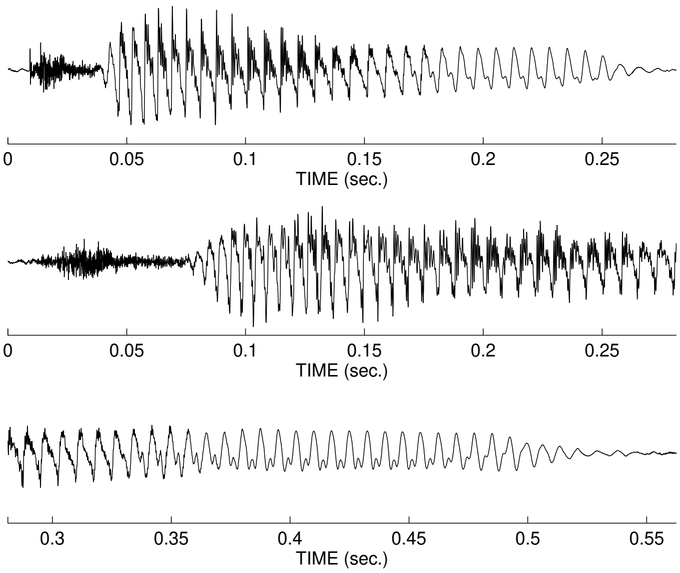

Figure 7.25: Time expansion of a female speech passage, *'time'*, by a factor of two. Top: original signal. Middle and bottom: processed signal. The original sampling rate is 16 kHz but the processing was performed using 32 kHz sampling rate and $N = 1024$.

segment corresponds to the utterance of the word *'time'* and was sub-sampled to 16 kHz. In order to avoid bandwidth reduction, as explained in Sub-section 7.4.3.1, it was up-sampled to 32 kHz prior to time-scale expansion. The processing was performed with $N = 1024$. It can be seen that the time-expanded signal mirrors the main features of the original signal, particularly in the voiced region. When listening to the result, it sounds free from obvious artifacts. The naturalness of the unvoiced regions as well as of the original pitch of the speech are absolutely maintained, despite a few subtle distortions only reasonably noticeable when using headphones.

Another example of time-expansion implemented in the same conditions as above, is shown in Figure 7.26. This figure illustrates the time-expanded version of the word *'to'* by a factor of two and three.

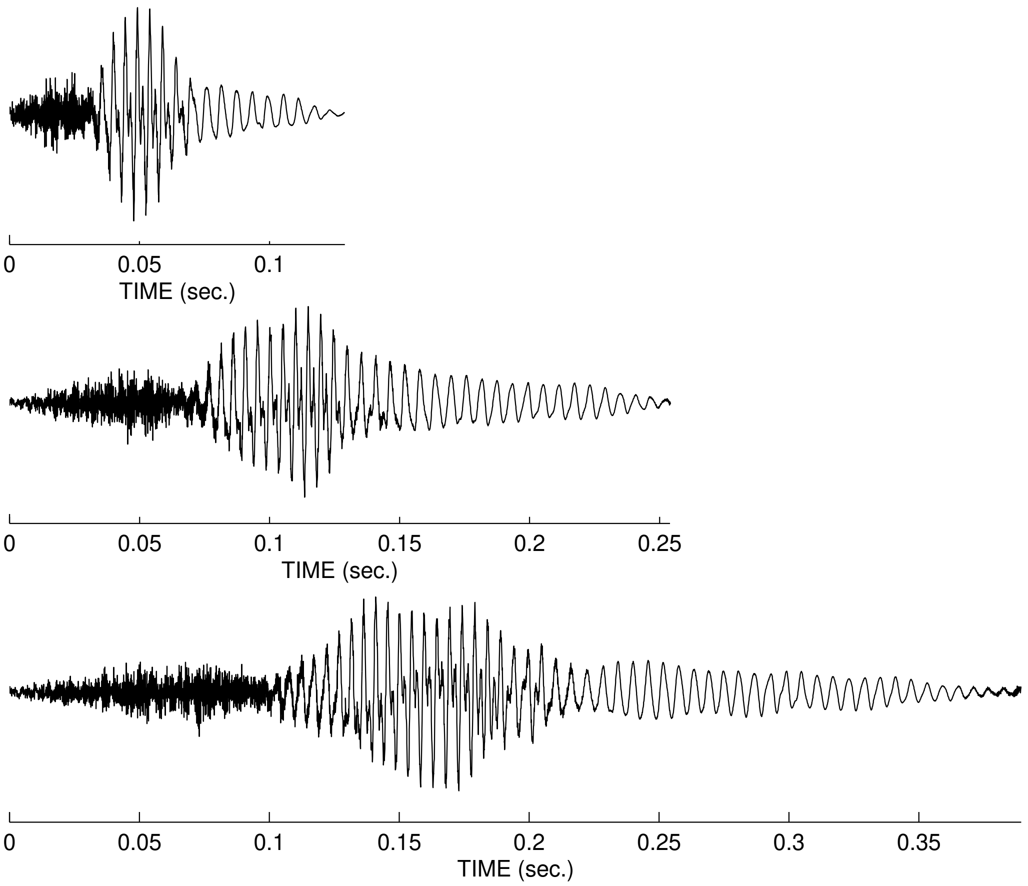

Figure 7.26: Time expansion of a female speech passage, *'to'*, by a factor of two and three. Top: original signal. Middle: time-expanded version by a factor of two. Bottom: time-expanded version by a factor of three. The original sampling rate is 16 kHz but the processing was performed using 32 kHz sampling rate for two-times expansion, 48 kHz sampling rate for three-times expansion, and in both cases $N = 1024$.

A few examples of time-scaled wideband speech are available on the Internet that reflect time scaling factors as large as four. The Web address is http://www.inescn.pt/~ajf/timescale.html. The simulation conditions are also detailed in the Web page.

7.5 Summary

We have shown in this Chapter that using a convenient reformulation of the analysis/synthesis framework of ASC, other functionalities besides compression can be efficiently implemented.

Three functionalities have been addressed: graphic equalization, noise modulation, and time-scale expansion.

Graphic equalization is obtained by means of an adaptation of the classic methods of filtering in the frequency domain. The new equalization technique is well suited for real-time implementation. Further research will involve parametric equalization.

Noise modulation aims at generating impaired audio material from unimpaired audio material. This has been achieved by means of a controlled filtering, and contamination of the audio signal with artificial noise. Further research will include other forms of spectral manipulation such as selectable filtering between minimum-phase and maximum-phase FIR filters.

Time-scale expansion has been implemented using a new technique of frequency-scale expansion, and using a conventional sample-rate conversion post-processing in the time domain. Further research will include fractional time-scale expansion, as well as time-scale compression.

Chapter 8

Conclusion

Hello world !

8.1 The Dissertation: Results and Contributions

In this Dissertation, we have covered a multidisciplinary spectrum of topics, which included tools for algorithm development and evaluation, hearing research and psychoacoustic modeling, multirate filter banks for audio analysis and manipulation, and techniques for special effects in audio. These topics have been addressed in separate Chapters and have all received original contributions.

The main focus of this Dissertation, which builds on the previous topics, has been concentrated on techniques for the coding of audio signals. In particular, we have reviewed in Chapter 5, speech, wideband speech, and high quality audio coding techniques. This review has been modulated by two concerns: to give an impressionistic picture of the historical evolution of coding paradigms, concepts and techniques, and to provide a representative snapshot of the current state-of-the-art in high quality audio coding.

In Chapter 6 we have described the central contribution of this Dissertation: ASC - a new perceptual audio coder. ASC condenses our three claims:

1. the optimum coding performance is only achieved if the observability space spanned by microscopic, macroscopic, objective and subjective angles of view, precedes the coding system controllability, in a global perspective,

2. harmonic signals invoke a special perception mechanism of the human auditory system, and elicit specific masking properties that should be dealt with in perceptual audio coding by activating a dedicated coding mode, inasmuch as non-stationary regions of audio signals

are dealt with relative to the quasi-stationary regions,

3. gains in complexity reduction can be achieved and a sparing of independent post-processing modules is obtained if the existing coding tools in a perceptual audio coder such as ASC, are judiciously reutilized, or exploited in order to implement multiple functionalities.

The coding quality delivered by ASC has been evaluated in formal subjective tests carried out by the MPEG-4 Audio Sub-Group. The results of these tests have been described and the performance of ASC relative to the MPEG-1 Audio standard has been characterized.

8.2 The Future: Research Topics and New Areas

After a great deal of investment in the area of algorithm development and optimization, which has given rise to a few relevant results, and to several papers submitted for publication with success, we sense we are at the verge of two lines of evolution that elicit the primacy of a pragmatism oriented to the implementation over the perfectionism oriented strictly to subjective quality and coding bit rate.

A first line of evolution concerns naturally the continuity of the research described in this Dissertation. ASC is a promising coding algorithm that may benefit from further research regarding:

- its efficient implementation for real-time coding,
- the characterization of additional coding gains provided by different types of quantizers,
- the incorporation of error concealment and correction techniques,
- its adaptation to the coding of speech specific signals,
- the evaluation of tandeming effects,
- its extension to the coding of stereo signals and surround sound.

The theme of integration of other functionalities, besides coding, allowing to share the analysis/synthesis framework of ASC, also inspires further research:

- an in depth study is necessary to evaluate the effects of accumulating coding with other co-existing functionalities,
- new functionalities such as voice disguising and voice substitution for the protection of victims will certainly improve the versatility and practical interest of ASC,

- the synthesis of additional sound channels representing a desired spatialization is an interesting move towards virtual reality.

Any one of the above topics has the potential to resume our active participation within the MPEG-4 standardization activities, especially in the context of the recently announced extension of the MPEG-4, the so-called MPEG-4 Phase 2, scheduled from January 1999 till December 1999.

A second line of evolution may represent an inflection of the first line, but broadens the scope and thus our challenges in audio processing. A first topic deals with the automatic classification of audio documents in order to ease its *semantic* annotation, access, and retrieval. This topic is in line with the MPEG-7 standardization activities. A second topic represents to us a new area of signal analysis, synthesis and coding, that could not make a better use of *applied audio research*: hearing diagnosis and rehabilitation.

Bibliography

[Adams91] John W. Adams. A New Optimal Window. *IEEE Transactions on Signal Processing*, 39(8):1753–1769, August 1991.

[Begault94] Durand R. Begault. *3-D Sound for Virtual Reality and Multimedia.* Academic Press, 1994.

[Bellan.89] Maurice Bellanger. *Digital Processing of Signals.* John Willey & Sons, 1989.

[Békésy60] Georg Von Békésy. *Experiments in Hearing.* American Institute of Physics and McGraw-Hill, 1960.

[Beranek90] Leo L. Beranek. *Acoustics.* American Institute of Physics, 1990.

[Blauert83] Jens Blauert. *Spatial Hearing.* The MIT Press, 1983.

[Bosi96] Marina Bosi et al. ISO/IEC MPEG-2 Advanced Audio Coding. *101st Convention of the Audio Engineering Society*, November 1996. Preprint n. 4382.

[Bosi97] Marina Bosi. Perceptual Audio Coding. *IEEE Signal Processing Magazine*, pages 43–49, September 1997.

[Brand.87] Karleinz Brandenburg. OCF-A New Coding Algorithm for High Quality Sound Signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 141–144, 1987.

[Brand.88] K. Brandenburg and D. Seitzer. OCF: Coding High Quality Audio with Data Rates of 64 Kbit/sec. *85th Convention of the Audio Engineering Society*, November 1988. Preprint n. 2723.

[Brand.89] Karlheinz Brandenburg and Dieter Seitzer. Low Bit Rate Coding of High-Quality Digital Audio: Algorithms and evaluation of quality. *7th International Conference of the Audio Engineering Society*, May 1989. Paper 6F.

[Brand.90] K. Brandenburg and J. D. Johnston. Second Generation Perceptual Audio Coding: The Hybrid Coder. *88th Convention of the Audio Engineering Society*, March 1990. Preprint n. 2937.

[Brand.91] K. Brandenburg, Jurgen Herre, James D. Johnston, Yannick Mahieux, and Ernst F. Schroeder. Aspec: Adaptive spectral entropy coding of high quality music signals. *90th Convention of the Audio Engineering Society*, February 1991. Preprint n. 3011.

[Brand.92a] K. Brandenburg and T. Sporer. NMR and Masking Flag: Evaluation of Quality using Perceptual Criteria. In *11th International AES Conference on Audio Test and Measurement*, pages 169–179, 1992.

[Brand.92b] Karlheinz Brandenburg and Gerhard Stoll. The ISO/MPEG-Audio Codec: A generic Standard for Coding of High Quality Digital Audio. *92nd Convention of the Audio Engineering Society*, March 1992. Preprint n. 3336.

[Brown96] Judith C. Brown. Frequency Ratios of Spectral Components of Musical Sounds. *Journal of the Acoustical Society of America*, 99(2):1210–1218, February 1996.

[Chen96] Juin-Hwey Chen and Dongmei Wang. Transform Predictive Coding of Wideband Speech Signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 275–278, 1996.

[Chen97] Juin-Hwey Chen. A Candidate Coder for the ITU-T's New Wideband Speech Coding Standard. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1359–1362, 1997.

[Cheung95] Shiufun Cheung and Jae S. Lim. Incorporation of Biorthogonality Into Lapped Transforms for Audio Compression. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3079–3082, 1995.

[Cheung96] Shiufun Cheung. *Biorthogonality in Lapped Transforms: A Study in High-Quality Audio Compression.* PhD thesis, Massachusetts Institute of Technology, 1996.

[Chu85] Peter L. Chu. Quadrature Mirror Filter Design for an Arbitrary Number of Equal Bandwidth Channels. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(1):203–218, February 1985.

[Crochi.79] R. E. Crochiere and J. M. Tribolet. Frequency Domain Techniques for Speech Coding. *Journal of the Acoustical Society of America*, 66(6):1642–1646, December 1979.

[Crochi.83] Ronald E. Crochiere and Lawrence R. Rabiner. *Multirate Digital Signal Processing.* Prentice-Hall, 1983.

[Davis93] Mark Davis. The AC-3 Multichannel Coder. *95th Convention of the Audio Engineering Society*, October 1993. Preprint n. 3774.

[Dehery91] Yves François Dehery, Michel Lever, and Pierre Urcun. A Musicam Source Codec for Digital Audio Broadcasting and Storage. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3605–3608, 1991.

[Dolson86] Mark Dolson. The Phase Vocoder: A Tutorial. *Journal of Computer Music*, 84(4):14–27, 1986.

[Fieldi.92] L. Fieldier and M. Bosi. AC-2: High Quality Digital Audio Coding for Broadcasting and Storage. In *Broadcast Engineering Conference*, pages 98–105, April 1992. Las Vegas.

[Ferre.92] Aníbal J. S. Ferreira. Codificação Perceptual de Áudio Digital Estereofónico. Master's thesis, Faculdade de Engenharia da Universidade do Porto, Portugal, January 1992.

[Ferre.93] Aníbal J. S. Ferreira. Wideband Speech and Music Coding with PAC. *AT&T Bell Laboratories internal report*, December 1993.

[Ferre.94] Aníbal J. S. Ferreira. Codificação de Áudio de Alta Qualidade usando MPEG: Enquadramento, Resultados e Perspectivas Futuras. In 1^{o} *Encontro Nacional do Colégio de Engenharia Electrotécnica*, pages 459–464, 1994. Ordem do Engenheiros-Lisboa-Maio.

[Ferre.95a] Aníbal J. S. Ferreira. Optimizing High Quality Audio Coding: Advantages of Full System Observability. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3063–3066, 1995. Detroit, MI.

[Ferre.95b] Aníbal J. S. Ferreira. Tonality Detection in Perceptual Coding of Audio. *98th Convention of the Audio Engineering Society*, February 1995. Preprint n. 3947.

[Ferre.95c] Aníbal J. S. Ferreira and José M. N. Vieira. An Efficient 20 Band Audio Equalizer. *98th Convention of the Audio Engineering Society*, February 1995. Preprint n. 3992.

[Ferre.96a] Aníbal J. S. Ferreira. Audio Spectral Coder. *100th Convention of the Audio Engineering Society*, May 1996. Preprint n. 4201.

[Ferre.96b] Aníbal J. S. Ferreira. Convolutional Effects in Transform Coding with TDAC: An Optimal Window. *IEEE Transactions on Speech and Audio Processing*, 4(2):104–114, March 1996.

[Ferre.96c] Aníbal J. S. Ferreira. Perceptual Coding of Harmonic Signals. *100th Convention of the Audio Engineering Society*, May 1996. Preprint n. 4177.

[Ferre.98] Aníbal J. S. Ferreira. Perceptual Audio Coding and the Choice of an Analysis/Synthesis Filter Bank and Psychoacoustic Model. *104th Convention of the Audio Engineering Society*, May 1998. Preprint n. 4671.

[Flanag.66] J. L. Flanagan and R. M. Golden. Phase Vocoder. *Bell Systems Technical Journal*, 45(9):1493–1509, November 1966.

[Fletch.40] H. Fletcher. Auditory Patterns. *Revue of Modern Physics*, (12):47–65, 1940.

[Gelfand90] Stanley A. Gelfand. *Hearing*. Marcel Dekker, 1990.

[Gersho94] Allen Gersho. Advances in Speech and Audio Compression. *Proceedings of the IEEE*, 82(6):900–918, June 1994.

[Gibson84] Jerry D. Gibson. Adaptive Prediction for Speech Encoding. *IEEE ASSP Magazine*, pages 12–26, July 1984.

[Gosse97] K. Gosse, F. Moreau de Saint Martin, X. Durot, P. Duhamel, and J. B. Rault. Subband Audio Coding with Synthesis Filters Minimizing a Perceptual Distortion. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 347–350, 1997.

[Gray92] A. Gersho and R. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Press, 1992.

[Greenw.90] D. Greenwood. A cochlear frequency-position function for several species–20 years later. *Journal of the Acoustical Society of America*, 87(5):2592–2605, May 1990.

[Griffin84] D. W. Griffin and J. S. Lim. Signal Estimation from the Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2):236–243, April 1984.

[Hartma.96] William M. Hartmann. Pitch, periodicity, and auditory organization. *Journal of the Acoustical Society of America*, 100(6):3491–3502, December 1996.

[Heinle97] Frank Heinle. A New Approach to the Compensation of Aliasing in Transform and Subband Coders. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2457–2460, 1997.

[Hellman72] Rhona P. Hellman. Asymmetry of masking between noise and tone. *Perception and Psychoacoustics*, 11(3):241–246, 1972.

[Henriq.90] Juan A. Henriquez, Terry E. Riemer, and Russel E. Trahan. A Phase-Linear Audio Equalizer: Design and Implementation. *Journal of the Audio Engineering Society*, 38(9):653–666, September 1990.

[Herre96] Jürgen Herre and James D. Johnston. Enhancing the Performance of Perceptual Audio Coders by Using temporal Noise Shaping (TNS). *101st Convention of the Audio Engineering Society*, November 1996. Preprint n. 4384.

[Hoogen.94] Abraham Hoogendoorn. Digital Compact Cassette. *Proceedings of the IEEE*, 82(10):1479–1489, October 1994.

[Isabel.90] S. H. Isabelle and J. S. Lim. On Modulated Filter Banks for Image Coding Applications. In *IEEE Digital Signal Processing Workshop*, September 1990. New Platz, NY.

[ISO93] ISO/IEC JTC 1/SC 29/WG 11. ISO/IEC 11172-3: Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s- Part 3: Audio, Mars 1993.

[ISO94] ISO/IEC JTC 1/SC 29/WG 11. ISO/IEC 13818-3: Information Technology - Generic Coding of Moving Pictures and Associated Audio Information- Part 3: Audio, December 1994.

[ISO96a] ISO/IEC JTC 1/SC 29/WG 11. ISO/IEC N1144. 33rd MPEG Meeting, January 1996.

[ISO96b] International Organization for Standardization. ISO 389-7: Acoustics-Reference Zero for the Calibration of Audiometric Equipment. Part 7: Reference Threshold of Hearing under Free-Field and Diffuse-Field Listening Conditions, 1996. Geneva, Switzerland.

[ITU72] CCITT Recommendation G.711. Pulse code modulation (PCM) of voice frequencies. Bue Book Fascicle III.4, October 1972.

[ITU88a] CCITT Recommendation G.721. 32Kbit/s Adaptive Differential Pulse Code Modulation. Bue Book Fascicle III.4, October 1988.

[ITU88b] CCITT Recommendation G.722. 7 KHz Audio-coding within 64 Kbit/s. Bue Book Fascicle III.4, November 1988.

[ITU88c] CCITT Recommendation G.723. Extensions of recommendation 32Kbit/s adaptive differential pulse code modulation to 24 and 40 Kbit/s for digital circuit multiplication equipment application. Bue Book Fascicle III.4, October 1988.

[ITU88d] CCITT Recommendation J.41. Characteristics of Equipment for the Coding of Analogue High Quality Sound-Programme Signals for the Transmission on 384 Kbit/s Channels. Bue Book Fascicle III.6, November 1988.

[ITU88e] CCITT Recommendation J.43. Characteristics of Equipment for the Coding of Analogue High Quality Sound-Programme Signals for the Transmission on 320 Kbit/s Channels. Bue Book Fascicle III.6, November 1988.

[Iwadar.92] M. Iwadare et al. A 128 Kbit/s Hi-Fi Audio Codec Based on Adaptive Transform Coding with Adaptive Block Size MDCT. *IEEE Journal on Selected Areas in Communication*, 10(1), January 1992.

[Iwakam.95] N. Iwakami, T. Moriya, and S. Miki. High-Quality Audio-Coding at Less than 64 Kbit/s by Using Transform-Domain Weighted Interleave vector Quantization (Twin-VQ). In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3095–3098, 1995.

[Jakob95] Conrad F. Jakob and Alan B. Bradley. Quantization Noise in Perfect Reconstruction Filter Banks and Its Consequences for High Fidelity Audio Compression. *5th Australian Regional Convention of the Audio Engineering Society*, April 1995. Preprint n. 4046.

[Jakob96] Conrad Jakob and Alan Bradley. Minimising the Effects of Subband Quantization of the Time Domain Aliasing Cancellation Filter Bank. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1033–1036, 1996.

[Jakubo.88] H. Jakubouske and G. Spikofski. SQUAM-The EBU Compact Disc for Subjective Assessment of Audio Systems. *EBU Review*, (227), February 1988.

[Jayant84] N. S. Jayant and Peter Noll. *Digital Coding of Waveforms*. Prentice-Hall, 1984.

[Jayant93] Nikil Jayant, James Johnston, and Robert Safranek. Signal Compression Based on Models of Human Perception. *Proceedings of the IEEE*, 81(10):1385–1422, October 1993.

[Johns.80] J. D. Johnston. A Filter Family Designed for Use in Quadrature Mirror Filter Banks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 291–294, April 1980.

[Johns.88a] J. D. Johnston. Method of Estimating the Perceptual Entropy of an Audio Signal. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1988.

[Johns.88b] James D. Johnston. Transform Coding of Audio Signals Using Perceptual Noise Criteria. *IEEE Journal on Selected Areas in Communication*, 6(2), February 1988.

[Johns.89] James D. Johnston. Perceptual Transform Coding of Wideband Stereo Signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1993–1996, 1989.

[Johns.92] J. D. Johnston and A. J. Ferreira. Sum-Difference Stereo Transform Coding. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages II–569 II–572, 1992. San Francisco, CA.

[Johns.96] James D. Johnston, Juergen Herre, Mark Davis, and Uwe Gbur. MPEG-2 NBC Audio-Stereo and Multichannel Coding Methods. In *101st AES Convention*, November 1996. Preprint no. 4383.

[Keyhl95] Keyhl M., Herre J., Ferreira A., and Gilchrist N. The Effect of Low Bit-rate Coding Upon Impaired Audio Material. In *International Broadcasting Convention*, pages 151–158, 1995. Conf. Pub. No. 413.

[Kondoz94] A. M. Kondoz. *Digital Speech (Coding for Low Bit Rate Communication Systems)*. John Wiley & Sons, Inc., 1994.

[Kraght92] Paul H. Kraght. A Linear-Phase Digital Equalizer with Cubic-Spline Frequency Response. *Journal of the Audio Engineering Society*, 40(5):403–414, May 1992.

[Kroon92] Peter Kroon. A High-Quality Multirate Real-Time CELP Coder. *IEEE Journal on Selected Areas in Communications*, 10(5):850–857, June 1992.

[Kulp88] Barry D. Kulp. Digital Equalization Using Fourier Transform Techniques. *85th Convention of the Audio Engineering Society*, November 1988. Preprint n. 2694.

[Laroch.97] Jean Laroche and Mark Dolson. Improved Phase Vocoder Time-Scale Modification of Audio. submitted to IEEE Transactions on Speech and Audio Processing.

[LeRoux95] J-M. LeRoux, R. Lefebvre, and J-P. Adoul. Comparison of the Wavelet Decomposition and the Fourier Transform in TCX Enconding of Wideband Speech and Audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3083–3086, 1995.

[Lin95] Gang-Janp Lin, Sau-Gee Chen, and Terry Wu. High Quality and Low Complexity Pitch Modification of Acoustic Signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2987–2990, 1995.

[Lokhoff91] Gerard Lokhoff. DCC - Digital Compact Cassette. *IEEE Transactions on Consumer Electronics*, 37(3):702–706, August 1991.

[Macon96] Michael W. Macon and Mark A. Clements. Speech Concatenation and Synthesis using an Overlap-Add Sinusoidal Model. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[Malvar92] Henrique Malvar. *Signal Processing with Lapped Transforms.* Artech House, Inc., 1992.

[Malvar97] Henrique Malvar. Lapped Biorthogonal Transforms for Transform Coding with Reduced Blocking and Ringing Artifacts. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2421–2424, 1997.

[McCle.73a] James H. McClellan and Thomas W. Parks. A Unified Approach to the Design of Optimum FIR Linear-Phase Digital Filters. *IEEE Transactions on Circuit Theory*, CT-20(6):697–701, November 1973.

[McCle.73b] James H. McClellan, Thomas W. Parks, and Lawrence R. Rabiner. A Computer Program for Designing Optimum FIR Linear Phase Digital Filters. *IEEE Transactions on Audio and Electroacoustics*, AU-21(6):506–526, November 1973.

[Moore89] Brian C. J. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, 1989.

[Moore97] Brian C. J. Moore, Brian R. Glasberg, and Thomas Baer. A Model for the Prediction of Thresholds, Loudness, and Partial Loudness. *Journal of the Audio Engineering Society*, 45(4):224–240, April 1997.

[Moriya97] T. Moriya, N. Iwakami, K. Ikeda, and S. Miki. A Design of Transform Coder for Both Speech and Audio Siganls at 1 bit/sample. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1371–1374, 1997.

[Moulin.95] Eric Moulines and Jean Laroche. Non-Parametric Techniques for Pitch-Scale and Time-Scale Modification of Speech. *Speech Communication*, 16:175–205, 1995.

[Noll95] Peter Noll. Digital Audio Coding for Visual Communications. *Proceedings of the IEEE*, 83(6):925–943, June 1995.

[Noll97] Peter Noll. MPEG Digital Audio Coding. *IEEE Signal Processing Magazine*, pages 59–81, September 1997.

[Oppenh.75] Alan V. Oppenheim and Ronald W. Schafer. *Digital Signal Processing.* Prentice-Hall Inc., 1975.

[Ordent.91] Eric Ordentlich and Yair Shoham. Low-Delay Code-Excited Linear-Predictive Coding of Wideband Speech at 32 Kbps. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 9–12, 1991. paper S1.3.

[Pailla.92] B. Paillard, P. Mabilleau, S. Morissette, and J. Soumagne. PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals. *Journal of the Audio Engineering Society*, 40:21–31, January/February 1992.

[Parks87] T. W. Parks and C. S. Burrus. *Digital Filter Design.* John Wiley and Sons, Inc., 1987.

[Pielem.96] W. J. Pielemeier, G. H. Wakefield, and M. H. Simon. Time-Frequency Analysis of Musical Signals. *Proceedings of the IEEE*, 84(9):1216–1230, September 1996.

[Portnof81] Michael R. Portnoff. Time-Scale Modification of Speech based on Short Time Fourier Analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3):374–390, June 1981.

[Princen86] John P. Princen and Alan Bernard. Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-34(5):1153–1161, October 1986.

[Princen87] J. P. Princen, A. W. Johnson, and A. B. Bradley. Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2161–2164, April 1987.

[Princen95] John Princen and James D. Johnston. Audio Coding with Signal Adaptive Filter Banks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3071–3074, 1995.

[Quacke.91] Schuyler Quackenbush. A 7 KHz Bandwith, 32 Kbit/s Speech Coder for ISDN. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–4, 1991.

[Quatie.92] Thomas. F. Quatieri and Robert J. McAulay. Shape Invariant Time-Scale and Pitch Modification of Speech. *IEEE Transactions on Signal Processing*, 40(3):497–510, March 1992.

[Quatie.95] T. F. Quatieri, R. B. Dunn, and T. E. Hanna. A Subband Approach to Time-Scale Expansion of Complex Acoustic Signals. *IEEE Transactions on Speech and Audio Processing*, 3(6):515–519, November 1995.

[Rabiner72] Lawrence R. Rabiner. Linear Program Design of Finite Impulse Response (FIR) Digital Filters. *IEEE Transactions on Audio and Electroacoustics*, AU-20(4):280–288, October 1972.

[Rabiner73] Lawrence R. Rabiner and Otto Herrmann. On the Design of Optimum FIR Low-Pass Filters with Even Impulse Response Duration. *IEEE Transactions on Audio and Electroacoustics*, AU-21(4):329–336, August 1973.

[Rabiner78] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[Rioul91] Olivier Rioul and Martin Vetterli. Wavelets and Signal Processing. *IEEE ASSP Magazine*, pages 14–37, October 1991.

[Roeder.95] Juan G. Roederer. *The Physics and Psychophysics of Music-An Introduction*. Springer-Verlag, 1995.

[Roucos85] Salim Roucos and Alexander M. Wilgus. High Quality Time-Scale Modification of Speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 13.6.1–13.6.4, 1985.

[Salami97] R. Salami, C. Laflamme, B. Bessette, and J-P Adoul. Description of the ITU-T Recommendation G.729 Annex A: Reduced Complexity 8 Kbit/s CS-ACELP Codec. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 775–778, 1997.

[Scharf70] Bertram Scharf. Critical Bands. In Jerry V. Tobias, editor, *Foundations of Modern Auditory Theory*, chapter 5. Academic Press, 1970.

[Schroe.75] Manfred R. Schroeder. Models of Hearing. *Proceedings of the IEEE*, 63(9):1332–1350, September 1975.

[Schroe.79] M. S. Schroeder, B. S. Atal, and J. L. Hall. Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear. *Journal of the Acoustical Society of America*, 66(6):1647–1652, December 1979.

[Schroe.85] M. Schroeder and B. Atal. Code-Excited Linear Prediction: High Quality Speech at Low Bit Rates. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 937–940, 1985.

[Silver.92] Herbert Silverstein, Robert J. Wolfson, and Seth Rosenberg. Diagnosis and Management of Hearing Loss. *Clinical Symposia (Ciba-Geigy)*, 44(3), 1992.

[Sinha93] Deepen Sinha and Ahmed Tewfik. Low Bit Rate Transparent Audio Compression Using Adapted Wavelets. *IEEE Transactions on Signal Processing*, 41(12):3464–3479, December 1993.

[Sinha96] Deepen Sinha and James D. Johnston. Audio Compression at Low Bit Rates Using a Signal Signal Adaptive Switched Filter Bank. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[Soman93] Anand K. Soman and P. P. Vaidyanathan. Coding Gain in Paraunitary Analysis/Synthesis Systems. *IEEE Transactions on Signal Processing*, 41(5):1824–1835, May 1993.

[Soumag.86] Joel Soumagne et al. A Comparative Study of the Proposed High Quality Coding Schemes for Digital Music. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1.6.1–1.6.4, 1986.

[Spanias94] Andreas S. Spanias. Speech Coding: A Tutorial Review. *Proceedings of the IEEE*, 82(10):1541–1582, October 1994.

[Sporer92] T. Sporer, K. Brandenburg, and B. Edler. The Use of Multirate FilterBanks for Coding of High Quality Audio. In *6th European Signal Processing Conference*, volume 1, pages 211–214, June 1992.

[Stevens57] S. S. Stevens. On the Psychophysical Law. *Psychology Revue*, 64(3):153–181, 1957.

[Sugiya.90] A. Sugiyama, F. Hazu, M. Iwadare, and T. Nishitani. Adaptive Transform Coding with Adaptive Block Size (ATC-ABS). In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1093–1096, 1990.

[Tobias72] Jerry V. Tobias. *Foundations of Modern Auditory Theory*. Academic Press, 1972. Volume II.

[Tranco.88] I. Trancoso, L. Almeida, J. Rodrigues, J. Marques, and J. Tribolet. Harmonic Coding: State-of-the-Art and Future Trends. *Speeech Communication*, 7(2):239–245, July 1988.

[Tribol.77] José M. Tribolet. A New Phase Unwrapping Algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(2):170–177, April 1977.

[Tribol.79] J. M. Tribolet and R. E. Crochiere. Frequency Domain Coding of Speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(5):512–530, October 1979.

[Tribol.80] J. M. Tribolet and R. E. Crochiere. A Modified Transform Coding Scheme with Post-Processing Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 336–339, 1980.

[Tsutsui92] Kyoya Tsutsui, Hiroshi Suzuki, Mito Sonohara Osamu Shimoyoshi, Kenzo Akagiri, and Robert M. Heddle. ATRAC: Adaptive Transform Acoustic Coding for Mini-Disc. *93rd Convention of the Audio Engineering Society*, October 1992. Preprint n. 3456.

[Uzun95] Necdet Uzun and Richard A. Haddad. Cyclostationary Modeling, Analysis, and Optimal Compensation of Quantization Errors in Subband Codecs. *IEEE Transactions on Signal Processing*, 43(9), September 1995.

[Vaidy.87a] P. P. Vaidyanathan. Quadrature Mirror Filter Banks, M-Band Extensions and Perfect Reconstruction Techniques. *IEEE ASSP Magazine*, 4:4–20, July 1987.

[Vaidy.87b] P. P. Vaidyanathan. Theory and Design of M-Channel Maximally Decimated Quadrature Mirror Filter Banks with Arbitrary M, Having Perfect Reconstruction Property. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35:476–492, April 1987.

[Vaidy.93] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice-Hall, 1993.

[Vargas93] Luis C. Vargas and Henrique S. Malvar. ELT-Based Wavelet Coding of High-Fidelity Audio Signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 124–127, 1993.

[Veldhu.92] Raymond N. J. Veldhuis. Bit Rates in Audio Source Coding. *IEEE Journal on Selected Areas in Communication*, 10(1), January 1992.

[Verhel.93] Werner Verhelst and Marc Roelands. An Overlap-add Technique based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages II.554–II.557, 1993.

[Wei97] X. Wei, M. Shaw, and M. Varley. Optimum Bit Allocation and Decomposition for High Quality Audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 315–318, 1997.

[Wylie96] Fred Wylie. Predictive or Perceptual Coding – apt-X and apt-Q. *100th Convention of the Audio Engineering Society*, May 1996. Preprint n. 4200.

[Yim96] S. Yim and B.I. Pawate. Computationally Efficient Algorithm for Time Scale Modification (GLS-TSM). In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[Yoshida94] Tadao Yoshida. The Rewritable MiniDisc System. *Proceedings of the IEEE*, 82(10):1492–1500, October 1994.

[Zelins.77] Rainer Zelinsky and Peter Noll. Adaptive Transform Coding of Speech Signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(4):299–309, August 1977.

[Zwicker61] E. Zwicker. Subdivision of the Audible Frequency Range into Critical Bands. *Journal of the Acoustical Society of America*, 33(2):248, February 1961.

[Zwicker90] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models*. Springer-Verlag, 1990.